Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITÀ

Ciclo XXXII

Settore Concorsuale: 10/M1 – LINGUE, LETTERATURE E CULTURE GERMANICHE
01/B – INFORMATICA

Settore Scientifico Disciplinare: L-LIN/14 – LINGUA E TRADUZIONE – LINGUA TEDESCA
INF/01 – INFORMATICA

USING DATA MINING TO REPURPOSE GERMAN LANGUAGE CORPORA
An evaluation of data-driven analysis methods for corpus linguistics

Presentata da: Jennifer-Carmen Frey

Coordinatore Dottorato
Prof.ssa Raffaella Baccolini

Supervisore
Prof. Marcello Soffritti

Co-supervisore
Dr. Aivars Glaznieks

Esame finale anno 2020

Abstract

The potential to exploit existing data resources is one of the biggest drivers of research and innovation. Recent advances in artificial intelligence, machine learning and data mining have led to a general interest in the possibilities provided when applying these methods to existing datasets of other research fields. A growing number of studies report interesting insights gained from existing data resources. Among those, there are analyses on textual data, giving reason to consider such methods for linguistics as well. However, corpus linguistics, defined as the empirical study of real-life language use (McEnery & Hardie, 2012), usually works with purposefully collected, representative language samples that aim to answer only a limited set of research questions (Hunston, 2009). Methods of data re-utilization and exploitation, like data mining and knowledge discovery are rarely considered in corpus linguistics (but see Degaetano-Ortlieb & Fankhauser, 2014; or Pölitz, 2016 for some examples), although the fast advances made in natural language processing and text mining strengthen the case for the use of machine-learning-based data-driven analysis in this field.

In this study I aim to shed some light on the potentials of data-driven analysis based on machine learning and predictive modelling for corpus linguistic studies. In particular, I investigate the possibility to repurpose existing German language corpora for linguistic inquiry by using methodologies developed for data science and computational linguistics. The study focuses on predictive modelling and machine-learning-based data mining and gives a detailed overview and evaluation of currently popular strategies and methods for analysing language corpora with computational methods.

Part I introduces strategies and methods that have already been used on language data, discusses how they can assist corpus linguistic analysis and refers to available toolkits and software as well as to state-of-the-art research and further references whenever possible, in order to allow the reader to use this overview as an entry point for personal endeavours. Part II evaluates the previously introduced methodological toolset by applying it in two differently shaped corpus studies that utilize already existing, readily available language corpora of medium size and primarily composed of German texts. Both studies are based on computational linguistic tasks that have evolved over the last few years and are increasingly used for linguistic analyses. Corpus study one explores linguistic correlates of holistic text quality ratings on student essays and is conceptually similar to automated essay scoring or predicting language competence levels, both common tasks in computation linguistics. Study two deals with age-related language features in computer-mediated communication and interprets age prediction models to answer a set of research questions that are based on previous research in the field. While both studies contribute to the study of German language by giving linguistic insights that integrate into the current understanding of the investigated phenomena, they are also conceptualized to be realistic case studies for testing the methodological toolset introduced in part I, in order to allow a detailed discussion of added values and remaining challenges of machine-learning-based data mining methods in corpus linguistics (cf. part III).

The results show that there are potential added values to using machine-learning-based data mining methods for corpus linguistics. However, the repurposing of available but relatively small corpora is difficult. Although new methodologies have been developed in order to prepare, select, transform, analyse and interpret data more efficiently, many of these techniques are still experimental, require a high background knowledge and technical skills and often depend on tools and resources that were developed for the English language. Furthermore, although strategies exist to extract more information from data or address more complex research questions, small data sizes often do not allow to observe phenomena in higher resolution, revealing few insights besides the main trends of the data. In terms of methodological rigour and efficiency, however, the methods can be an improvement over previous, mainly manual techniques, even when using existing language corpora of small to medium size.

Table of contents

Part I
Theory and application: Methods for data science in corpus linguistics

Part II
Repurposing German language corpora with data science: Two case studies

Empirical corpus study 1
Exploring holistic text quality ratings

Empirical corpus study 2
Age-related language in South Tyrolean Social Media

Part III

Evaluation: A critical discussion of data-driven analysis methods for corpus repurposing

Appendix

Introduction

The potential to exploit existing data resources is one of the biggest drivers for research and innovation. The advances that have been made in artificial intelligence and machine learning led to a general interest in the possibilities provided when feeding our continuously growing data masses to computers. While self-driving cars, weather forecasting and machine translation become reality, machines have been found to perform increasingly well in many other useful tasks including answering questions of clients through chatbots, recommending movies or predicting user preferences, and detecting fraudulent activity in business transactions or harmful content on the web. Indeed, the use cases are so manifold that over the last few years, new disciplines have emerged that occupy themselves exclusively with how to deal with and make use of data in general. *Data science, interpretable machine learning* and *explainable artificial intelligence* research ways of utilizing data resources for knowledge discovery and/or intelligent applications and provide skills and toolsets to collect, prepare, transform, analyse and interpret data with computational means, making it feasible to use the new technologies even with a minimum of knowledge of the underlying calculations. Ultimately, having computational methods for dealing with data and being able to peek into the black box of well-performing predictive systems could also enhance research in other fields, by allowing researchers to make use of the knowledge encoded in intelligent systems.

At the core of these developments are circular approaches to data, where data is re-used for new purposes, extending, repositioning or exchanging its original function, as for example in the Combined Life Cycle Model of the Data Documentation Initiative (Vardigan et al., 2008) depicted in Figure 1 (Corti et al., 2019; Kitchin, 2014; Kitchin & Lauriault, 2015; Vardigan et al., 2008).

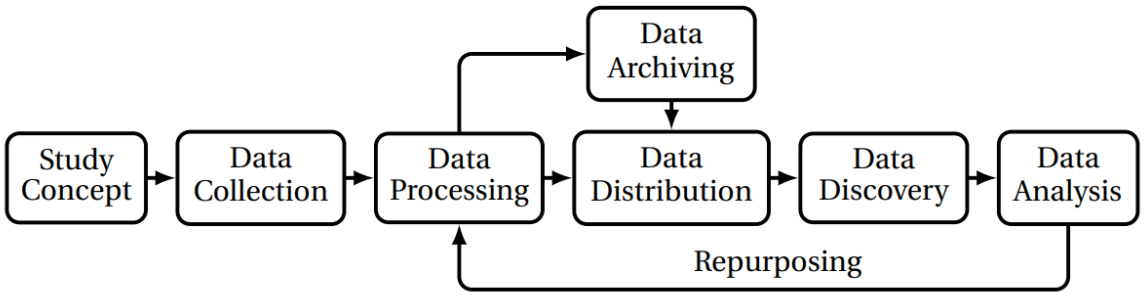


Figure 1: Data documentation initiative version 3.0 Combined Life Cycle Model (Image source: Ball, 2012; adapted from Vardigan et al., 2008)

These trends concern in particular also language data that is used in many research fields as empirical basis for investigation and is increasingly gathered and archived and distributed in the form of language corpora, as well as analysed with computational means[1]. There is therefore reason to consider such approaches also for corpus linguistics, repurposing linguistically annotated language corpora for linguistic inquiry.

However, there are often methodological challenges when re-utilizing research data, especially if it was not designed for multi-purpose studies but to answer specific research questions. Data skewedness, incomparability of data sources or missing information can easily impede the analysis

[1] Making use of the fact that natural language processing, computational linguistics and other approaches to textual analysis are some of the fields that benefitted most from the new advances in machine learning and artificial intelligence over the last few years.

and need to be taken care of by appropriate analysis methods or carefully subsampling the data (and thus potentially decreasing its size to a point that is not worth applying computational methods on). Hence, examples for successful instances of data mining that report interesting and unknown insights extracted from existing data sources are also subject to a publication bias, as unsuccessful attempts barely get published.

As with any kind of real-life observational data, this is also true for language corpora. Although data science could potentially be used to repurpose costly created and laboriously curated language corpora, corpus linguistics, i.e. the empirical study of real-life language use, has for long been based on carefully designed, representative language samples that serve to answer a single or at least a very limited predefined set of research questions. Methods of data re-utilization, like data mining and knowledge discovery have barely been discussed in corpus linguistics, although the use of machine-learning-based data-driven analysis seems reasonable given the new developments in data science and computational linguistics.

In this study I aim to shed some light on the potentials of data-driven analysis based on machine learning and predictive modelling for corpus linguistic studies. In particular, I investigate the possibility to repurpose existing language corpora for linguistic inquiry by using methodologies developed for data science and computational linguistics. The study focuses on predictive modelling and machine-learning-based data mining and gives a detailed overview and evaluation of currently popular strategies and methods for analysing language corpora with computational methods.

Part I introduces strategies and methods that have already been used on language data, discusses how they can assist corpus linguistic analysis and refers to available toolkits and software as well as to state-of-the-art research and further references whenever possible, to allow the reader to use this overview as an entry point for personal endeavours. It first introduces data mining, and text mining, as well as standard methods of predictive modelling and machine learning and proceeds by giving an interdisciplinary overview on existing strategies and previous work using predictive modelling and machine-learning-based analysis methods in applied linguistics. Following that, it discusses current needs for quantitative analysis in corpus linguistics and points out how computational methods can be used to assist the steps of a corpus linguistic study on existing data.

Part II evaluates the previously introduced methodological toolset by applying it in two differently-shaped corpus studies to two already existing, readily available German language corpora. The studies are based on computational linguistic tasks that have evolved over the last few years and are increasingly used also for more linguistic analyses[2]. While both studies contribute to German linguistics by giving new insights on the investigated phenomena, they are conceptualized to be realistic case studies for testing the methodological toolset introduced in part I, and thus allow to evaluate the potential benefits and challenges posed by data mining on language corpora.

The corpus studies are complementing each other in their research design. Corpus study one explores linguistic correlates of holistic text quality ratings on student essays and is conceptually similar to automated essay scoring or predicting language competence levels, both common tasks in computation linguistics. It introduces strategies that can be used for broad scale explorations a corpus by computational means starting from a simple naïve approach on machine-learning-based data-driven analysis to advanced methods for interpreting complex neural network models. Study two deals with age-related language features in computer-mediated communication and interprets age prediction models

[2] However, as with most methods developed in natural language processing, the previous research in the field mainly focused on English data and only few non-English resources can be found.

to answer a set of research questions that are based on previous research in the field. Compared to the first study, the analysis is more targeted and tests individual hypothesis, while attempting to account for methodological issues in corpus linguistic analyses.

The results of the case studies point out possibilities for using methods for machine-learning-based data-driven analysis on existing data resources and allow to critically discuss the generated outcomes in terms of the potentials, restrictions and the added value of these methods for corpus linguistics in part III, which concludes the thesis summarizing the outcomes of the first two parts and giving an outlook on future developments.

Part I

Theory and application:
Methods for data science in corpus linguistics

1 From existing data to data-driven analysis

Many scientific fields have changed substantially in the last few years. Some even speak of a paradigm shift in science that moves from theoretical research and the subsequent first revolutions coming with computational methods to *data-driven* or *exploratory science* (Hey et al., 2009; Kitchin, 2014). Many of these changes are relatable or even retraceable to the increased availability of data. *Big data* in the sense of huge volumes of fast-paced, extensible and still exhaustively describing data (cf. Kitchin, 2013) is particularly important in this context and has been frequently stated as being the main driver for new developments in artificial intelligence and machine learning (cf. Jordan & Mitchell, 2015). However, also well-curated sets of richly annotated, multi-faceted and diverse relational data have, in recent years, been taken into consideration and have been re-evaluated on their contribution to scientific progress[3]. The availability of massive amounts of new data or highly annotated data and their successful utilization in artificial intelligence applications like image recognition or self-driving cars have led to us to believe in the role of data as invaluable resource for the industry but also for research.

This can be seen by the growing importance of (research) data management (Corti et al., 2019; Wilkinson et al., 2016) but also by the fast and extensive development of new, powerful methods for the extraction of information from unstructured data or detection of useful knowledge in databases (i.e. *information extraction, knowledge discovery*), which serves the need to process and make sense of collected data.

Indeed, we can assign many of the new methods in statistics, artificial intelligence, machine learning and data mining to this expansion of data and to the increased computational power. Moreover, despite evolving more or less disjointly from each other in separate communities, the aforementioned disciplines overlap in terms of their abstract methodological and increasingly computational nature and the fact that they all deal with data. At the cross-section of these disciplines we find a set of computationally enhanced methods that allow researchers to make use of data resources and that can be applied in a variety of fields. In such a context, the term *data science* is frequently used to summarize the whole set of widely applicable new methodological options and the resulting advanced analytical processes that include strategies from statistics, data mining, text mining, knowledge discovery, information extraction, artificial intelligence and machine learning, information theory, and information visualization. It is, however, important to point out that many of the principles and components of data science have been "around for a long time" (Agarwal & Dhar, 2014), although sometimes in less refined forms, and that the main difference that makes the new term relevant (or justified) is the epistemological change: the fact that new types of research questions, new types of problems as well as new levels of scale, scope and precision that are significantly different from previous approaches can be tackled (Agarwal & Dhar, 2014; Dhar, 2013; Kitchin, 2014). This entails the observation of "micro, individual-level data on comprehensive scale" (Agarwal & Dhar, 2014); the integration of analytical and predictive systems in real life settings, allowing for applied research in many fields (e.g. Berland et al., 2014; Weir et al., 2016); and the scope of holistic analyses of data, e.g. by explaining multifactorial (causal) relationships that generalize over the given dataset.

In the following I will point out the interdisciplinary character of data science methods and their possible application to linguistics and give a concrete definition of the terms data science, data mining and text mining that are used throughout the thesis.

[3] To point out its value in opposition to big data, this type of data is also called "small" data by Kitchin and Lauriault (2015).

1.1 Data-driven analysis of textual data

As mentioned before, the methods used in data science – in the sense of data and computationally enhanced methods for scientific inquiry – are by no means exclusive to computer science or any other discipline. They can be applied to practically any field that deals with the empirical analysis of data, including the analysis of language corpora. Indeed, one of the major contributors to data-driven analysis methods at this moment are techniques built for language data, mostly coming from the fields of *natural language processing* and *computational linguistics*. In that context, strategies for the identification of patterns and interesting insights on textual data, such as data mining on already structured corpora or text mining on unstructured streams of text, have emerged. Such strategies might substantially change the way research approaches textual data, as can be seen by the many fields that have started to apply text mining or data science methods to make use of textual data, such as the social sciences and (digital) humanities, biomedicine, as well as business intelligence and marketing. However, in applied linguistics and in particular corpus linguistics, which by definition work with empirical language data, these methods can also be utilized, helping researchers to analyse existing research data and more specifically language corpora with the toolset provided.

1.2 Data science

The term *data science* was originally proposed as a new name for both computer science (information science) (Peter Naur in 1960[4]) and statistics (Jeff Wu in 1997), referring in both instances to the changes those fields experienced with the increased availability of data. This fact illustrates that the term is essentially just an evolution of existing fields, bringing two different communities together that are very different in philosophies and aims, however much they are related by their approaches and both contribute individually to the methodological toolset of data science.

Data science means in general any sort of analytics done by researchers or practitioners that uses '*big data*' to draw conclusions (Agarwal & Dhar, 2014; Dhar, 2013). It encompasses all (computational) techniques that can be used to do state-of-the-art research based on or driven by available empirical data like unstructured or linguistically annotated, and thus structured, language corpora. It thus addresses issues of research design and problem definition; the selection and retrieval of variables to analyse and their preparation and subsetting; analysis techniques like *text classification*, *regression modelling*, *clustering* and *outlier detection*; as well as the interpretation of all preceding steps to draw insights from the data or to reformulate the research question and refine the process.

Hence, data science can be defined as the aim to address complex research questions by using computational methods

- to retrieve, prepare, and filter related variables,
- to analyse them with advanced statistical models,
- to interpret the resulting models and
- to repeat and refine the previous steps

in order to get relevant insights from the data that confirm existing or generate new hypotheses.

[4] Later published in Naur (1966, 1974).

1.3 Data mining

Data mining as a term is closely related to data science. However, the two terms are not completely synonymous. Data mining was originally a denominator for one of the steps of the KDD (*knowledge discovery in databases*) process (Dunham, 2006; Fayyad et al., 1996) and was only attempted by a limited group of computer scientists. However, it has since become a well-known concept that refers to the process of finding useful, previously unknown patterns and relationships in datasets (Witten et al., 2016) and is used in many application scenarios such as *business intelligence, educational data mining*, *social media analysis* and others[5]. Hence, it stands for the general act of *information extraction* or *knowledge discovery* and is used by data scientists to get new insights from data. The instruments used to extract these insights are manifold and reach from standard statistical measures, from *pattern mining algorithms* (e.g. Agrawal & Srikant, 1995) and *information extraction techniques* (Sarawagi, 2008) to *machine learning* and *artificial intelligence* (Han et al., 2012). Even within these frameworks, various adaptations and implementations provide a big variety of techniques to somehow extract the supposedly hidden "research treasures" in society's continuously growing data repositories. While researchers or decision-makers from many different disciplines show interest in applying data science methods in their own fields, computer science itself does its best to further advance and refine the methodological toolset for special cases.

1.4 Text mining

Many applications of data mining nowadays are not exclusively based on structured data from databases (e.g. business transactions) but they also make use of unstructured data such as textual data, i.e. text mining[6]. *Text (data) mining* (Hearst, 1999) can be seen as a special case of data mining, where textual data is used instead of structured databases (Feldman & Sanger, 2007). This introduces various difficulties for analysis that originate in the multi-layered and ambiguous character of language. Originally, the term *text mining* was used for the analysis of completely unstructured streams of texts (Feldman & Sanger, 2007; Hearst, 2003; Kroeze et al., 2003). However, nowadays also semi-structured language resources are available for example through the use of pre-annotated corpora or the use of social media data that often offer information on the users, publishing time or user reactions. Next to computer science and computational linguistics, the main driving forces in data mining on textual data often come from business intelligence, public security and forensics as well as from biomedical sciences. Their aim is to find new market possibilities, monitor customer opinions and competitors (Chen et al., 2012), identify criminal behaviour (Coulthard et al., 2016) or find new hypotheses from existing medical literature or patient files in order to avoid costly and or risk-related primary studies (cf. Holzinger et al., 2014; Lamurias & Couto, 2019; Rinaldi et al., 2007). However, also the social sciences and humanities show increasing interest in text mining methodologies and are contributing with new summarization and visualization techniques for distant and blended reading (e.g. Jänicke et al., 2017; Liu et al., 2019; Stulpe & Lemke, 2016).

In this thesis the term *data science* is used for whenever the interdisciplinary use of computational techniques for any step in the analysis of empirical data is meant, including the selection and preparation of data, monofactorial and multifactorial, statistical and/or visual methods for data investigation, exploration and hypothesis testing as well as their interpretation and further refinement. *Data*

[5] See for example Baker and Inventado (2014), Chen et al. (2012) or He et al. (2013).

[6] While structured data is available in well-defined databases, usually having a limited amount of possible values for each attribute, unstructured data does not provide any additional information (e.g. metadata), possible categorizations or predefined possibilities of which values occur in the data.

mining, on the other hand, is used for one of the tasks in data science, namely the act of finding interesting new patterns in partly structured corpus data through computational methods (i.e. predictive modelling and machine learning). The term *text mining* is used, when the data to extract patterns from is raw, unstructured text (e.g. in the case of feature extraction methods see section 4.3.2.1).

2 Corpus linguistics

Corpus linguistics is the empirical study of language through the use of real-life examples of naturally occurring (spoken or written) language (McEnery & Wilson, 2001). After initial controversies in the early 2000s on whether corpus linguistics should be seen as a discipline in itself or as a methodological framework (Tognini-Bonelli, 2001), the corpus linguistic approach, and more generally the use of language corpora for linguistic inquiry, has spread widely over many domains that deal with natural language. Corpus-based research is, according to Laurence Anthony (Anthony, 2013), "one of the dominant methods to study language" today and is used for lexicography and lexical studies, grammatical studies, register studies and genre analysis, studies on dialects and language varieties, contrastive and translation studies, diachronic studies of language change, language learning and teaching, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and literary studies as well as forensic linguistics (McEnery et al., 2006).

Most generally, a *corpus* can be defined as a "body of naturally occurring language" (McEnery et al., 2006). However, there are various more refined definitions that highlight certain aspects of corpora as we know them today. An important aspect is made clear in Tognini-Bonelli's definition(2001) when she states that a corpus is "a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis". Kilgarriff and Grefenstette (2003), state that a corpus is "a collection of texts when considered as an object of language or literary study". The definitions point out that a corpus usually contains more than one text and a "body of text" is usually only called a corpus, when it is used to study its language. All these definitions hint to the fact that corpora need to fulfil a number of characteristics in order to be suited for linguistic or literary research. They should be *representative* for a particular language, variety or register in order to claim that research results are valid for the whole of the language, variety or register instead of only describing the dataset, i.e. to facilitate *generalizability* (cf. Biber, 1993a). Hence, corpora also need to be balanced and relatively big. It requires corpora to be machine-readable, to facilitate fast access to the data, including possibilities for automation of annotation and analysis processes.

However, although there is a general consensus in corpus linguistics about the necessity of the aforementioned corpus design criteria (Biber & Reppen, 2015; Gries & Newman, 2014; Lüdeling & Kytö, 2008; McEnery et al., 2006), many corpora differ from an ideal design (Gries & Newman, 2014). The collection of perfectly balanced, large and representative corpora is not always feasible in practice and, with the increasing complexity of linguistic studies, not even theoretically possible. Instead, the more pragmatic stance of the corpus having to match the research goal is often taken (Gries & Newman, 2014).

We can see a multitude of different corpora and different approaches to corpus linguistics. There are *general-purpose corpora* made with the aim of being representative of a whole language allowing many different analyses on the same data, and corpora specifically created and designed for a particular research question. Researchers pursue *quantitative* as well as *qualitative* research on corpora. Increasingly often, studies try to combine both methods in a mixed approach (Dörnyei, 2009; Hashemi

& Babaii, 2013). *Corpus-based* studies approach the data top-down, finding examples or counterexamples for known linguistic theories (McEnery et al., 2006), while *corpus-driven* approaches aim to extract theories from the data bottom-up, without considering prior linguistic knowledge (Tognini-Bonelli, 2001). In addition, a third methodological approach, the so-called *data-driven* approach (Bubenhofer, 2009; Rayson, 2002), aims at evolving theory iteratively through the empirical study of corpora, by claiming that no purely corpus-driven induction without intuitions from theory is possible (see also Kitchin, 2014), thereby stepping away from the strong division of both worlds.[7] This approach to quantitative corpus analysis is adopted in this thesis.

2.1 The stages of a corpus linguistic study

No matter if corpus-driven, corpus-based or data-driven, no matter if quantitative or qualitative, there are in general three different steps or phases in the corpus linguistic workflow: 1) the phase of *corpus design*, 2) the phase of *corpus creation* including *data collection*, *annotation* and *corpus construction* and 3) the *corpus analysis* and interpretation phase. The type of research approach and which research questions one is able to tackle in a corpus linguistic study depends on the way these steps are executed, combined and repeated.[8]

Corpus design
Before the construction of a new corpus, the purpose, collection methods and sampling criteria are defined in a detailed corpus design phase. The range of research questions to study on the corpus as well as the target language or variety are detailed in this step and provide the basis for the subsequent corpus creation (see also Biber, 1993a; Hunston, 2009; Schäfer et al., 2013).

Corpus creation
The actual corpus building or creation can be further divided into corpus collection and corpus construction and annotation. *Corpus collection* or *compilation* means the actual gathering of the text or language data and comprises of steps such as recruiting and contacting text or speech data donors, transcribing spoken or handwritten material, scanning or digitalizing non-machine-readable materials or scraping from online sources, as well as corpus sampling. *Corpus annotation and construction* on the other hand is the phase where collected materials are cleaned, processed and annotated, linked together and made accessible for later analysis. It comprises all manual, semi-automatic and automatic methods taken to "identify textual and linguistic characteristics" (Rayson, 2008a) on lexical, syntactic, morphologic, pragmatic, semantic etc. levels and usually also entails the provision of automatic means to query the corpus and retrieve occurrences or text samples.

Corpus analysis and interpretation
The final step deals with the actual investigation of the corpus data and the linguistic phenomena displayed in it. *Corpus analysis and interpretation* aims at answering research questions by testing linguistic theories or generating new ones by exploring the data. Along with methods for retrieving frequency data, concordance lines and co-occurrence data (see below), statistical methods and visualizations are employed for inference and interpretation.

[7] We can also see this trend towards more integrated exploratory and confirmatory research designs in other scientific fields aiming for abductive research that iteratively evolve on hypothesis building and testing cycles (Haig, 2018; Kitchin, 2014).
[8] Rayson (Rayson, 2008b), for example, gives a good overview on the various process models in the corpus linguistic workflow.

Although each of these steps is equally important in corpus linguistic studies, this thesis focuses solely on the step of corpus analysis and interpretation, investigating and evaluating computational methods that can assist the process of inferring linguistic insights from empirical language data stored in language corpora in a data-driven, mainly quantitative fashion. The thesis thus refrains from discussing matters of corpus design, collection, compilation or annotation but focuses on those methods that can be used for linguistic inquiry on available data. The remainder of this section describes traditional corpus linguistic methods, names the main statistical methods used in quantitative corpus analysis and discusses the status quo alongside current needs in quantitative corpus analysis, as well as the potential added value of computational methods from data science and data mining for dealing with these needs.

2.2 Methods in quantitative corpus analysis

2.2.1 Traditional corpus linguistic methods: Frequency lists, concordances and collocations

The main investigative methods in corpus analysis build upon the observation of linguistic contexts, frequencies of occurrence and frequencies of co-occurrence of words, word patterns or other linguistic units. As such, a first step of any corpus linguistic analysis is to extract frequency lists, so-called concordances or collocations (colligations/collostructions) for a given corpus.

Context: Concordances
Concordances show occurrences of a word (or linguistic pattern) of interest within the context in which it occurs in the corpus. The typical format for such concordances is the Keyword-in-Context (KWIC) format. It shows each occurrence of the searched item with a specified number of preceding and succeeding words in one line. Concordance lines are usually used for qualitative exploratory analysis. However, they can also be the basis for quantitative investigations including data mining approaches (as, for example, in Pölitz, 2016b).

Occurrence: Frequency lists
Probably the most common corpus linguistic method used in quantitative studies is to extract frequencies of occurrences. This can concern the frequency of words, bundles of one or more words in a sequence (n-grams), keywords or any other linguistic unit or phenomenon present in the corpus. Frequency lists can furthermore display all words (word forms or lemmas, part-of-speech categories, etc.) of a corpus or just a custom list of items of interest (e.g. the variants of a linguistic category). Frequency lists are often the basis for further quantitative analysis, e.g. comparing frequencies for different (sub)corpora, but they are also used to describe the corpus qualitatively.

Co-occurrence: Collocations, collostructions, colligations
Closely related to both previous methods is the analysis of frequently co-occurring words or linguistic structures using statistical association measures like log-likelihood or mutual information. While the term *collocation* is usually only used for co-occurring lexical items, *colligation* or *collostructions* refer to lexico-grammatical co-occurrences between words and grammatical patterns or constructions (cf. Gries & Durrant, Forthc.; McEnery et al., 2006). Co-occurrence analyses are frequently used for lexicographic, didactic and contrastive purposes. They allow for the extraction of word profiles and for pragmatic investigations to be conducted, as well as for the identification of idiomatic formulaic sequences.

These three methods of retrieving information from corpora are the basis for most corpus linguistic studies, both qualitative and quantitative. However, while concordance lines are prevalently used in

qualitative studies, frequencies and collocations are the main basis for quantitative corpus analysis and are used to compare (sub)corpora and perform inferential statistics.

2.2.2 Statistical methods of inference in quantitative corpus analysis

Looking at the nature of the data used in corpus linguistics (i.e. frequency lists, collocations) it becomes clear that in many cases corpus linguistics is seen as an "inherently distributional discipline" (Gries & Newman, 2014) that relies on statistical methods to analyse the (relative) frequencies of occurrence, (relative) frequencies of co-occurrence and (relative) dispersion of linguistic phenomena (Paquot & Plonsky, 2017).

Summarizing all the statistical methods used in corpus linguistics would exceed the scope of this work. However, the list below shows examples of common methods in quantitative corpus linguistic studies which are also frequently discussed in introductory works on statistics in corpus linguistics (Biber & Jones, 2009; Desagulier, 2017; Gries, 2009, 2013; Johnson, 2011; Moisl, 2015; Oakes, 2005) and used in many studies. Those are:

- comparisons of central tendencies including statistical significance tests to control for group differences[9];
- calculation of monofactorial correlations (between numeric variables) and associations (for categorical variables)[10];
- simple and multiple linear regression models that can predict outcomes for new observations using one or more input variables[11]; and
- clustering techniques that allow observations or linguistic phenomena to be grouped based on the data[12].

In general, most analyses aim to test if there is any *relationship* between a concept of interest (i.e. the *dependent variable*) and linguistic or non-linguistic features (i.e. *predictor variables*) that are assumed to be relevant (e.g. because there are theoretical accounts for them)[13]. An existing relationship does not necessarily imply that one variable is responsible for the value of the other (i.e. *causal relationship* or *dependence* between the two variables). The relationship can also be of *symmetric* nature, i.e. a co-occurrence that is caused by either the *effect* of the predictor on the dependent variable or also by the effect of another, possibly unknown factor that influences both the predictor and the descriptor variable (i.e. a *confounder*).

Symmetric relationships or co-occurrences are called *correlation* (if the related variables are numeric or at least ordinal) or *association* (if they are categorical). They are typically calculated between two variables in a *bivariate* or *monofactorial analysis*, using a correlation (e.g. *Pearson-product-moment correlation* or *Spearman rank correlation*) or association measure (e.g. *Chi-Square test of*

[9] Comparing the frequency of occurrence of a linguistic phenomenon between different groups of writers is a typical example of such an approach.
[10] An example of a correlation could be to test whether the frequency of productive use of a lexical item increases with exposure to it. An example of an association could be to test whether the presence of a phenomenon is more frequent in one type of text than in others.
[11] Regression models are used for example to predict the sales of a book based on its language, thereby identifying also the factors that influence sales.
[12] E.g. grouping types of text based on their vocabulary.
[13] For example, it could be that one variable is higher when another variable is higher as well, or that one group has more occurrences of a certain type than another.

independence, *Log-likelihood test*) while controlling if there is enough evidence to consider that the variables are related[14] (i.e. *significance testing*)[15].

In the presence of known confounders, other already known factors/predictors, or with an a-priori complex research design, however, statistical models like regression models or machine-learning-based predictive models can be used to account for the needs of multifactorial analysis. Using predictive modelling, which (possibly wrongly) implies a causal relationship, *combinatorial effects* of the predictor variables as well as confounding effects between them can be taken into account. This makes it possible to evaluate the joint effect of variables on the dependent variable, and to single out the effect of one variable when possible confounders are not held constant a priori. It furthermore allows for the detection of *interactions* between the variables, which indicate combinatorial effects, e.g. one predictor enhances or reverts the effect of another predictor on the output variable.

The use of multifactorial methods that are based on predictive modelling and machine learning is, however, not yet common in corpus linguistics. Some studies still use frequency lists, concordances and collocations even without any statistical test or control mechanism, although most corpus linguistic research nowadays includes at least simple statistical comparisons and monofactorial analysis. Only recently are a number of researchers also experimenting with advanced regression models like *generalized (additive) mixed-effects models*, or *naïve discrimination learning* inspired by psycholinguistic or cognitive linguistic research; fields that are closer to traditional statistics. Others venture to explore the interpretability of complex machine learning models (often defined as black box models) that have become popular in natural language processing, data mining and artificial intelligence[16]. In general, the aim to conduct methodologically correct and timely research as well as the aim to make use of existing corpus data advocate the further exploration of predictive modelling and machine learning for quantitative corpus analysis.

3 Predictive modelling and machine learning for data science

> "Many of the tools used to perform data mining are standard statistical methods that have been around for decades […] However, data mining also includes a wide range of other techniques for analysing data that grew out of research into artificial intelligence (machine learning), evolutionary computing and game theory" (Finlay, 2014)

In general, most often when we talk about data mining and data science, we talk about predictive modelling or about building a model. A *model* can be described in broad terms as an abstract, reduced representation of the world, of a system or of a given set of data. While *association-based statistical models* present the relationships in the data in a parsimonious but precise way (cf. Breiman, 2001b; Shmueli, 2010), *predictive models* focus on those relationships that allow generalization and

[14] In general this is done by falsifying that they are unrelated, defining the Null-Hypothesis (H0) so that it tests the case of unrelated variables and stating that there is a significant deviance from the expected values for unrelated samples to assume that they are not unrelated (and thus, according to logic, related).
[15] See the first two methods listed above.
[16] Indeed there are specifically developed software tools for statistical analysis, text or data mining, visual analysis methods and interpretation methods for corpus linguistics (e.g. Kupietz et al., 2018; Periñán-pascual, 2017; Pölitz, 2016b; Rayson et al., 2017; Siirtola et al., 2014).

abstraction from the actual data, and that predict outcomes for new (similar) data given the rules or structures learned from the input data.

A predictive model is thus "any method that produces predictions, regardless of its underlying approach: Bayesian or frequentist [statistics], parametric or non-parametric, data mining algorithm or statistical model, etc." (Shmueli, 2010).

3.1 Machine learning

Although predictive modelling originates in statistics as the "earliest and most prevalent form of statistical inference" (Geisser, 2017), the field then advanced, in particular, through machine learning research that has stronger ties to computer science. *Machine learning* is the term and the corresponding research field for using computational methods to build statistical models. The "machine*"* (i.e. computer) provides the power, memory and precision to automate complex calculations that would not be feasible with human means. The machine "learns" to predict outcome values or group affiliations through the experience it gained from the data (i.e. the resulting predictive model). In fact, machine learning is not one method, but a multitude of methods, such as *linear* or *logistic regression*, *Naïve Bayes classification*, *decision trees*, *rule learners*, memory-based learning, *support vector machines*, *ensemble learners* or *neural networks* and numerous derivations and adaptations of these base methods. While we also count *clustering* methods and *anomaly detection* methods as machine learning (see section 3.4.1)[17], predictive modelling usually implies supervised machine learning methods for *classification* (the prediction of a target class) and *regression* (the prediction of continuous values).

As machine learning was soon taken over by computer science and valued especially for its practical relevance in applied prediction and automatization scenarios, its original connection to statistical modelling and data analysis is almost lost in modern definitions of the term. Mohri et al. for example define machine learning as: "computational methods using experience to improve performance or to make accurate predictions". Van den Bosch says in his introduction on machine learning for corpus linguistics that the "prime goal of machine learning is to develop automatic learning algorithms by which a computer can learn to perform real-world tasks, not by being told (programmed) beforehand how the problem is solved, but by discovering the solution on the basis of examples" (Van den Bosch, 2009). These definitions focus strongly on predictive accuracy and the application of machine learning to solve practical tasks, but do not necessarily imply that analysis could be based on the models.[18]

3.2 Predictive vs. descriptive vs. explanatory modelling

Not all potentially *predictive* models are built with the intention to predict. Witten et al. (2016) emphasize this aspect, when they point out that machine learning is "the technical basis of data mining" and can be understood as "the acquisition of structural descriptions from examples." They

[17] These two are usually referred to as unsupervised learning.

[18] However, recent literature in statistics (in addition to data mining and data science) propagates the possibility to use predictive models and thus machine learning for data analysis. Hastie et al. (2011) and James et al. (2013) use the term *statistical learning* to emphasize the scientific, statistical use of learning algorithms. The definition found in the famous introductory book on machine learning and data mining by Witten et al. (2016) also mentions the value of machine learning for scientific analysis when saying: "We interpret machine learning as the acquisition of structural descriptions from examples. The kind of descriptions found can be used for prediction, explanation, and understanding." The main part of machine learning research over the last decades, however, focuses on prediction and application.

furthermore say that the "kind of descriptions found can be used for prediction, explanation, and understanding."

In general, there is a methodological difference between *predictive modelling*, *descriptive modelling* and *explanatory modelling* that is still under debate in statistics (cf. Shmueli, 2010).

Classical statistical models like linear and logistic regression are applied in many research fields for so-called *descriptive or explanatory modelling* (e.g. in psychology, sociology or biomedicine*)*. Such studies use the statistical models for data description and causal explanation.

The term d*escriptive modelling* is used when the goal is to build a precise model of the data that describes associations in order to represent the data in a reduced format. This *data modelling culture* (Breiman, 2001b) is the most common form of statistical analysis performed by statisticians according to Breiman (2001b) and Shmueli (2010).

In other scientific fields (e.g. the social sciences) we can observe the "almost exclusive" use of statistical models for causal explanation (Shmueli, 2010). These *explanatory models* assume (based on strong suggestions from theory) a causal relationship between the *predictor* variables and the *response*. Consequently, we can define explanatory modelling as testing specific causal hypotheses about theoretical constructs[19] by using association-based statistical models on observational data (Shmueli, 2010). Although this approach is controversial (for various reasons including frequently unmet assumptions or lacking model validity), in practice it is the most common approach to quantitative analysis.[20]

Although the models would be suited for it, neither descriptive modelling nor explanatory modelling necessarily aim to predict the outcome for new observations. For both descriptive models and explanatory models, the performance is measured on the exact same data that was used for building the model (henceforth also *training data*)[21]. The strict definition of the term *predictive modelling*, on the contrary, is used whenever a (predictive) statistical model is applied to data in order to predict the outcome for new (at least slightly different) data. This assumes generalizability of the model to further cases and can give further relevance to the relationships found, but it also shifts the perspective and the modelling goals. The main values of interest in predictive modelling are the input and the output (as well as the error made during prediction). The statistical model in between those values is treated as a *black box* that does not need to disclose the relationships it learned, contributions of individual or combined variables and special cases (see Figure 2). This makes it possible use more complex and performant algorithmic methods that do not need to be interpretable for human beings.

[19] i.e. an abstraction that describes "a phenomenon of theoretical interest" (Edwards & Bagozzi, 2000)
[20] Although there would be statistically more solid methods like randomized controlled trials (cf. Pearl, 2009)
[21] E.g. with goodness-of-fit tests and residual examination

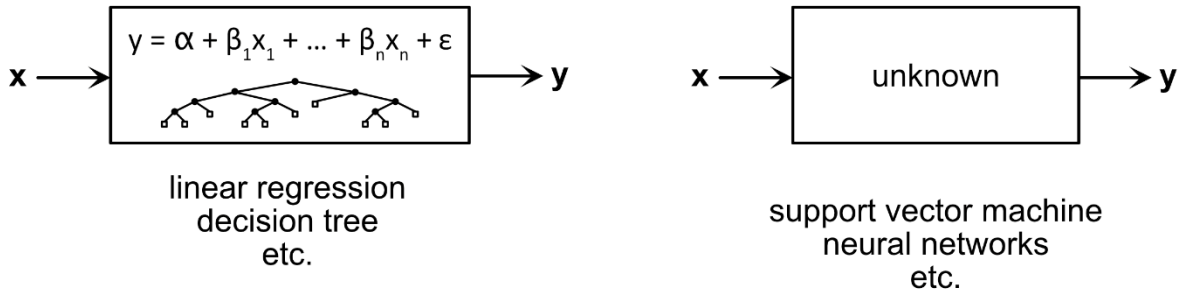


Figure 2: Two schematic illustrations of modelling approaches. Left: The model is intrinsically explanatory. It's input-output mappings are transparent and can be interpreted directly. Right: The model is (treated as) a black box. What happens between input and output is not interpretable or at least not taken into consideration (Figure adapted from Breimann, 2001)

Instead of *goodness-of-fit* tests and *F-tests* that show how much of the variance in the response variable can be explained by the predictor variables (explanatory power) and in-depth interpretation of regression coefficients and effect sizes, the model performance in predictive modelling is evaluated through prediction accuracy (or similar measures for predictive power) on new, previously unseen data (cf. section 4.3.3.2). Hence *generalizability* is an important issue that is addressed by making sure that the model is neither too specific nor too reduced but still presents the necessary structures that help to predict a construct of interest.

This slight difference in perspective between descriptive/explanatory modelling and predictive modelling changed the focus from simple and interpretable models (that are in turn rigid, less powerful in the sense of how well they can extract relationships from the data, and have many assumptions) to the application of completely inscrutable systems (that achieve high prediction accuracies). This disparity that has grown over the last decades, has, however, been questioned in recent years. On one side, interpretable, explanatory models are often not sufficient to model complex problems in explanatory studies as they do not manage to find structure in the data. On the other side, society and the scientific community ask for explainability and interpretability of complex models or predictive systems that are deployed in real life (Gilpin et al., 2018; Honegger, 2018; Shmueli, 2010) and the artificial intelligence and machine learning community recognizes the need for interpretability in order to facilitate debugging and further model improvements (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016b).

The emerging fields of interpretable machine learning and explainable AI contribute to making the explanation and interpretation of complex predictive models feasible (Gilpin et al., 2018; Miller, 2018). Through these new developments, predictive modelling in general and machine learning methods in particular can be applied for data analysis (i.e. data mining) in order to explore datasets, generate or test hypotheses about the data and its inherent relations and thus derive relevant insights for research. By using knowledge about

a) more interpretable methods like regression models, decision trees, rule learners or Naïve Bayes classifiers and
b) post-hoc black box interpretation methods for complex models,

predictive models can furthermore give insights not only into the generalizability of the relationships found, but also into the importance of individual variables for the learned class or outcome, on

individual and combinatorial variable effects (both direction and effect size), and existing variable interactions. [22]

This combination of predictive modelling for explanatory purposes, however, needs a set of skills and knowledge about the underlying logic of different machine learning methods as well as of principal concepts and typologies of machine learning which are introduced in the following sections. I first introduce major concepts and terminology in the context of machine learning and predictive modelling. After different typologies of methods are introduced, a brief description of common machine learning methods, their underlying logic, as well as examples for their use in linguistic research and their interpretation is given. Finally, this section also discusses the interpretability of machine learning models, giving references to recent research fields that contribute to making predictive modelling relevant in data science and thus for corpus linguistics.

3.3 Concepts and terminology

There is a substantial overlap between the term predictive modelling and other terms such as statistical modelling, statistical learning, and machine learning.[23] While the terms could be used interchangeably, *statistical modelling* and *statistical learning* are more related to quantitative statistical analysis. The former often refers to simpler, traditional models like linear and logistic regression. The latter enhances the relevance of more complex, algorithmic methods often connected with classical machine learning. *Predictive modelling* and *machine learning* on the other hand focus on the prediction step and are naturally connected to theoretical and applied computer science. However, predictive modelling is also often used to describe data mining endeavours and consequently more related to learning tasks that want to draw new insights from data.

The act of modelling, more precisely the creation of a model, is called *model building*, *training* or *fitting*. Generally speaking, it means to apply a learning algorithm to a dataset. While *model fitting* is the main term used in statistics (see also the evaluation criterion '*goodness-of-fit*' below) *training* often refers to "teaching" a machine learning algorithm. *Model building* is the most generic denominator.

The output is, in general, a *model*, an abstraction of the data that is able to summarize the data and generalize from it. The abstract (or in the case of intelligent, applied systems, deliverable) product that we obtain through a modelling process is, however, also often called a *predictive system* or, in the case of classification, a *classifier*.

Depending on the type of learning algorithm, the model can be used to produce summaries of the data or to generate predictions for new, unseen data. There is, however, a high redundancy in terms used to refer to elements of the input data, i.e. the *training set* and, in the case of prediction the possible prediction output. Observation, data point, example, case, or row refer to one instance in the dataset, while feature, variable, attribute, independent variable, predictor or predictor variable, factor, X or column are names for the values used to describe the instance. The concept or construct being predicted can be named: class, target, dependent variable, descriptor or descriptor variable, response or response variable, outcome, y, and label. It can be either a categorical (two or more levels) or a numerical value.

[22] An added value that was previously only assigned to explanatory modelling.

[23] For the sake of completeness, artificial intelligence should also be stated here. Although we did not touch on this term yet, it is strongly related to machine learning, often even used interchangeably. However, it is common to use the term artificial intelligence mainly when a broader, more abstract goal of building intelligent machines for example through machine learning is meant (e.g. robotics, self-driving cars, natural language understanding).

There is also some confusion about how to refer to the machine learning method or algorithm chosen to train or build a model, as they are often also called model or classifier but do not refer to the abstract outcome. Instead they refer to the mathematical or algorithmic rules used to obtain this outcome. Therefore, *learning algorithm*, *machine learning algorithm*, *method* or *technique* are often preferred. Besides, r*egressor* and *classifier* are used in order to refer to classification or regression algorithms. For clarity, in this work the term learning algorithm is used to refer to the machine learning method employed.

Once a model has been trained, its relevance can be evaluated and compared by measuring its *predictive power*, i.e. whether and how well it can predict values or class labels for new data and therefore be used for generalizations, or its *explanatory power*, i.e. whether and how well it can be used to explain the target variable. The model's predictive power, or "performance", is evaluated by predicting outcomes for previously unknown data from a held-out test set or with cross-validation.[24] Explanatory power on the other hand is established by goodness-of-fit and interpretability of the model, including whether one can estimate variable importance, variable response, individual and combinatorial variable effects and effect sizes.

Depending on the aim of the study, the selected model can be that with the highest predictive power (most generalizability), the simplest model structure (cf. Occam's razor) or the highest explanatory value (highest R^2). However, most studies make a trade-off between prediction performance (e.g. the percentage of correctly predicted instances, also called *accuracy*), model *complexity* (how many probably redundant or unnecessary predictor variables it contains, how many variable transformations are done, how many classes are predicted, etc.), and *interpretability* (how can the model be used for explanation and interpretation of the inherent structures).

Every learning algorithm has to find a good balance between generalizing over the dataset and accounting for special cases in the dataset in order to create a model that cannot only be applied to the very same data but is useful also for other data. In general, one wants the learning algorithm to abstract from the given data and find generalizations that apply also to slightly different data points. But of course, the more general a model is, the less it will be able to capture the fine distinctions in the data. Hence, any kind of exception from the general rule will not be accounted for. Such an *over-generalization* is called *underfitting*, as the generated model barely fits (i.e. represents the depth of information provided in) the training data. A model that fits the training data too well on the other side will only be able to deal with the training data but will fail as soon as any deviation occurs. This is called *overfitting* or *under-generalization* and brings equally bad results in prediction task on new data.

The dilemma of finding the optimal balance between generalization and presenting the information as fine-grained as possible is often also called the *bias-variance trade-off* (Alpaydin, 2014), where *bias* refers to overfitting to the training data and *variance* refers to the potential noisy groups in very general models. Usually the best bias-variance trade-off is estimated by investigating the predictive performance of a model. A model with a high predictive performance is supposed to be better than one with a lower performance. However, in any case a certain threshold of predictive performance (i.e. the *baseline*) has to be exceeded in order to claim that the model is able to generalize at all. This is an important validation criterion as any predictive model can also give random predictions if there is not enough structure to learn from the data.

A general problem of explanatory or predictive modelling is also *the multitude of possible models* (McCullagh, 2018). No matter which concept we want to model, there is usually more than one way

[24] For detailed descriptions see section 4.3.3.2.

to build the model[25]. Different operationalizations of the concept, different predictor variables or different representations of those variables, or different modelling methods might lead to similarly good models. However, there are some strategies that we can apply to make informed decisions about which models we take for analysis and interpretation that vary depending on our needs and priorities. After various competing models have been built, *model selection*, *model comparison* or *classifier comparison techniques* help to interpret the differences between the models and/or choose the final model that will then be interpreted in detail.

The interpretation of the final model depends on the learning algorithm used. *Intrinsically interpretable methods* like *linear* or *logistic regression, decision trees* or *rule induction methods* create models that can be inspected without further interpretation methods just by looking at the symbolic representations of the model internals. For *black box models,* additional helper methods are needed for interpretation. However, the model can be uninterpretable, even if the method used to build it is in general intrinsically interpretable. This can happen if the actual model is too complex (e.g. has too many possible predictor variables, too many levels of the target variable or uses non-linear transformations for variables or learning algorithms. Recent research on the *explainability* or *interpretability* of complex and/or black box models, however, is developing methods and strategies for *model-agnostic* and *model-specific* explanation and interpretation of models (cf. Guidotti et al., 2019; Linzen et al., 2019; Molnar, 2018b).

To sum up, m*odelling* in general comprises all steps from defining the actual problem, retrieving and preparing the necessary data and variables, training one or more predictive models and selecting or at least comparing the models in terms of some evaluation criterion.

3.4 Machine learning typologies

Machine learning methods can be categorized on different dimensions. Apart from the notion of supervised, unsupervised and semi-supervised methods, we differentiate depending on the relationships learned (e.g. linear vs. non-linear), the assumptions on the data (e.g. parametric vs. non-parametric), the time of learning (eager learners vs. lazy learners) and the degree of abstraction in the model (instance-based or memory-based learning vs. model-based learning, deep learning and neural networks), and the way decision boundaries are drawn or the inherent interpretability of a trained model (e.g. intrinsically interpretable models vs. black-box models, symbolic vs. non-symbolic models). In the following a selection of frequently referred to typologies is introduced in order to build the theoretical basis for the description of methods in section 3.5.

3.4.1 Supervised, unsupervised and semi-supervised

Machine learning methods can be roughly divided into supervised, unsupervised and semi-supervised methods.

In *supervised methods* the algorithm learns how to predict new (target) values after it has trained on a set of instances where the correct answers (target information) were given; it learns on so-called labelled data. Depending on the type of the target value, this is either called ***classification*** (when categorical class labels are given) or ***regression*** (when the output variable is a continuous value that needs to be estimated or predicted).

[25] Breiman (2001b) calls this the *Rashomon effect*.

In *unsupervised machine learning*, no correct output labels are provided. Unsupervised methods aim to report on the structure of the data without prior knowledge of its potential meaning, so the training data is entirely unlabelled. Methods in unsupervised machine learning try to cluster the data based on its features (***clustering***) or detect anomalies in the data that lead to interesting or potentially problematic instances (***anomaly detection***).

Semi-supervised methods are located between supervised and unsupervised methods and make use of (a small set of) labelled and (a large set of) unlabelled data during training.

3.4.2 Regression vs. classification problems

As stated above, predictive models in supervised machine learning can model either classification or regression problems, depending on the type of variable they predict.

Regression problems are supervised machine learning tasks that try to predict a continuous, numerical outcome for a new data point given the relationships and interactions of variables and their outcomes in the training data. Usually regression problems are modelled with some variant of the linear regression method (see section 3.5.1). However, there are also more algorithmic methods for regression that are often adapted versions of algorithms for classification (e.g. support vector regression, neural network regression and regression trees or random forest regression).

A *classification* problem on the other hand is the supervised machine learning task of predicting one of two (binary classification) or more (multi-class classification) categorical class levels based on a given set of variables for a new observation. Almost all learning algorithms are suited for binary classification[26], but workarounds are often needed for multi-class classification (e.g. training individual models for each level and then choosing the class with the highest probability for the final prediction as for multi-class support vector machines or logistic regression).

3.4.3 Linear vs. non-linear

Models can be divided into *linear* and *non-linear models* depending on the complexity of the underlying mathematical function that is used to describe the data. Linear models apply linear functions on the feature values in order to predict the class label, e.g. *linear regression* or *linear kernels of support vector machine*s. Non-linear functions can also separate data that is not linearly separable. Examples for such non-linear models are *support vector machines* with non-linear kernel functions, non-linear *neural networks* or *polynomial regression* (James et al., 2013).

3.4.4 Parametric methods vs. non-parametric methods

Parametric methods such as *logistic regression* (section 3.5.2), *Naïve Bayes classification* (section 3.5.5), or *simple neural networks* (section 3.5.9) simplify the learning function by making assumptions about the data (e.g. that it is distributed normally and that data points are independent). Because of these assumptions, algorithms can be easier to understand and interpret, are faster, and need less data for training. However, this complexity reduction can only be applied when all the assumptions are met. Otherwise, *non-parametric* methods should be chosen, which do not make strict assumptions about the form of the mapping function or about the data. Non-parametric methods such as the *k-nearest neighbour algorithm* (section 3.5.6), *decision tree algorithms* (section 3.5.3), or *support vector machines* (section 3.5.7) usually perform better when big amounts of training data are available and

[26] An exception are linear regression models that need a logistic transformation in order to predict classes instead of numerical values (see section 3.5.2).

there is no prior knowledge or no time or resources to optimize the algorithm and choose the right features. This comes with the downside of reduced interpretability, slower training times and the risk of overfitting the data. While parametric functions usually only perform well for simple, well-defined problems, non-parametric methods offer possibilities to deal with more complex problems.

3.4.5 Instance-based learning vs. model-based learning

While *model-based learning* builds an abstract model (a *mapping function*) of the data on the basis of the whole training set and then predicts only on the basis of that model (e.g. support vector machines, Naïve Bayes classifiers or decision trees), *instance-based learning* (also *memory-based learning* or *case-based learning*) saves the whole training set as such and searches only at the time of prediction for the closest class of a new observation (by searching for the most similar observation in the training set). The processing effort is thus transferred from the training phase to the prediction phase with less time needed for training, but higher memory requirements and slower prediction times. Examples for instance-based learning methods are the k-nearest neighbour algorithm and its descendant *TiMBL* (Daelemans et al., 2007).

3.4.6 Lazy learners vs. eager learners

The distinction between lazy learners and eager learners is equivalent to instance-based learning and model-based learning but focuses on the time of learning and decision-making instead of the information used to decide when predicting. *Lazy learners*, i.e. instance-based methods only "learn" when the actual prediction has to be made, while *eager learners*, i.e. model-based methods already "learn" and abstract before by creating the model during training (Kotsiantis et al., 2006).

3.4.7 Probabilistic vs. non-probabilistic

Probabilistic models are statistical models that make use of prior probabilities and distributions of the data in order to make predictions. Alongside the simple Naïve Bayes models there are many variations based on Bayes rule of prior probabilities (e.g. *Multinomial Naïve Bayes*, *Bayesian Linear Regression*) as well as other methods like *graphical models, relevance vector machines*, *probabilistic principal component analysis* or *probability-based clustering* (Witten et al., 2016). Algorithms that are not based on probability data include *support vector machines*, *k-means clustering* or traditional *principal component analysis*.

3.4.8 Symbolic vs. non-symbolic (statistical/numerical) models

Models are called *symbolic* when they learn declarative knowledge representations which are inherently interpretable. *Rule learning* algorithms, *decision tree* algorithms or *instance-based classifiers*, for example, learn how to classify using very simple, intuitive approaches that display structure and knowledge (Mooney, 2003). *Non-symbolic models* are bases on mathematical functions to abstract from the data. Instead of interpretable rules, the model learns a numeric statistical representation, which is easy to process for computers in predictive scenarios but remains a black box, and therefore difficult to interpret without additional interpretation methods. Most modern machine learning methods such as *neural networks* and its derivatives as well as *support vector machines* are based on non-symbolic representations.

3.4.9 Intrinsically interpretable vs. black-box models

Related to the concept of symbolic and non-symbolic models is the division made on the basis of the interpretability of models. Machine learning models used for data mining are frequently classified as

intrinsically interpretable models (e.g. regression models, decision trees and rule learning algorithms) or *complex black-box models* (e.g. ensemble methods, support vector machines or neural networks) (cf. Molnar, 2018b).

3.4.10 Decision boundaries

One interesting division of machine learning models is the type of *decision boundary* used. There are models that draw two-axis-orthogonal lines to split the data (e.g. *decision trees)*, models that draw hyperplanes to maximize the margin between classes (e.g. *perceptrons* and *multi-layer perceptrons* or *support vector machines*), and models that look at adjacent examples like the *k-nearest neighbour* algorithm (Van den Bosch, 2009).

3.5 Standard methods for predictive modelling and their interpretation

This section introduces some of the principal machine learning methods used in data mining, focusing on supervised (predictive) methods (i.e. regression and classification) [27].

The method descriptions below introduce the general logic or idea behind the method, give references to implementations and adaptations, name known strengths and weaknesses of the method and provide examples for their use in linguistic studies. Whereas some of the methods are also frequently used in explanation-oriented communities (cf. regression and classification methods based on linear or logistic regression and its many derivatives like *generalized linear models*, *mixed-effects models*, *regularized regression* methods as described in section 3.5), most of the methods have been primarily developed and used in the machine learning and artificial intelligence community and are therefore oriented towards predictive uses without interpretation. The methods are sorted in increasing order of model complexity and decreasing order of interpretability. Although the model's actual performance and interpretability depends of course on the specific modelling task and the dataset used, indications on the average predictive power and interpretability of the methods are given. Finally, each method description comprises concrete and, if possible, linguistically-oriented examples of how to interpret them that shall allow judgment of their utility in data mining scenarios.

3.5.1 Linear regression and its derivatives

Regression methods are among the core analytical tools in traditional statistics. While the principal logic behind is based on summing weighted features to achieve an estimate close to the observed value for an instance (cf. simple linear regression as described below), recent variants like *generalized (additive) mixed-effects models* with manual or automatic *stepwise model selection* techniques are highly valued for their ability to deal with complex data (e.g. Gries, 2015c).

Simple linear regression

[27] For all of the methods, tested implementations are available via data mining software, machine learning frameworks or libraries for statistical software like R. Most software tools also provide sensical default values for hyperparameters and configurations that are expected to give stable results for any problem. Open source data mining software like WEKA or RapidMiner (see section 4.3.3.1) allow to apply them to one's own datasets without further programming skills. Furthermore, most statistical software (R, SPSS) offers libraries or built-in methods for many models as well, allowing to use familiar interfaces and minimize time spent on the transfer and conversion of data from one program to the other. Additionally, machine learning or text mining libraries for different programming languages offer convenient tools to apply these methods with minimal effort and basic programming skills of the respective language (e.g. *scikit-learn* for Python, *WEKA* or *ELKI* for Java or *Torch* for C).

In the most basic form of a regression model, the *simple linear regression*, a linear function is identified that explains the relationship between predictor variables and dependent variable best by minimizing the error between the predictions and the real observed values (see Figure 3).

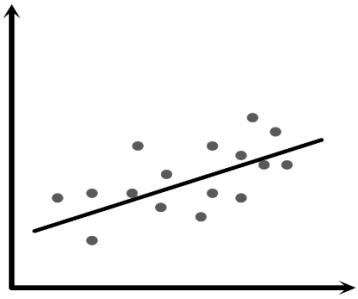


Figure 3: The regression line is the linear line that best fits the relationship between x and y in the data.

The function is a weighted sum of all feature values adjusted with an intercept value to account for values that do not start at zero. The model can thus be represented with the following function (Figure 4).

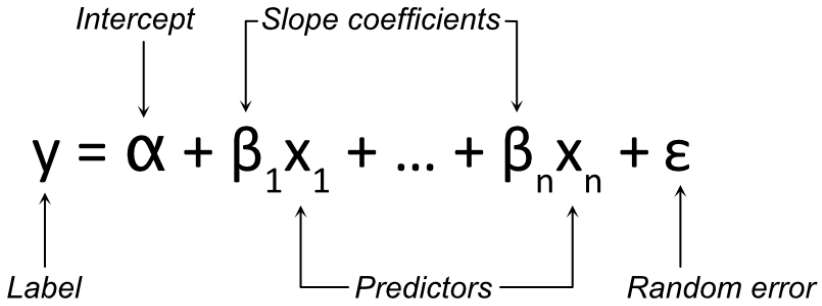


Figure 4: Prediction function for linear regression.

In order to identify the function that can best present (or *link*) the inherent relationships between predictor variables and dependent variable, regression uses a method to find the combination of weights that minimizes the error for the examples in the training set[28]. However, the method makes strong assumptions on the data, e.g. independent observations, normal distribution of errors, heteroscedasticity (i.e. when the variance of the residuals of a regression model is not constant), and linearity of the relationship between predictors and dependent variable that have to be controlled when using it.

Derivative methods of linear regression models

To deal with this issue, other methods have been developed that are less influenced by violated assumptions and can deal for example with non-normal distributions (*Poisson regression*), non-linear relationships (e.g. *polynomial regression* or *regression splines*) or sparse features (regression using regularization methods like Lasso regression, ridge regression or *Elastic Net regression*). Nelder and

[28] The most common method and basis for simple linear regression models is to use the so-called *ordinary least square* to find the "best fit".

Wedderburn (1972) proposed a technique called *generalized linear models* (McCullagh, 2018) which makes it possible to have non-normally distributed errors in the response variable with regressions within one framework by disconnecting the model building and the *link function* used. Another very popular adaptation of simple or generalized linear models are so-called mixed-effects models, which account for random effects introduced, for example, by hierarchical data and repeated measurements. Mixed-effects models allow for the definition of a different intercept and/or different slope for particular groups of data points, thereby removing the bias introduced through a certain random variable. This approach is highly advocated (Barth & Kapatsinski, 2018; Gries, 2015c) for corpus linguistic studies and particularly relevant when individual variation is possibly biasing the results.[29] Hence many modern sociolinguistic corpus studies refer to this method (see Hilte et al., 2017; Murakami, 2016; Spina, 2019a; Vajjala et al., 2016).

Model selection techniques

Linear regression modelling often refers to model selection techniques to identify variable or model configurations that have the lowest error, highest explainability or highest accuracy. Forward or backward model selection are common methods for variable selection. They iteratively search through the set of variables to find, in each step, the one variable that contributes least to the model and should therefore be excluded or contributes most to the model and should therefore be included (see also section 3.5.1). The models are compared with F-test ANOVA to establish that the difference in model performance is not due to chance.

Interpretation

Linear regression is considered one of the most interpretable predictive modelling techniques and is widely used in statistical analysis. This is mainly due to its linearity, which makes it easy for humans to understand the effects of a single feature. Apart from the numerical interpretation of regression coefficients, t-values, and related p-values, 2- or 3-dimensional effects (or interaction) plots help to show dependencies and interactions visually (see Figure 5 and Figure 6). The multiple and adjusted R^2 and the overall p-value of the model and other measures like prediction accuracy or MSE make it possible to judge the relevance, significance and performance of the built linear or logistic regression model. This further allows the building and comparison of different models using manual or automatic model selection techniques that search for the best combination of predictor variables similarly to model comparison and ablation studies in computer-science-oriented machine learning. Additionally, using regularization methods like Ridge regression or LASSO, sparse feature sets are created that also give insights into the relevance of individual features.

[29] Apart from the bias introduced by non-equally-represented writers or speakers, having more than one observation per writer or speaker usually introduces hierarchy in the data and produces "repeated measurements" that are problematic for simple linear regression models.

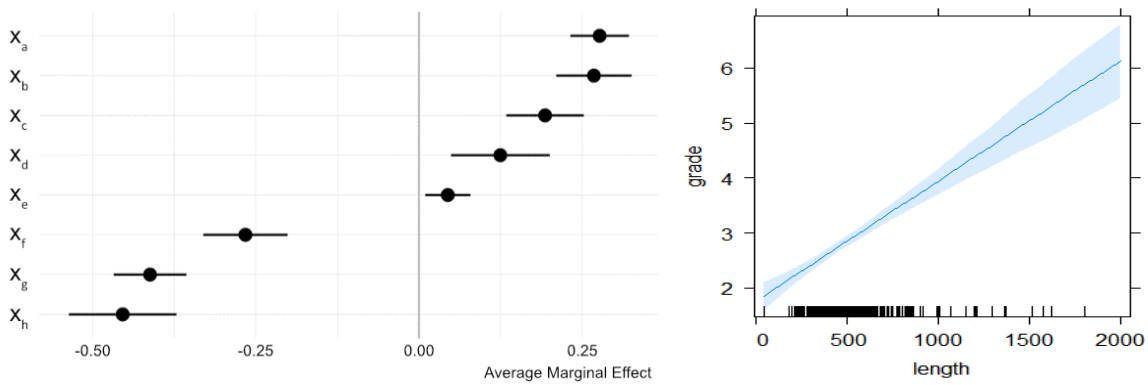


Figure 5: Graphical representations of variable effects plots in regression modelling.

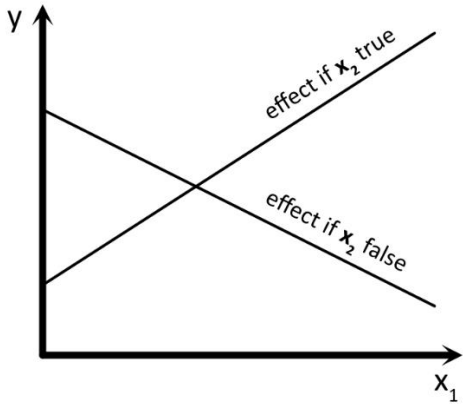


Figure 6: Graphical representation of an interaction effects.

However, the linearity assumption made in the model is also one of the shortcomings of linear regression models. It limits the applicability of the method and often lacks predictive power compared to other better performing but more complex models. Moreover, the other model assumptions might also lead to problems if they are violated. Multicollinearity, i.e. the presence of features that are correlated not only with the dependent variable but also with other independent variables, is a problem that can lead to weakened model performance as well as incorrect regression coefficients, and in turn to incorrect interpretations. Unfortunately, it is often difficult to avoid multicollinearity completely. Regression models are also built to single out and analyse the effect of individual features while considering possible confounders, i.e. other variables that have a moderating, enhancing or maybe even reversing effect on the relationship between the predictor and descriptor variables. Such confounding effects by variables that we are primarily not interested in are considered noise. They are often modelled as random effects in mixed-effects models (contrary to the main effects that are the focus of the analysis)[30]. However, if unknown or not regarded, confounding variables can have substantial influence on the conclusions drawn from the data. Confounders are per definition correlated with the dependent and the independent variable and therefore always introduce multicollinearity. Therefore, when using linear regression models, one usually has to control for bias caused by multicollinearity by

[30] Confounders, or rather, in this case, combinatorial effects of two or more variables that are actually of interest for the analysis are called (variable) interactions, the interaction effects of which we want to observe and interpret in the analysis.

calculating the variance inflation factor (VIF) for regression models (James et al., 2013)[31] or at least checking individual bivariate correlations using a correlation matrix. Moreover, because linear regression models estimate using additive weighted feature effects, the method uses per definition all features in the model (Molnar, 2018b). This can lead to models that do not perform well, just because there is too much noise in the data as all features are considered, no matter if they are useful or not. It is therefore crucial to remove unnecessary features from the models. Having many features, as is common for state-of-the-art research designs, using manual stepwise model selection to select relevant features is often not feasible anymore and automatic methods for sparse regression models like LASSO regression or automatic stepwise regression are needed.

3.5.2 Logistic regression

Logistic regression makes it possible to transfer linear regression analysis to classification problems, i.e. when the predicted value is not continuous. By transforming the output of a linear regression with a logistic function, thereby retrieving a number between 0 and 1, the predictions no longer present continuous values but can be interpreted as class probabilities, where prediction values above 0.5 represent predictions for one class (i.e. the reference class) and those under 0.5 for the other. The model then shows coefficients for all continuous variables indicating how the probability of the reference class changes when the variable value is changed by one. Categorical variables can be used by transforming them into one-hot-encoding, i.e. for each category of the variable a new *Boolean* binary variable is created (i.e. binary *true* or *false* values) that is set to *true* only if the instance belongs to that category. There are methods for both binomial as well as multinomial logistic regression suitable for binary or multi-class classification problems respectively. However, although the internal mechanics of the logistic regression model can be investigated in detail, one usually needs to transform the values in order to make them comprehensible for humans. Interactions can be modelled but must be indicated manually, increasing the complexity of the model immensely. It is possible to use a mixed-effects model structure (see section 3.5.1) for logistic regression as well. However, logistic mixed-effects models, especially when there are multiple modelled classes and more than a few variables are very difficult to evaluate (Barth & Kapatsinski, 2018) and even more difficult to interpret[32].

Use in linguistics

As in the case of linear regression models, logistic regression models have recently found broad diffusion in corpus linguistic research (see e.g. Desagulier, 2018; Gries, 2015c; Gries & Deshors, 2014; Speelman et al., 2018).

Interpretation

The interpretation of logistic regression models is similar to linear regression models. However, the additional log-transformation and the resulting shift from direct variable effects to odds ratios can be confusing, especially if variables with various categories are involved. Instead of linearly interpretable weights that are summed, the reported weights for logistic regression present probabilities between 0 and 1 that are multiplied. Furthermore, logistic regression models suffer from the same disadvantages as linear regression. Their performance on complex tasks is relatively low, possible interactions have to be inserted manually and might only make sense up to a certain depth of combination.

[31] Typically a VIF above 10 is considered high in most studies, however, it is suggested to use even lower thresholds as a VIF of 10 usually means highly correlated factors ($r > 0.9$) and already factors with a VIF of 2 have been observed to lead to falsely non-significant parameter estimates (Zuur et al., 2010). Zuur suggests a general threshold of 3.

[32] See also Speelman et al. (2018) or Gries (2015c) for more information on mixed effects models in quantitative corpus linguistic research.

However, logistic regression is still considered one of the most interpretable classification methods and allows binary as well as multi-class distinctions to be modelled. This interpretability is decreased by any modification that is done on the model to make it more powerful or to allow non-linearity in the target variable, residuals or relationships (e.g. via non-linear functions in generalized linear models) (Molnar, 2018b).

3.5.3 Decision trees

Decision tree learning is one of the simplest and most widely-used approaches in machine learning (Mitchell, 1997). While linear and logistic regression fail when there are non-defined interactions or non-linear relationships, decision trees are perfectly suited for those cases (Molnar, 2018b). The concept of decision trees is to iteratively split the training data into subsets that have a reduced variance of class labels. Each split is made using the variable (and a value within it) that best divides the data of the training set into classes.[33] This way a tree of if-then rules is generated that is able to represent the data on the basis of the input features given. During the prediction phase, the new instance just follows the rule tree down to the leaf it belongs to.

Figure 7 shows a simple example for a *regression tree* (i.e. decision tree for regression problems) used in the sociolinguistic study of Mairesse et al. (2007). The tree models extroversion of speakers (on a numerical scale from 1 to 5.5) based on the length of their utterances (word count), average pitch of their voice measured in Hertz and the observed variability of volume in decibel (intensity). The data is first split on the word length variable, with higher values leading to higher extroversion rates in the corresponding leaf nodes (bold numbers). By following the edges down to the leaf nodes, one can trace back the decision path for a prediction. The tree has an overall size of 9 nodes and 5 leaves and a maximum depth (or height) of 3.

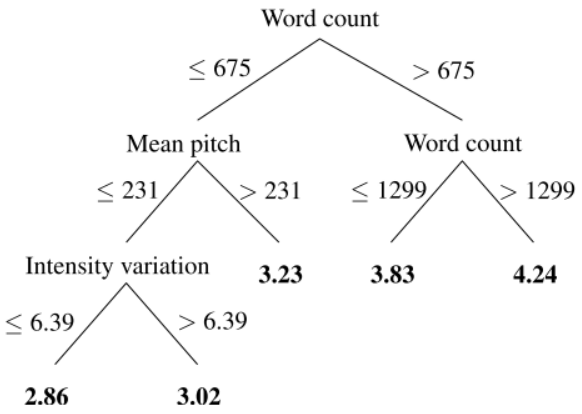


Figure 7: Regression tree model predicting the speaker's extroversion (scale of 1 to 5.5 where 5.5 indicates a 'strongly extrovert' speaker).

In principle, the splitting procedure would be repeated until all instances are sorted into noise-free classes, resulting in a tree that represents all the information it the training data and has very little generalizability for new data. In order to prevent this overfitting, tree-building algorithms usually provide means to define a so-called stopping criterion[34], which prevents the tree from growing too big

[33] There are different strategies or metrics to decide which of the features shall be taken as division criteria, most of them based on information-theoretic measures of entropy.
[34] This procedure is also called *pruning*.

and getting too detailed (e.g. by defining a minimum number of instances for the leaf nodes or a maximum depth of the tree).

Well known decision tree algorithms include C4.5 (Quinlan, 2014) and its WEKA implementation J48 (Hall et al., 2009) or CART (Breiman et al., 1984), the predominant algorithm for decision tree learning in R programming language[35]. Decision trees usually work well for classification problems. They are fast and flexible as they are robust to noisy data, errors and missing values, can handle numerical and categorical data and can even represent multifactorial, combinatorial relationships. Although decision trees are prone to overfit the data, in data science they are still highly valued for their rather straightforward interpretability (Alpaydin, 2014). Moreover, some of the best-performing ensemble methods achieve such high accuracy by combining the outcomes of various decision trees (see section 3.5.8).

Use in linguistics

Decision trees have rarely been used in linguistic studies up until now. Multi-dimensional analysis with various variables usually refer to the more efficient *tree ensembles, random forest* or *gradient boosting machines*, or to *support vector machines* that can model linear and non-linear relationships. Analysis with fewer variables mainly uses linear models to interpret linear (partial) regression lines for individual variable contributions instead of the combinatorial effects displayed by decision trees. However, the early linguistic data mining study of Daelemans et al. (Daelemans et al., 1997), the comprehensive data mining example of Mairesse et al. on linguistic cues of personality (Mairesse et al., 2007) and, more recently, Bernaisch et al.'s study on dative alternation with extensive use of predictive models (Bernaisch et al., 2014) are examples were decision trees where used for model investigation and statistical inference. Gries (Forthc.) gives a methodological account on decision tree models in corpus linguistics.

Interpretation

Decision trees are one of the most intuitive models to interpret by their internal representations. In order to see the decision process that leads to a certain prediction, one only needs to follow the decision path (all the branches of the tree that lead to the leaf where the prediction falls). The higher up a variable occurs in the decision tree, the more important it is considered to be for the model. Advanced decision tree visualisations allow the inspection of the remaining class distribution in the leaf nodes, the impurity of the subsample for each intermediate node and visual highlighting of all tree paths that lead to a certain classification. The model summary output of most implementations gives additional information about the tree model built, such as information on its size, number of leaf nodes or maximum/average height. The `rpart` package for R additionally provides a complexity parameter, with which the tree can be pruned (stopped from growing more detailed) and the model complexity-model performance trade-off can be estimated.

Figure 8, Figure 9, and Figure 10 show common visualization approaches for decision trees. The conditional inference tree in Figure 8 allows us to see the remaining variance in the response variable for the leaf nodes (represented as box plots for regression trees or as stacked or unstacked bar plots for classification trees).

[35] For implementations see the `rpart` package (Therneau et al., 2015) as an example.

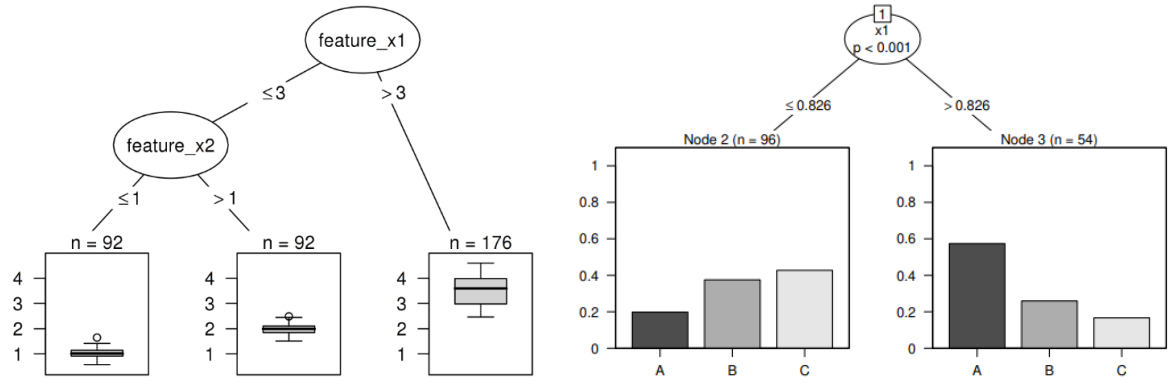


Figure 8: Examples of a decision tree visualization for a conditional inference tree for regression (left graph, source: Molnar, 2018b) and classification (right graph, source: Hothorn et al., 2015) built with the ctree package for R.

The default decision tree visualizations for the R package `rpart` or for the decision tree estimator in *scikit-learn* (Ellson et al., 2001) show additional information such as node impurity, probability for each class to occur in a node (and corresponding subtree), the sample size of the subtree belonging to one node, and more. The colour of the tree nodes and leaf nodes indicates the (prevalent) class (see Figure 9).

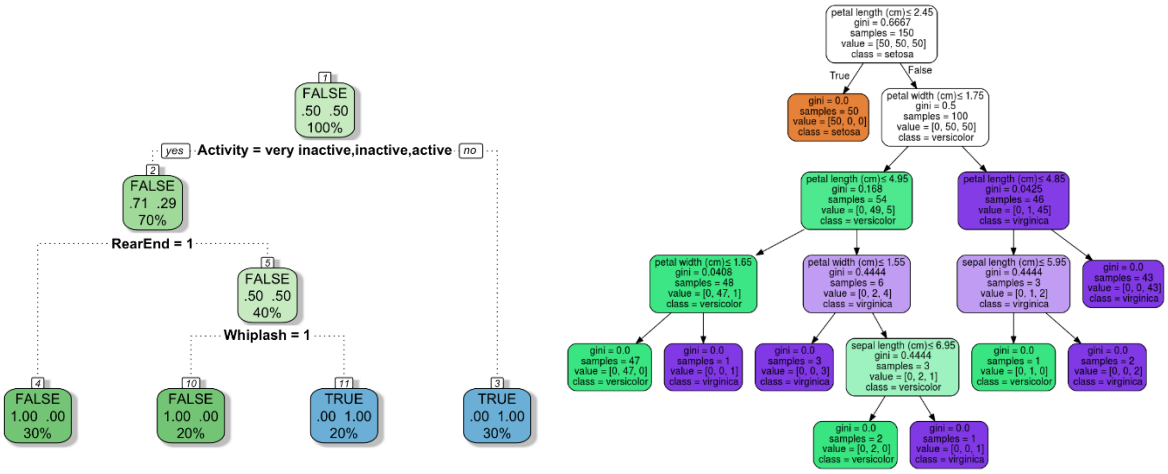


Figure 9: Examples for binary and multi-class classification tree built and visualized with `rpart` (left) and `scikit-learn` (right).[36]

Furthermore, the python module `dtreeviz`[37], offers an even more detailed visualization for decision trees that illustrates class distributions and splitting criteria for every tree node (Figure 10).

[36] Image sources: https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/ (`rpart` graph), and https://scikit-learn.org/stable/modules/tree.html (*scikit-learn* graph).
[37] https://github.com/parrt/dtreeviz

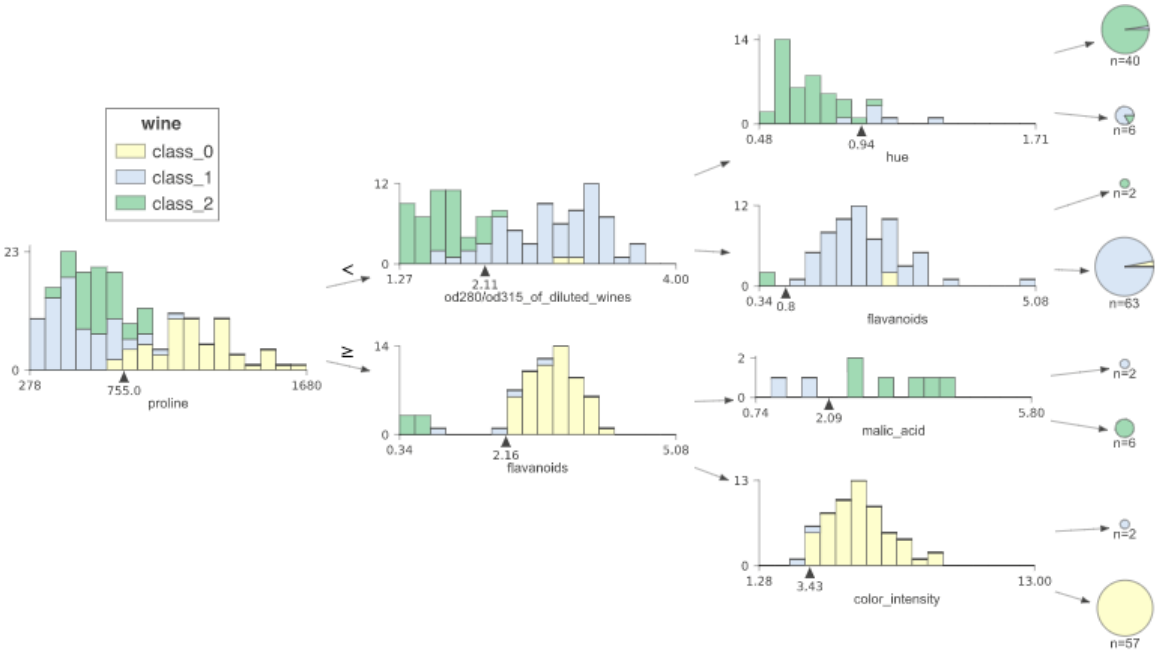


Figure 10: Example of a dtreeviz visualisation.[38]

3.5.4 Rule induction algorithms

Rule learning algorithms summarize the data by extracting rules based on correlations and associations in the data. The algorithms either work top-down by rule inductions or Bayesian rule list algorithms, or bottom-up by sequential covering algorithms (Mitchell, 1997). Covering algorithms learn one rule that describes a specific subset of data perfectly, then separate this data from the rest and continue iteratively with the remaining data. Rule induction methods on the other hand are similar to decision tree learning. They search for rules to split data on highly expressive variables to create subsets that are then split further and further until the data is ordered. Rule learning algorithms have the most expressive output to human readers and can be interpreted immediately (Mitchell, 1997) which makes them interesting for data mining approaches, although they are not among the most efficient methods and can even create contradicting rules. Common rule learning algorithms are e.g. OneR (Holte, 1993), RIPPER (Cohen, 1995) or PART (Frank & Witten, 1998).

Interpretation

The rules can be interpreted intuitively as if-then rules. However, interpreting rules with more than a few components becomes easily unfeasible. Additionally, very large rule sets can hardly be overviewed with human means. Furthermore, rule learners cannot display linear relationships, nor can they make use of continuous predictor variables. Numeric values have to be discretized or binned into intervals.

Figure 11 shows an example of extracted classification rules through the rule learning algorithm RIPPER in the implementation of WEKA (*JRip*) from Mairesse et al.'s study on linguistic cues of personality (Mairesse et al., 2007). The model was trained to classify openness to experience on a binary distinction, based on an essay corpus.

[38] Image source: https://explained.ai/decision-tree-viz/images/samples/wine-LR-3.svg

#	Ordered rules
1	(School $\geq$ 1.47) and (Motion $\geq$ 1.71) $\Rightarrow$ NOT OPEN
2	(Occup $\geq$ 2.49) and (Sixltr $\leq$ 13.11) and (School $\geq$ 1.9) and (I $\geq$ 10.5) $\Rightarrow$ NOT OPEN
3	(Fam $\geq$ 600.335106) and (Friends $\geq$ 0.67) $\Rightarrow$ NOT OPEN
4	(Nlet $\leq$ 3.502543) and (Number $\geq$ 1.13) $\Rightarrow$ NOT OPEN
5	(School $\geq$ 0.98) and (You $\leq$ 0) and (AllPct $\leq$ 13.4) $\Rightarrow$ NOT OPEN
6	Any other feature values $\Rightarrow$ OPEN

Figure 11: Example rule set from Mairesse et al. (2007)

The rules were induced when trying to classify openness to experience (as binary category) based on an essay corpus. The model predicts that people who refer to school, work or friends as well as those using longer, more familiar words in their essays are not open to experience.

Use in linguistics
Besides the comprehensive data mining study of (Mairesse et al., 2007), examples of recent linguistic studies using rule learning for data mining include Porhet et al.'s study on multimodal doctor-patient interaction that leads to patient confidence (Porhet et al., 2017) and the *CollOrder* system for detecting and correcting odd collocations in L2 writing (Varghese et al., 2013).

3.5.5 Naïve Bayes classification

Naïve Bayes classification calculates the probability that an observation belongs to one class using the distributions of the variables of the training data. This calculation is done based on Bayes' rule of conditional probabilities with a strongly simplifying assumption that is responsible for the name '*naïve'*: it assumes that all variables are independent and therefore also their conditional probabilities are independent. This probabilistic, generative classification approach is fast and easy to build as very little explicit training is needed. It can deal well with big datasets and can achieve good performances if the data is not too noisy and the features have been chosen wisely (Witten et al., 2016). It also cannot model interactions or dependencies between features as it treats each predictor variable as independent. The classic Naïve Base algorithm is the basis for a number of adaptations and evolutions following *Bayesian concepts*, e.g. *Bayesian Belief networks*, *Multinomial Bayesian*, or *Bayesian Networks*.

Use in linguistics
In linguistic applications the Naïve Bayes algorithm is one of the base algorithms for text classification and is often used as one of a set of different algorithms to compare classification results. For example, Pastor (2016) achieves her best results for an authorship attribution task based on phraseological features with a Naïve Bayes model. Simaki et al. (2016a) report particularly high *specificity*, i.e. the ratio of points that were correctly predicted as not belonging to a class from all data points that indeed did not belong to a certain class for the *Bayes Net* model in an age prediction task. Moreover, the automatic essay scoring system *BETSY* (Rudner & Liang, 2002) is based on a Bayesian architecture.

Interpretation
The interpretation of simple Naïve Bayes models is rather straightforward as the calculated probability distributions can be inspected (Molnar, 2018b). However, the Naïve Bayes classification algorithm needs additional data transformations when sparse features sets are used (i.e. smoothing) and usually does not deal well with noise or too many predictors.

3.5.6 Instance-based learning

Instance-based learning, also called example-based, memory-based or case-based learning (cf. section 3.4.5), does not extrapolate or abstract from the data but decides at the time of prediction which class to choose depending on the closest point in the whole dataset (Mitchell, 1997). It can thus be considered a non-parametric method because no assumptions about the data are made (Alpaydin, 2014). The method is also called *lazy learning* (Daelemans & Hoste, 2002), as practically the whole process is postponed to prediction time and training is only needed to store the data in efficient ways. Hence, instance-based learning algorithms are fast at training but usually very slow at prediction time. They need a lot of storage and processing power in order to be able to save and access the training data and are sensitive to the choice of similarity measure. They are also sensitive to noise in the data and do not deal well high-dimensional data, as they consider all attributes even if the target function is based on very few. The most famous instance-based learning algorithm, the k-nearest neighbour algorithm, searches for the k data points in the training data that are most similar, using *distance metrics* such as *Euclidean distance*, *Canberra* or *Chebychev distance* and then decides based on those data points which class the new instance should be labelled as. Other algorithms such as the *Tilburg memory-based learning* algorithms (Daelemans et al., 2007) are most often derivations of the k-nearest neighbour algorithm.

Use in linguistics
In computational linguistic studies on text classification the k-nearest neighbour algorithm is often used to complement other classification algorithms when comparing classification results (Alabbas et al., 2016; Rocha et al., 2017). However, the method is usually not among the best-performing models and has rarely been considered in more recent studies.

Interpretation
Moreover, as this classification algorithm does not really build a model of the training data, the explanatory power of the method is very low. Information can only be inferred on a local level, by investigating the neighbours of a given data point. Global structures like interactions or feature importances cannot be concluded and interpretation via close data points is restricted when many (e.g. continuous) features are involved.

3.5.7 Support vector machines

Support vector machine (SVM) is the name for a technique that is widely used for different modelling tasks and data mining scenarios. In general, the aim of support vector machines is to draw a decision boundary (see section 3.4.10) with p-1 dimensions[39], i.e. a *hyperplane,* between the data points which maximizes the margin between classes by "creating the largest possible distance between the hyperplane and the instances on either side of it" (Cristianini & Shawe-Taylor, 2000). After drawing the hyperplane that separates the data points best, the algorithm searches for a set of data points in the training data (so-called *support vectors*) that can be used to recreate the hyperplane while ignoring the rest of the points in the training data. These data points are then used to summarize the data (cf. Figure 12).

[39] P is the number of parameters.

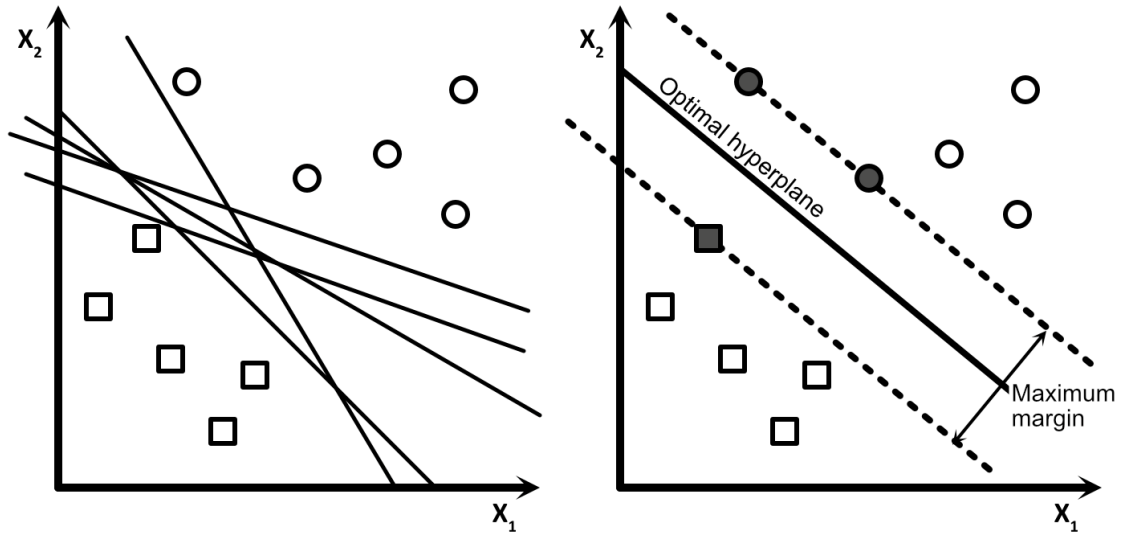


Figure 12: Left: possible decision boundaries for a binary classification task. Right: the best decision boundary separating the data points with the maximum margin.

There is, however, an important additional step, which makes this technique very flexible and yields good prediction performances for many modelling tasks. When trying to draw the discriminating hyperplane, different functions can be used. The simplest form of SVMs uses linear hyperplanes that describe linear relationships between a variable and the class. However, there is a way to use non-linear functions (e.g. polynomial, radial basis function, gaussian) to define the decision boundary, thereby transforming the data to a higher dimensionality and allowing non-linear decision boundaries to be drawn (cf. Figure 13). This so-called *kernel trick* [40] makes SVMs very flexible and efficient for complex classification problems with large feature sets (high-dimensional data) (Abbasi & Chen, 2007; Harish et al., 2010; Kotsiantis et al., 2007). SVM models are usually characterized by a high *precision*, i.e. how many of the positively labelled data points were correctly labelled, which is counterbalanced by a poor *recall*, i.e. how many data points belonging to the class were found by the algorithm[41].

[40] A kernel takes data input and transforms it using a mathematical function (see Figure 13).
[41] For a detailed description see section 4.3.3.2.

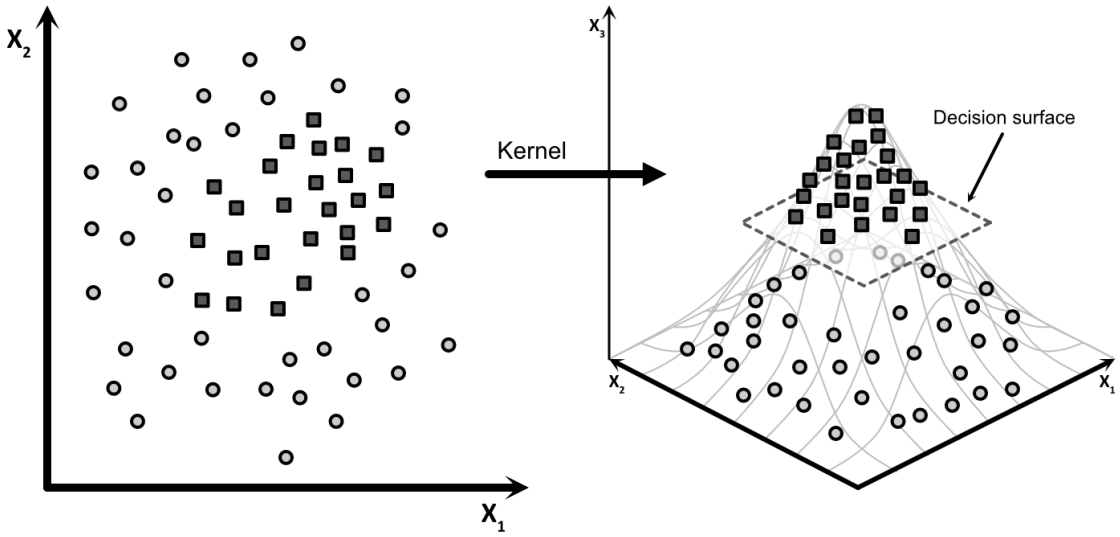


Figure 13: Schematic illustration of the "kernel trick" to model non-linear decision boundaries.

Interpretation

Because of the complex mathematical transformations involved in building well-performing SVM models, SVMs are usually not interpretable and are categorized as black-box models. In order to interpret variable importances and effects, SVMs need additional interpretation methods. Moreover, as the classifier tries to draw a single decision surface or boundary, the SVMs are not able to model multi-class problems in their basic form. However, they often refer to workarounds where various binary classifiers are used to approximate multi-class decisions, using the class with the highest combined probability. The interpretability of multi-class support vector machines is thus additionally hampered.

Use in linguistics

Nevertheless, SVMs are one of the most popular methods for predictive modelling on textual data as they have been found to perform well in many text classification tasks. In computational linguistic studies with classifier comparison they are often reported as the best performing algorithm. Recently they are also used in applied linguistics (e.g. computational sociolinguistics or writing research) to predict author characteristics like age or first language or language competence and inspect frequent prediction errors or compare performance measures for different feature sets in order to gain linguistic insights from linguistic corpora (e.g. Bykh & Meurers, 2016; Rabinovich et al., 2016; Simaki et al., 2016a).

3.5.8 Ensemble methods

The idea behind ensemble methods is to account for the fact that no learning algorithm or machine learning method can achieve perfect results for every domain or modelling task. This is also called *No Free Lunch Theorem* (Wolpert & Macready, 1995) and has its origin in the different strategies and assumptions different algorithms make. As stated by Alpaydin (2014),

> "with finite data, each algorithm converges to a different solution and fails under different circumstances. […] The performance of a learner may be fine-tuned to get the highest possible accuracy on a validation set, but this fine-tuning is a complex task and still there are instances on which even the best learner is not accurate enough."

Ensemble methods try to compensate for this by training and combining multiple models. An ensemble is composed of a set of *base learners* that are (usually designed to be) diverse and mutually informative, where some work better on some part of the data while others work better on other parts. The goal is thus to find and combine a series of models that complement each other, rather than optimizing the individual performance of each model. A simple example case would be when different learning algorithms are used for the same classification task and the class label most classifiers voted for is chosen as an end result. Choosing different algorithms is, however, only one approach of many. There are various strategies to modulate the base learners and also to combine them later on.

Base learners can differ in their hyperparameters, the algorithms, the input representations or the choice and splitting strategy of the training sets. For example, in *bagging* (short for *'bootstrap aggregating')* training is done on various, only slightly different training sets that are chosen randomly. *Boosting* and *cascading* on the other hand make use of the linearity of various training rounds. Boosting emphasizes problematic parts from earlier rounds in the later ones, while cascading starts with simple learners and only uses more complex ones on the examples where the learner is not confident. *Mixture-of-expert-models* only train classifiers on specific feature sets or domains to make base learners that focus on different aspects of the data. This is very efficient when different feature domains are weighted differently. Another method is the so-called *error correcting output codes*, in which the main learning task is divided into subtasks and each learner is only responsible for one subtask.[42]

As we already saw in the bagging and boosting approaches of training set splits, the base learners can be combined using a parallel or serial approach. *Multi-stage combination schemes* take a serial approach and train for example on instances where simpler learners did fail in previous training sessions. In *multi-expert combination schemes,* learners work in parallel and are either selected locally for different feature domains or fused globally with *voting* or *stacking* approaches ('*learner fusion*') (Alpaydin, 2014). The voting schemes can range from simple voting, through weighted voting approaches with different weights for different learners, to Bayesian model combination (making use of model-conditional likelihoods and prior model probabilities) or stacked generalization (i.e. when a subsequent learner learns the target label on the basis of the class predictions of previous base learners).

Below, I briefly introduce two of the most popular ensemble techniques that have outstanding results in data mining settings: *random forests* (Breiman, 2001a) and *gradient boosting methods* (Friedman, 2001).

3.5.8.1 Random forest

The random forest algorithm trains a high number of decision trees with different random subsets of instances or variables in parallel and then aggregates them to find a decision for a given prediction (usually by choosing the class label that has been assigned most often). It is very fast as the trees can be built in parallel and can avoid the problem of overfitting that is usually caused by decision trees. It can deal with a high number of features and successfully ignores noisy features without needing further feature selection methods. Furthermore, it can account for interactions and variable dependencies and is thus one of the most valued and best-performing algorithms used in data mining tasks.[43]

[42] For a more detailed description see Alpaydin (2014).

[43] Various people even called them the number one "go-to method" in recent times.

3.5.8.2 Gradient boosting methods

Gradient boosting methods on the other hand use and combine various models in a sequential manner. They first train a model with any machine learning method of choice, evaluate it, and then use difficult observations that had a high error for further training. In this subsequent training step(s) the residuals from the previous model are predicted, thereby finding adjustments to the base prediction that account for the errors. This procedure is repeated iteratively until the overall accuracy of the model does not increase anymore when used on a held-out test set. The continuous re-evaluation of the model on the test set is thereby crucial to establish generalizability as otherwise, the model would be refined until it reaches perfect accuracy on the train set and overfits the data.

Gradient boosting methods like *XGboost* or *Gradient boosting machines* are also very fast, because the training problem gets iteratively simpler with every training step (one only trains the errors of the last model). Moreover, they are efficient and high-performant in data mining tasks (Goldbloom, 2016). They have, for example, successfully been applied to predict the helpfulness of online reviews based on linguistic features of the texts (Singh, Irani, et al., 2017) or to classify Twitter users (Pennacchiotti & Popescu, 2011).

Interpretation
However, in terms of interpretability, ensemble methods are in general black box models, meaning that they don't allow immediate interpretability of the model internals. Additional methods must be applied in order to retrieve variable importance measures, retrieve interactions or explain individual predictions. This is why they usually only occur in corpus linguistic studies that are closely related to computational linguistics, where they are interpreted mainly based on classification errors and the comparison of model performances for different variable sets (cf. support vector machines). However, recently developed post-hoc interpretation methods help to investigate the black box models created with ensemble methods. Some of them focus particularly on tree-based ensemble methods like random forest or gradient boosting methods (Biecek, 2018; Louppe, 2014).

3.5.9 Neural networks, multilayer perceptron and deep learning

The internal logics of neural networks are often compared to the mechanism found in our brains. However, the connection between brain functions and neural networks is mainly of metaphorical nature. Biology thereby inspired and lent terminology to the computer scientists developing neural networks (Nielsen, 2015). A neural network is a complex system of connected transformation functions that learns to predict outcomes by input features.

The algorithmic idea of neural networks is somehow similar to both gradient boosting machines (as they depend on the iterative refinement of the model by evaluating its performance) and linear regression models (in the core mechanics of individual calculation units).

A neural network essentially consists of various layers of processing units (also called *neurons*) that transform given numeric input vectors into some output using input weights similar to linear regression models. However, there are various aspects of neural networks that differ substantially from simple additive function mapping. First, apart from the basic addition of weighted inputs, each neuron adds an *activation function* that can transform the inputs in order to produce non-linear input-output mappings. Secondly, there is not only one mapping function for all inputs into one output, but various such functions are combined together and aligned in so-called *hidden layers*, where each layer consists of various neurons that take a set of inputs and transform it into (intermediate) outputs that can then serve to inform (as new inputs) further neurons in subsequent layers. Consequently, each intermediate layer of neurons can take the output of previous layers in order to produce the input of subsequent

layers (or final output) – and each of those neurons can model a non-linear relationship. Thirdly, the inputs can be passed through the layers in a linear way, where each neuron feeds its output to the units of the next layer (also called feed-forward networks as schematically displayed in Figure 14). However, there are also possibilities to go back and forth in the hidden layer structure, inserting loops in the feed-forward structure (e.g. *recurrent neural networks*, *convolutional neural networks*). Finally, the actual training of the network happens by iteratively updating the weights backwards through all the layers in the network according to the prediction error produced with the current weights (i.e. *backpropagation*).

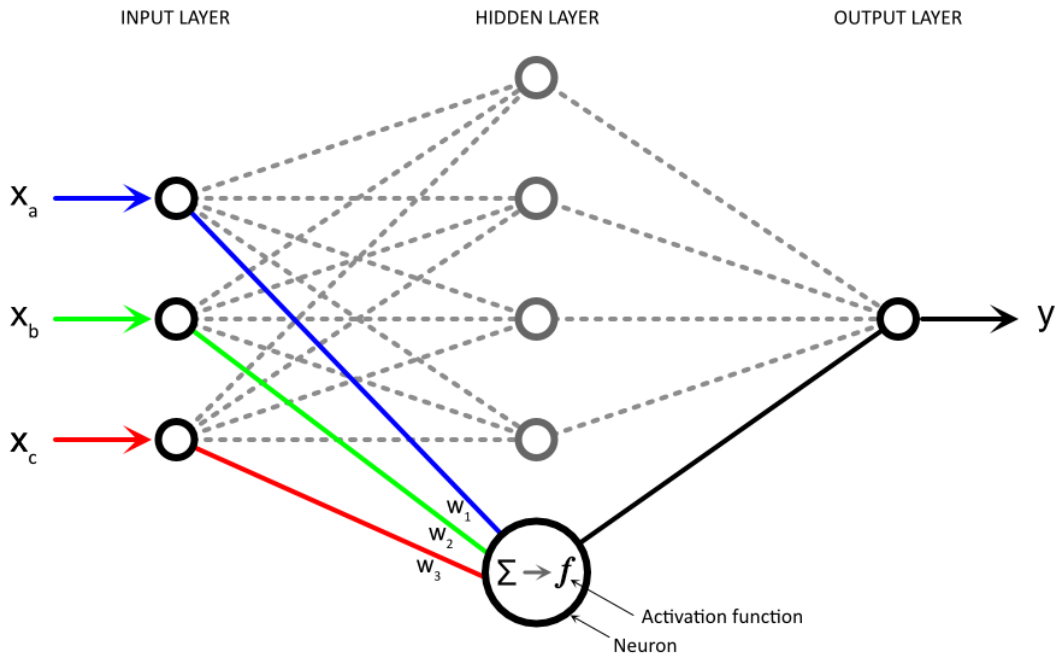


Figure 14: A feed forward network with one hidden layer.

While the general architecture and internal function of one neuron is called perceptron, a *multi-layer perceptron* (a frequently used term to denominate a neural network) is the combination of more than one perceptron into networks of various layers. Hence, the multi-layer perceptron is the basis of every algorithm that falls under the category neural network.

Over the last years, a myriad of different network types has been developed (Kelleher & Tierney, 2018) that make use of different activation functions, different network topologies (e.g. *recurrent neural networks*, *long-short-term-memory networks)*, different loss-functions to calculate error during training (e.g. *mean squared error*) and optimization algorithms (e.g. *gradient descent*, *attention-based neural networks with self-attention*), etc. The most important structural change of the simple multi-layer perceptron illustrated above usually derives from adding further hidden layers to abstract over the inputs, which is generally referred to as *deep learning* (i.e. networks that have more than one hidden layer to do calculations) (Goodfellow et al., 2016; LeCun et al., 2015).

Interpretation

Neural networks, and in particular, deep neural networks can yield very good performances in prediction tasks, especially when non-linear, multi-facetted relationships are present in large datasets. However, because of all the computation performed and the many abstractions and potentially distorting calculations they also need a lot of computational power, as well as rather large datasets that have

enough information to be distributed over the neurons and are also very difficult (or even impossible) to interpret without post-hoc interpretation techniques (Molnar, 2018b; Weld & Bansal, 2018).

3.6 Explainable artificial intelligence and interpretable machine learning

As can be understood from the descriptions above, the interpretability of machine learning methods, and thus their suitedness for text or data mining, can be restricted by the nature of the method (e.g. when using uninterpretable, complex, non-symbolic methods like neural networks), or the complexity of the problem it is applied to (e.g. when non-linear relationships, too many features or too many class levels are used).

However, the higher prediction performance of complex methods or complex model setups in model evaluation suggests that they match the actual problem that we want to model more reliably. Moreover, the success of intelligent applications that are based on such complex models (e.g. in automatic assessment or crime monitoring) depends not only on the numerical prediction performance, but also on the trust people put into them and on whether they are fair or known to have systematic biases that favour or disfavour certain cases. Hence, the interpretation of complex machine learning models is a research desideratum of crucial importance for many fields that use data science or machine learning methods.

The amount of research on the interpretability of machine learning and on explanation of predictive models has exploded in the last three years. Analytical, technical and societal needs have encouraged many scholars from different fields to develop strategies, methods and evaluation criteria for the interpretation of complex models (Guidotti et al., 2019). Although the topic is relatively new, and many denominators were used in the very first publications[44], two main fields have evolved almost in parallel and drive most developments in the topic (Adadi & Berrada, 2018):

- Explainable artificial intelligence (XAI)
- Interpretable machine learning (IML)

Explainable artificial intelligence (Došilović et al., 2018; Gilpin et al., 2018) is, as the name denotes, more related to the artificial intelligence community, focusing on deployable predictive systems, and a more general notion of AI. The main research agenda in this community is the development of transparent, unbiased and fair, explainable predictions for automatic decision-making, as well as the improvement of decision-making systems.

Interpretable machine learning (Doshi-Velez & Kim, 2017; Molnar, 2018b) on the other hand is more related to the machine learning and statistics community, has wide applications in image recognition and, less frequently, natural language processing, and is often more focused on gaining insights and understanding global and local relationships in the data.

[44] E.g. *intelligible intelligence* (Weld & Bansal, 2018), *comprehensible classifications* (Freitas, 2014), *black box understanding* (Koh & Liang, 2017; Louppe, 2014) or *interpretability* (Lisboa, 2013; Ribeiro et al., 2016a); See Doran et al. (2018), Miller (2018), Guidotti et al. (2019) or Gilpin et al. (2018) for discussions on the controversial terminology of interpretable machine learning.

3.6.1 Motives for research in explainable artificial intelligence and interpretable machine learning

As stated before, these developments are driven by a variety of reasons that can be clustered into societal (ethical and legal), technical and analytical or scientific needs.

Ethical and legal reasons: More and more processes in our daily lives are guided by intelligent systems that influence decision-making. Bank lending as well as decisions made on health treatment or insurance fees need to be unbiased and fair towards all parts of the society (Boyd & Crawford, 2012). However, intelligent systems are known to perform less reliably when there are huge imbalances in the data and there is only little data for a certain group. This can lead to systematic errors when minority groups are concerned (Krawczyk, 2016). Furthermore, systems trained on natural data, especially on textual data, have been shown to pick up and reproduce already existing biases in the data, leading to discriminating systems (Caliskan et al., 2017). In order to raise trust in a deployed predictive system as well as in order to ensure its correctness, systems need to be as transparent as possible (i.e. decisions must be validatable and explainable) for ethical but also more recently for legal reasons[45].

Technical reasons: While society urgently needs methods to explain the decisions of intelligent systems, the developers of such systems also have a certain interest in explaining the behaviour of the system. Only by opening the "black box" can they find systematic biases, debug algorithms and prevent hidden technical debts (Sculley et al., 2015) in order to improve the models and deployed systems. In addition, only systems that are trusted by their users will be accepted eventually.

Scientific reasons: Lastly, the scientific community has a strong interest in interpretability and explainability of complex predictive models for data mining applications. Given the trained models are a) predictive and b) interpretable, they can facilitate scientific inquiry by uncovering internal structures, raise the accountability of models, and allow to direct future data collection (Adadi & Berrada, 2018; Guidotti et al., 2019).

3.6.2 Model interpretability

In general, what is meant by interpretability of machine learning models is to "explain or to present in understandable terms to a human" (Doshi-Velez & Kim, 2017)[46]. An interpretable model is thus one that allows to understand the general behaviour of the model as well as why the model made a specific decision, or in more technical terms to understand "how the input is mapped to the output" (Doran et al., 2018).

As already discussed in section 3.4.9, we can divide predictive models into intrinsically interpretable models and complex black-box models, the former offering immediate access to symbolic representations that allow to understand the model internals (e.g. regression models), the latter not being immediately comprehensible for a human because of complex mathematical transformations during the modelling step (e.g. support vector machines or neural networks).

The main purpose of research into model interpretability is to develop *model-specific* or *model-agnostic post-hoc interpretation methods* (cf. Guidotti et al., 2019; Molnar, 2018b) that allow to interpret

[45] For example the EU GDPR law, released in 2018, set down a "right to explanation" of any automatically enhanced decisions (Goodman & Flaxman, 2017)

[46] Although originally, the term *explanation* was often used in opposition to interpretation for explaining a single prediction, and *interpretation* for explaining or interpreting global model functionality, the terms are increasingly used interchangeably. In this work I will therefore mainly refer to interpretation and interpretability as the terms are more frequently used in NLP-related machine learning.

global and *local* model behaviour as well as to find strategies on how to operationalize and evaluate the interpretability of a predictive model.

Global model interpretability concerns the whole model and aims to

- identify features or groups of features that have a significant effect on the dependent variable;
- estimate the importance of these features for predicting the dependent variable;
- evaluate the direction and size of the effect of features on the dependent variable;
- identify interactions among the features;

Local model interpretability on the other hand focuses on one instance or a limited number of close instances, giving local explanations for the corresponding model prediction(s), including which and how individual feature values of the observation influenced the prediction(s) (i.e. importance, effect size and direction and possible interactions of individual features on the prediction of one observation).

3.6.3 Methods for post-hoc interpretation of complex models

Post-hoc interpretation methods for complex models can be specific to the model or model-agnostic.

Model-specific methods compute feature importances and feature effects (that are not immediately available) extracting internally saved values like entropy measures for individual trees or feature weights for neurons in a network and calculate variable importance or marginal effects based on them. They are usually more reliable and robust than model-agnostic measures, but also difficult to compare with other model-specific methods.

Model-agnostic do not depend on the type of model. They can be applied to any machine learning model as they do not base on internal values but solely on the observation of the model behaviour, usually using strategies like controlled permutation of input features to approximate the relevance, importance and effect of features or interactions. While being less robust, they have the advantage of providing comparable values for different types of machine learning models (cf. Ribeiro et al., 2016a).

In the following, recently proposed model-specific and model-agnostic methods for the interpretation of complex black box models are introduced.

4 Data science in corpus linguistics: From theory to application

While the prerequisites for data science and in particular predictive modelling in corpus linguistics seem to be given by the battery of existing machine learning methods and additional possibilities for model interpretation, the actual application of such methods in corpus linguistics is still limited. Theoretical knowledge about both predictive modelling and the linguistic domain, as well as technical skills are needed to perform such interdisciplinary analysis. However, there are more and more instances where data mining techniques have been applied to corpora and other language data, in particular when we consider the broader field of applied linguistics. Furthermore, text mining and NLP communities produce an increasing amount of methods, frameworks and tools that lower the need for technical skills, and theoretical background knowledge on NLP while offering abstract and

transferable approaches for language modelling. Both developments are leveraging corpus linguists who aim to use predictive modelling and machine learning for methodologically correct, multifactorial analysis and/or the re-utilization of language corpora.

Hence, this section looks at existing examples of data science for linguistic inquiry and at concrete methods, frameworks and tools that can be used for such approach. After a historical overview of the use of data mining, text mining and predictive modelling in applied linguistics and the discussion of the current situation, an interdisciplinary set of applicable strategies and available tools are identified for each step in the modelling process.

4.1 A historical perspective on data mining, text mining and predictive modelling in applied linguistics

There have been successful studies utilizing data-driven, computational methods for text analysis and linguistic inquiry in applied linguistics, although the terms *data mining, text mining* or *data science* have only been mentioned occasionally. Below the main cornerstones of using data mining, text mining and predictive modelling in applied linguistics are outlined in a historical review.

4.1.1 Data mining in applied linguistics

The first study that explicitly named the term *data mining* for linguistic research was (Daelemans et al., 1997). In this study, Daelemans et al. proposed a methodological framework using machine learning methods for confirmatory as well as exploratory corpus analysis. They stated that data mining can be used for the evaluation of hypotheses. In order to evaluate two competing theories and compare them against each other, the following strategy could be applied[47]:

1. Collecting a representative corpus
2. Annotating the *concepts deemed relevant for each of the two theories*
3. Analysing the corpus by computing the learnability (non-random prediction) of the linguistic phenomenon using a learning algorithm and by analysing the different annotations with statistical techniques

(4.) Making claims about the necessity of a particular variable for the explanation of a phenomenon, by comparing the performance of the models that were trained on the differently annotated corpora.

Second, they stated that data mining can be used for the discovery of theories. To build new hypotheses, the proposed strategy consists of three steps:

1. Collecting a representative corpus
2. Annotating *any possibly relevant* annotation
3. Extracting generalizations and categories using learning algorithms.

Daelemans et al. used a decision tree algorithm to demonstrate the proposed framework in an analysis of Dutch diminutives. Both types of analysis (although somewhat vaguely defined) are based on the application of a machine-learning-based predictive model with subsequent evaluation and performance comparison.

[47] The fourth step is not numbered but listed underneath in the original paper.

In the years following Daelemans et al.'s paper, the term *data mining* and the methodological framework of training, evaluating and interpreting a predictive text classification model was further used in 2005 by Baayen and Cutler (Baayen & Cutler, 2005) in a psycholinguistic study, and in 2009 by by Teich and Frankhauser (Teich & Frankhauser, 2009) in a study on linguistic registers as well as by Thelwall et al. (Thelwall et al., 2009) in a sociological investigation of emotion in social network communication. All of them claimed to use *data mining techniques* on linguistic corpora for a linguistically motivated analysis. However, their strategies differed clearly from each other.

Baayen and Cutler (2005) as well as Teich and Frankhauser (2009) used predictive models to investigate empirical linguistic data. Baayen and Cutler (Baayen & Cutler, 2005) used *stepwise model selection* for *multiple linear regression* in their exploratory investigation of linguistic predictors and therefore followed an approach originating in traditional statistics.

Teich and Frankhauser (Teich & Frankhauser, 2009) used machine learning methods primarily used in natural language processing and computer science to study the differentiability of scientific registers. They used *feature ranking, clustering, and classification techniques*, summarizing all three techniques under the umbrella term data mining. Instead of comparing the "learning" and performance of individual machine learning models for two competing linguistic theories (cf. Daelemans et al., 1997), their approach used the text classification scheme from machine learning to establish the difference between four subcorpora, testing if an automatic classification is possible. Furthermore, they evaluated misclassifications between the individual registers[48] and performed an exploratory analysis of task-related linguistic features using *feature ranking techniques* (see section 4.3.2.3). Compared to the strategy proposed by Daelemans et al. this method allowed to identify the features that are relevant to discriminate between subcorpora (registers) instead of analysing the features related with one register[49]. The university of Saarbrücken in Germany has since then concentrated on *data mining* for register studies (Teich & Frankhauser, 2009)[50], various research projects focused on the study of language, registers and linguistic variation using data science methods. Newer studies use information-theoretic methods to identify typical features and the amount and type of information they carry to address intra-textual variation, diachronic change, information density, and linguistic encoding (Crocker et al., 2015; Degaetano-Ortlieb et al., 2019; Degaetano-Ortlieb & Teich, 2016, 2017).

The last study by Thelwall et al. (2009), however, manually classified social networking texts for their strength and polarity of emotive language. Although they used a methodology similar to the one described by Daelemans et al. (Daelemans et al., 1997) (i.e. classifying a text in terms of different categories and then investigating the characteristics related to it), their "classifiers" were human annotators and their observation of related features were summarized findings during the annotation process. Although the authors used the terms classifier, text mining, data mining, opinion mining and computational linguistics, the study itself had an entirely qualitative approach und would thus, rather qualify as content analysis.

Between 2012 and 2015 the project KobRA (*Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining*) that was a joint project between various German research institutions focused on data mining for corpus linguistics and digital humanities. It investigated "the benefits and issues of

[48] Although an overall classification accuracy was stated, there was no basic evaluation of the "learnability" of the task (cf. the methodology of Daelemans (1997) described above). However, the accuracy was at 96% comparatively high and strongly suggests that the classifier successfully learned how to distinguish the registers.

[49] This study is also interesting, as it uses off-the-shelf implementations for machine learning provided by the data mining software WEKA (Hall et al., 2009) and therefore presents a very applicable approach, even for researchers without advanced programming skills.

[50] See also (Degaetano-Ortlieb, 2015; Degaetano-Ortlieb et al., 2014; Teich et al., 2013, 2015).

using machine learning technologies in order to perform after-retrieval cleaning and disambiguation tasks automatically" (Bartz et al., 2015). It thus focused more on practical issues and the automation of manual tasks during corpus analysis, than on the use of predictive models for analysis and interpretation itself. However, the project also produced plug-ins for the data mining software RapidMiner that can be used to analyse, e.g. diachronic change in word semantics using topic modelling on KWIC (keyword-in-context) lines (see section 2.2.1) that can be extracted from big German reference corpora. The plug-ins also include methods for customized visualisation and interpretation of the topic models[51].

4.1.2 Text mining in applied linguistics

In parallel, the first linguistic studies using *text mining* were published. The term *text mining* became for the most part relevant in the field of literary studies and the digital humanities.

Plaisant et al. (2006) published a highly cited article that explores erotics in literary study in the NORA project (see also Don et al., 2007). The study uses machine-learning-based classification in combination with a correlation analysis as well as a visual interface to explore predictions along with the text (including highlighting of indicative words for the category).

A similar approach was taken by Hota et al. (2007). They investigated lemma unigrams and trigrams that were particularly relevant for a well-performing classification model that predicts the gender of characters in a Shakespeare play. This machine-learning-based exploratory investigation of gender-specific n-grams was accompanied by a qualitative investigation of concordance lines for the best unigrams/trigrams. Both n-grams and concordance lines are frequently used techniques in corpus linguistics (see section 2.2.1), hinting to the general applicability of such methods for corpus analysis[52].

Archer and Rayson published an article on *key domain analysis* as a method of corpus linguistics that can perform text mining tasks in digital humanities and social sciences, thereby relating the general toolset of corpus linguistics with the data mining aim to find new interesting patterns in (text) data. However, they did not make use of predictive modelling in their analysis (Rayson & Archer, 2008).

4.1.3 Other applications of machine-learning-based predictive models

Although the examples of the explicit use of *data mining* or *text mining* are rare before 2010, the strategy of using machine-learning-based predictive models can be found also in linguistic studies that did not necessarily refer to the terms data mining or text mining.

Pennebaker et al. (Pennebaker et al., 2002) for example proposed the term *word pattern analysis* for bottom-up-analysis of typical word patterns in texts through *latent semantic analysis* (Landauer et al., 1998), and related this technique to artificial intelligence.

Van Halteren et al. (2005) used classification models to show the existence of a "*human stylome*", i.e. stylistic characteristics of language use that allow to identify the writer of the text. This approach contributed to the field of stylometry and to modern authorship attribution and author profiling (see section 4.3.1.1) that is almost exclusively based on machine learning methods that use linguistic features to classify some category of interest[53].

[51] See Pölitz (2016a) for a case study on language change using this framework.
[52] Although Hota et al. (2007) would probably classify as a literary study.
[53] However, the goal of these studies is usually not to investigate the relationships between the features and the constructs (author characteristics) but to build the most accurate system possible.

Furthermore, Baroni and Bernardini (2006) used machine learning and predictive models for register studies in translation studies. They called it a *machine-learning approach* using *shallow linguistic features* (i.e. not hand-coded but automatically annotated word, part-of-speech category and lemma n-grams) with a complex learning algorithm (SVM) to classify texts into translated and original texts. They also used a backwards selection technique of features that significantly contributed to prediction performance to investigate relevant features of *translationese,* the linguistic variety produced by translators. By doing so, they were one of the first linguistic studies to use a strategy that is now usually referred to as *ablation study* in machine-learning-based data-driven corpus analysis and computational sociolinguistics (cf. Ilisei, 2012; Malmasi & Cahill, 2015; Yannakoudakis et al., 2011).

Yu (2008) did a comprehensive study of prediction accuracy for automatic classification based on different text representations (varieties of word counts, stemming and stop word removal), different amounts of linguistic features used for classification and different learning algorithms, including statistical significance tests for accuracy differences[54] and can be seen as one of the first studies systematically using *feature engineering techniques* (see section 4.3.2) for linguistic inquiry.

Noteworthy is also Abbasi (2008), who published a detailed summary on previous attempts on text classification methods, used linguistic feature types and possible research designs. The study included an extensive discussion on visualization approaches to investigate important features for automatic text classification (*ink plots* and *writeprints*) and represents a first attempt towards better *interpretability* of text classification studies.

The feature engineering approaches in Yu's text classification study focused on various representations and sets of shallow features classification, more abstract (psycho-)linguistic measures (e.g. LWIC by Pennebaker et al., 2001) have been explored by Mairesse et al. (Mairesse et al., 2007) in their study of linguistic features for personality prediction. The study combined many strategies and methods from an interdisciplinary view that are until now widely used in data science[55]. The analysis is based on previously discovered (although weak) correlations between linguistic features of writing/speech and personality and aims to a) control if these features also have the power to predict personality traits and (i.e. testing the existing theoretical basis) b) identify hitherto unknown relationships and gain new insights (i.e. elaborate old and generate new theories). It selects an extended set of possible predictor and predicted variables, uses classification methods next to regression and ranking models of the time, evaluates if prediction results are above a baseline and therefore allow to accept the model for generalizations, systematically compares the effect of different feature sets and different operationalizations of the predicted features and interprets the models "qualitatively" using decision trees and rule-based algorithms.

In the digital humanities and social sciences the terms *macro-analysis* and *micro-analysis* (Jockers, 2013) as well as *close* and *distant reading* (Moretti, 2013) were developed and frequently used for methodologies that we would now call data mining or text mining (Jänicke et al., 2015; Wiedemann, 2013).

Coming from the social sciences, Schwartz et al. (Schwartz et al., 2013) proposed a *differential language analysis* with an *open-vocabulary approach* for analysing age-, gender- and personality-specific language. The authors use the term *open-vocabulary* (as opposed to closed-vocabulary approaches using word count features for restricted lexica) for word n-grams and topics extracted with NLP tools

[54] A step that is still often neglected in computational linguistic studies applying similar techniques.

[55] Apart from using of relatively high p-value for statistical significance and not reporting the use of any corrections or repeated testing which might, with such high number of performed tests, lead us to expect that some of the significant relationships are not correct.

and filtered according to their bivariate relationships to the dependent variable. These open-vocabulary lists were then used to perform regression analysis and evaluate predictive models. Subsequently, the authors used specially designed visualizations for analysis with many features to get insights from the data.

The field of *computational sociolinguistics* was then proposed to describe sociolinguistic analysis that make use of computational methods (Nguyen et al., 2016). First studies in this field were prevalently focused on feature ranking and classifier comparison for authorship-related text classification tasks. In learner corpus and second language acquisition research the terms *detection-based approach* (Jarvis, 2012)[56] and *key structure analysis* (Ivaska & Siitonen, 2017) were proposed for using text classification to find patterns of cross-linguistic inference in language learning.

Summary

In the past, learning algorithms have been used in applied linguistic studies to model linguistic theories or to explore the linguistic patterns in a corpus.

Within the context of text classification methods, both confirmatory and exploratory study designs have been tackled, and models have been trained to distinguish

- different registers, thereby informing variation studies and translation studies
- different author characteristics like gender, age, personality thereby information sociolinguistic research and linguistic forensics
- author's first language or second language competences, thereby informing second language acquisition, learner corpus research and research in academic writing development
- qualitative from non-qualitative texts, thereby building the basis for modern writing assessment
- etc.

Regression models, predicting numeric outcomes and investigating linear (or non-linear) effects of linguistic features as well as interactions, have been used primarily in psycholinguistic studies or cognitive linguistics that work with numerical measurements like response times.[57] Apart from that, clustering methods (e.g. topic modelling) have served for preliminary exploratory corpus analysis, allowing to group data points without pre-existing categories, while outlier analysis has been used to point to interesting and/or erroneous data.

Many studies have focused on the analysis of linguistic features related to some concept (e.g. by comparing results for different feature sets). Only rarely do studies test different operationalizations for a predicted class (e.g. of age, text quality etc.). Possible models have been tested by evaluating the prediction performance of a trained machine learning model, i.e. comparing it against random prediction and against other competing models. Misclassified observations have been investigated to infer inherent logic represented in the subcorpora or classes. Relationships between linguistic features and a class or subcorpus have been evaluated by *comparing mean values* for automatically classifiable categories or *most frequent words* for the individual classes. Alternatively, *ablation studies* or *feature ranking* approaches have been used to explore black box models. The investigation of variable effect, including type, direction and magnitude of the effect has, however, barely been included in the past.

[56] But see also Jarvis' (Jarvis, 2011) article from 2011 on data mining with learner corpora.

[57] However, regression model are also used to predict text quality scores or language proficiency levels which are not inherently continuous, but ordinal variables (Crossley, Kyle, Allen, et al., 2014; McNamara et al., 2010).

4.2 Current situation

Since the first studies in the late 1990s, data science methods can be observed with increasing frequency in applied linguistics (see also de Marneffe & Potts, 2017; Gries, Forthc.; Milin et al., 2016; Periñán-pascual, 2017; Zeroual & Lakhouaja, 2018). Many current studies

- evaluate if a predictive model is able to produce non-random predictions when trained on linguistic features that are supposedly relevant for a concept according to linguistic theories;
- evaluate which linguistic features give the best results when training and evaluating a machine learning model;
- evaluate the salient linguistic features in automatically distinguished classes;
- inspect importance, effect and response for variables and variable interactions in statistical models like linear or logistic regression;
- analyse most frequent misclassifications;
- rank the importance of linguistic features used in a modelling task;
- use clustering methods to explore similar observations or similar features;[58]

They build upon predictive modelling with methods coming either from traditional statistics or from machine learning research and artificial intelligence in order to analyse linguistic data. However, there is often a methodological difference between studies drawing from the field of traditional statistics (often related to cognitive linguistics and psycholinguistics, but also variationist studies) and those drawing from computational linguistics and computer science (often related to more applied linguistic disciplines like sociolinguistics, writing research but also register studies). The former focus on detailed interpretation and explanation of built models but are often limited to simple models on well-structured and designed data with few investigated features. The latter often use complex machine learning models as a black box and only interpret them superficially through the comparison of prediction results.

Approaches incorporating both sides, complex modelling techniques as well as in-depth interpretation are, however, still rare in applied linguistics[59].

4.3 Methods, frameworks and tools for data science in corpus linguistics

The following section discusses the steps of a data science approach to quantitative corpus analysis using predictive modelling, and how they can be leveraged by methods, existing frameworks, and available tools that have been developed in other disciplines. With a clear focus on applicability, it names recent developments in NLP, machine learning, artificial intelligence, statistical analysis, information extraction and other fields that can be used for linguistic inquiry on corpus data. Focusing on tools that are readily available for non-technical audiences as well, the individual sections refer to state-of-the-art research and recent trends to provide entry points for further investigation.

The section starts with the selection of the modelling task, operationalization of the target variable, and selection of relevant (linguistic and non-linguistic) predictor variables, showing common strategies for text mining research designs. Secondly, strategies and tools for extracting features and preparing the feature set are introduced, including feature extraction, feature transformation and feature selection. Then, the actual model building is addressed and user-friendly tools for training and

[58] However, although these strategies are not mutually exclusive, few studies combine more than one strategy.
[59] But see for example Desagulier (2014), Murakami (2016), Deshors and Gries (Forthc.), Gries (2019) or Spina(2019b).

evaluating a model with the feature set and target variables are presented. The model evaluation section explains major concepts for validating and comparing models based on their predictive performance. Finally, the section on model interpretation describes strategies to make sense of the built models that have been used in the studies presented in section 4.1 and presents new strategies developed for explainable artificial intelligence and interpretable machine learning that could substantially enhance the interpretation of model internals in corpus linguistic studies.

4.3.1 Research design: Choosing modelling task, descriptor and predictor variables

The research design defines the modelling task, the dependent and independent variables (i.e. features) including their operationalizations used. Although the choice of the construct or concept to model is practically random, there is a number of text mining tasks that have established as independent disciplines in NLP, of whose prior experience one can base further decisions on the choice of target variable operationalizations, feature sets and learning algorithms used for analysis.

In the following, common text mining tasks are introduced briefly, giving references to further reading as well as indications on typical operationalizations and features used for those purposes. Afterwards, general types of target variable operationalizations and feature types are discussed. The section gives pointers to recent studies that have relevance for corpus linguistics, whenever possible.

4.3.1.1 Common modelling tasks and frameworks in text mining

Since the first uses of data mining for textual data, a number of repeatedly tackled text mining problems emerged and are now established as independent disciplines that have their individual research communities including targeted conferences or regularly recurring shared tasks[60]. However, these tasks have also become frequently applied frameworks in other fields outside of NLP and machine learning research. The list below shows examples of common text mining tasks.

- Author attribution, author profiling, personality prediction or native language identification
- Automated essay scoring
- Register studies and genre discrimination
- Opinion mining and sentiment analysis
- Topic modelling and semantic analysis
- Argument mining, event mining and trend mining
- Virality, relevance or review helpfulness prediction
- Detection of spam, hate speech or other malicious text
- Social media analysis
- Business intelligence, market analysis, competitor analysis
- Mining of genome information, medical diagnoses, patient files and medical literature

Most of the tasks named above are supervised machine learning tasks, where the aim is to predict a certain construct of interest, like the author, the genre or some other characteristic of the text. A few text mining tasks are also of unsupervised nature, making use of clustering or anomaly detection techniques (e.g. when modelling the topic of a text or identifying outliers). While many tasks focus on writing *style* as discriminative factor, machine learning and predictive modelling can also be used to analyse the *content* dimension of a set of documents. For example, we can categorize *authorship attribution*, *author profiling*, *register* and *genre discrimination*, *native language identification* and

[60] i.e. competitions, where various people or groups of people try to solve the same modelling task aiming, in general, for the best prediction result.

language proficiency prediction as mainly style-oriented approaches that draw from so-called *stylistics* or *stylometry* (Lynch, 2009) to classify texts while intentionally excluding content features. *Sentiment analysis*, *opinion and argument mining, topic modelling, event detection* and *trend mining* on the other hand are mainly concerned with the content and the semantics of a text and thereby represent the opposite side of the spectrum. Other applications of text mining freely draw from both types of information and combine them depending on the underlying hypotheses. This regards for example *social media analysis*, the *detection of malicious or helpful text*, *relevance or virality prediction and* different types of *business intelligence, market analysis* and *competitor analysis*.

Below, some of the most important of these tasks are introduced briefly.

Authorship attribution and author profiling

Authorship attribution and the related tasks *authorship verification* and *author profiling* (Estival et al., 2007; Stamatatos, 2009) have been particularly fruitful tasks for many different fields in applied linguistics and text mining. The main idea behind those tasks is to predict and describe the author of a text based on features of the text or context.

While approaches in this area started with authorship attribution (i.e. the assignment of one of a set of possible authors to a given text) and authorship verification (i.e. a slightly refined task, in which the set of candidate authors is not given but the question is rather to classify if a given text was written by a given author or not), by using first statistical models and then increasingly complex machine learning methods, many recent approaches also capture more general author characteristics like gender, age, social class or similar author categories. The term *author profiling* (Argamon et al., 2009; Raghunadha Reddy et al., 2016) thus subsumes a variety of different subtasks like *gender discrimination, age prediction, personality prediction* or *native language identification* and others. Although a lot of the technological and methodological progress has been driven by the forensics community, the framework of authorship attribution and its various derivatives yielded a lot of interesting insights in other fields that work with text data and language corpora. It has been applied in the social sciences (Bamman et al., 2014; Kosinski et al., 2016) and humanities (e.g. Jockers & Mimno, 2013; Moretti, 2013; Rockwell & Berendt, 2016), political sciences (e.g. Karami et al., 2018; Tumasjan et al., 2010), media studies (e.g. He et al., 2013; Lu & Szymanski, 2018), and other disciplines including applied linguistics.

In general, authorship attribution or author profiling are most often approached using supervised machine learning with text classification or occasionally with regression models. The used linguistic features usually come from *stylometry or stylistics* (Lynch, 2009; Oakes, 2009), namely lexical, content-independent features like function words, newer approaches using unstructured, multilingual or multimodal data are barely used for studies that aim for insights drawn from the data. However, there are some studies that also look at semantic features of the texts (Kosinski et al., 2016; Rangel & Rosso, 2016). However, operationalization of dependent variables for author predicates seems to be a critical task in all these approaches. Research ethics have to be kept in mind, especially when results are not only used to satisfy the interest of the research community but can also be a used for decision-making either in general, e.g. when officials implement new laws or policies based on such research results or even on a personal level, e.g. when prediction models are used to identify potential risks, or sanction certain people (Emani et al., 2015; Kosinski et al., 2016; Lambiotte & Kosinski, 2014).

The subtask of *natural language identification* (Koppel et al., 2005) furthermore paved the way for a series of related work that was highly interesting also for language learning and language teaching, second language acquisition and learner corpus research. Koppel et al. already used a variety of error-related features for their prediction methods and gave tentative explanations for the discriminative

value of different error types for individual native languages. This idea was then extensively exploited for language learning research and second language acquisition, writing research and related fields, giving rise to the creation of specific methodological frameworks like the *detection-based approach* (Jarvis & Crossley, 2012) for the analysis of transfer effects or first language-dependent key structures in second language use with *key-structure analysis* (Ivaska & Siitonen, 2017). The creation text analysis tools like *Coh-Metrix* (McNamara & Graesser, 2012)*, TAACO* (Crossley et al., 2016)*, TAALES* (Kyle & Crossley, 2015) or *CTAP* (Chen & Meurers, 2016) that automatically extract linguistic features of text complexity, readability, cohesion, and other quality-related metrics additionally furthered these developments. The features proved to be useful over various studies and informed not only native language identification and the interpretation of transfer effects. Next to other stylometry features and different customizations of n-gram features, they also serve to *model writing proficiency* (e.g. Pilán, Volodina, et al., 2016; Present-Thomas et al., 2013), discriminate different types of non-native and native varieties (Crossley & McNamara, 2009; Ivaska et al., 2018; Rabinovich et al., 2016)[61] and grade student essays automatically (i.e. *automated essay scoring)* (Crossley, Roscoe, et al., 2014; Vajjala, 2018).

Genre discrimination and register studies

Genre discrimination or *register studies* are slightly different from the above-stated problems, as they do not pivot around the author but on other stylistic aspects of the text. Still the domain is closely related to the authorship attribution and profiling problems discussed above. As both frameworks focus on situational matters of style of a text instead of the actual content, advances in both frameworks have been mutually enabling each other (Stamatatos et al., 2000). This was facilitated also by scholars who researched into the transferability, or so-called universality of stylistic features over different tasks (Stamatatos, 2016; Van Halteren et al., 2005). First studies in genre and register analysis tried to solve impeding problems of complexity posed by this task with traditional quantitative analysis methods (e.g. Biber, 1993b) and his multidimensional analysis. However, especially because of the inherent complexity of the task, supervised machine learning methods also have a long tradition in this area (Finn & Kushmerick, 2006; Karlgren, 2004; Kessler et al., 1997). Teich and Frankhauser (2009), for example, analysed different scientific registers using data mining (see also Teich et al., 2015). (Degaetano-Ortlieb, 2015) analysed evaluative registers under the macro and micro-analysis framework (Jockers, 2013) widely diffused in the digital humanities (cf. section 4.1). (Lin et al., 2009) classified genre in online discussion threads in an educational setting under the framework of text mining. Similarly, the linguistic particularities of one language variety, namely translated texts, have been analysed using a machine learning approach, i.e. training and interpreting successful text classification models that can distinguish between translated and original texts (machine learning approach) e.g. by (Baroni & Bernardini, 2006; Ilisei, 2012; Volansky et al., 2015).[62]

Topic modelling, semantic analysis and text summarization

Topic modelling, semantic analysis and text summarization are trying to extract the abstract topic, key words and main concepts of a text. A topic model is a statistical model that tries to extract the prevalent topic, in the form of lists of clustered key words of a text or a collection of documents (Blei & Lafferty, 2009; Steyvers & Griffiths, 2007). There are different methods used for topic modelling, such as *non-negative matrix factorization*, l*atent semantic analysis* or *indexing* (LSA), and *latent dirichlet allocation* (LDA). The most popular and efficient among these is probably the LDA algorithm by (Blei et al., 2003), which assumes that each document has been created by a mixture of topics that are

[61] For other similar studies using n-grams and key-structures see Ivaska et al. (2017) or Rabinovich et al. (2016).
[62] For analysis on different learner varieties see above.

made up by a mixture of words assigned to each abstract topic through its probability distribution of co-occurrence over the corpus. Topic modelling is usually a key step for further processing or mining tasks. It can help to select topic-specific features, to enhance recommender systems that suggest further articles for a specific reader group (Wang & Blei, 2011), and facilitate trend mining, semantic analysis and text summarization (Barua et al., 2014; Chen, 2006).

Sentiment analysis and opinion mining

A semantic text mining task with a rather long history in NLP is *sentiment analysis* often also called *opinion mining*. In sentiment analysis texts are classified according to their polarity as either positive or negative (sometimes also as neutral) (Wilson et al., 2005). Next to early rule-based approaches, also called *lexicon-based approaches* (Taboada et al., 2011; Turney, 2002) that calculate the class depending on the ratio of positive and negative words considering negation particles and other moderators, machine learning models based on frequency-based NLP features like word vectors and semantic similarity-based word embeddings (Maas et al., 2011; Pang et al., 2002) are used for this task. Sentiment analysis is not only important for commercial applications like market analysis, customer service or brand monitoring but is also important as a preprocessing task that can be used as input for other text classification problems. Later studies extended the task to detecting not only the sentiment, i.e. polarity, but also the general opinion including the opinion holder as well as the subject or object being talked about (Kim & Hovy, 2006; Wiegand et al., 2016).

Argument mining

Similar to the broader version of sentiment analysis and opinion mining, which entails a whole analysis of opinions, involved actants and aspects, *argument mining* or *argumentation mining* (Lippi & Torroni, 2016; Peldszus & Stede, 2013) tries to identify argumentative structures like conclusions, premises, or pro- and counter arguments including their interrelations in free text. It is a rather recent, rapidly evolving research area and has already been applied to a diverse set of genres and text types (e.g. legal documents, product reviews, student essays, tweets or academic literature) in order to e.g. facilitate semantic search in the internet (Wachsmuth et al., 2017), evaluate student understanding based on their written essays (Persing & Ng, 2015) or for conversational search (Ida et al., 2019).

Event detection, trend mining

Other semantically oriented text mining tasks are concerned with the detection and extraction of events within a stream of text, e.g. log files, twitter streams, business logs or other data that can be aligned along a time axis (Dong & Li, 1999; Radinsky & Horvitz, 2013). While *trend mining* and is often used in industrial settings to monitor and customer reactions (Lazard et al., 2016; Ma et al., 2013), predict stock developments, find new business possibilities, or observe upcoming trends on the market, *event detection* tries to find, classify and relate events described in text datasets and is used e.g. in biomedical text mining (Byrd et al., 2016) or public security to monitor natural disasters or predict riots from Twitter streams (Burnap et al., 2015; Cresci et al., 2015; Singh, Dwivedi, et al., 2017).

Relevance prediction, recommender systems and review helpfulness prediction

Text mining is further used to predict the relevance of a text (or objects that are referred to in texts) and rank them compared to other texts. Such ranking approaches are used for example for content-based recommender systems (Lops et al., 2011), search engines and information retrieval (Croft & Lafferty, 2003) (e.g. Chinkina & Meurers, 2016; Weiss et al., 2018) or to rank user reviews, forum answers, news feeds or other social media texts (Dalip et al., 2013; Krishnamoorthy, 2015; Singh, Irani, et al., 2017).

Relevance prediction is often informed by crowdsourced information (Cao et al., 2011) of the readers themselves, feeding previous reactions to similar contents into the training data of a predictive system. It thus often bases the predictions on semantic, similarity-based text-features, sometimes also with stylistic features. Furthermore, these systems often have to deal with a high amount of multi-dimensional data, making the interpretation of such complex models rather difficult.

Prediction of virality, persuasiveness or user engagement

Another strand of research that can be approached with semantic as well as stylistic features is the *prediction of text virality, persuasiveness or engagement*. This text mining task is used for marketing purposes as well as in media sciences and applied linguistics to infer the elements of engaging communication (Guerini et al., 2012; Jaech et al., 2015; Tan et al., 2014; Wei et al., 2016).

Spam detection and detection of hate speech or other offensive, deceptive or abusive text

Although spam filters originally often based on manually curated sets of rules, keywords and blacklists for spammers, the fast-paced changes in spam make the use of machine learning particularly attractive. Newly proposed techniques for detecting spam use complex neural networks, genetic algorithms and extensive feature engineering to identify unwanted emails, twitter messages, weblinks or reviews (Faris et al., 2019; Wang et al., 2015; Wu et al., 2018). Moreover, abusive, offensive or deceptive texts (e.g. in social media) are tackled with complex classification models trained on big data (Davidson et al., 2017; Gröndahl & Asokan, 2019; Nobata et al., 2016; Zhang et al., 2014).

Social media analysis

The purpose of *social media analysis* is to monitor social media activities in order to predict trends, mine current opinions and find key topics or users (Phillips et al., 2017). It thus draws from different other tasks and refines them specifically for the social media environment. It is used especially for *business intelligence, market analyses and competitor analyses* but also for public security and administration. Social media analysis is furthermore often seen as a dynamic process that monitors streams of data and is thereby related to trend mining and event detection. More than in traditional research settings, this industry-driven approach aims to refine the model continuously with new data, instead of using and inspecting it for a non-recurring analysis[63].

4.3.1.2 Common concepts and operationalizations

In general, most text mining approaches used in the past have focused on the prediction of concrete and verifiable concepts like a text's author, his or her age, gender, first language, language competence, the polarity of a text's sentiment, the classification of a text in terms of appropriate and valuable or non-valuable content (e.g. spam, abusive language), or the text's user engagement or propagation (likes, clicks, retweets, etc.). The respective variable of interest is usually gathered via questionnaires, trained annotators, or other existing metadata and used as class label for the subsequent prediction tasks. However, in recent research, more complex concepts have also been approached using text mining techniques. This includes for example the prediction of personality traits of the author, the opinion transmitted in the text or its quality or persuasiveness, as well as irony or sarcasm. Moreover, especially through in-depth analysis in social sciences and humanities traditional clear-cut distinctions for previously used "easy" classification tasks have been questioned. Studies showed that allegedly verifiable categories like gender or age can be complex (social) constructs that depend on various factors. Hence, the analysis of linguistic phenomena related to those constructs should also

[63] But see Lipizzi et al. (2015) for research-driven analysis on conversational patterns in newly launched product reactions on Twitter.

be open to alternative operationalizations (Hovy & Spruit, 2016; Nguyen et al., 2014). Table 1 and Table 2 give examples on how concepts are commonly operationalized in classification or regression tasks of text mining.

Classification problems

Task	**Predicted concept**	**Operationalization**
Gender prediction	Author gender	Biological sex inferred from questionnaire or user profiles
Age prediction	Author age	Age group or generation according to defined splits of chronological age
Native language identification	Native language of the author	First language of author according to questionnaire or test group
Spam detection	Spam or not spam	Spam or valuable text according to manual classification
Language proficiency prediction	Grade or competence level of writer	Grade or competence level according to manual classification, grade level or test group
Sentiment analysis or opinion mining	Polarity of text (positive/neutral/negative)	Polarity according to manual classification

Table 1: Text mining tasks, their predicted concepts and possible operationalizations for classification problems.

Regression problems

Task	**Predicted concept (class)**	**Operationalization**
Age prediction	Author age	Numerical age in years according to questionnaire data
Automatic essay scoring	Test score	Test score according to trained annotators and scoring rubric
Review helpfulness prediction	Helpfulness score	User recommendations
Virality prediction	Virality	Message propagation in numbers of retweets, forwards, etc.
Business intelligence	Success	Number of sales, transactions, clicks,
Other	Probabilities for categorical classes	Percent probability for individual classes

Table 2: Text mining tasks, their predicted concepts and possible operationalizations for regression problems.

4.3.1.3 Common feature types in text mining

The main feature types commonly used to create feature sets for predictive modelling in text mining are (shallow) word frequency features, semantic representations, specifically designed linguistic features, metadata features and multi-modal features.

Word frequency features

Word frequency features are usually represented in a frequency matrix or so-called word vector and contain (raw or normalized) frequencies for e.g. words, all words except from a list of so-called stop words or words from a specific dictionary. Frequencies for word categories, constituents or annotations are also frequently used. The word vectors are usually extracted from manually or semi-automatically crafted frequency tables or computational linguistic feature extraction methods. This category comprises common variables typically used in corpus linguistics like frequency lists of words or other linguistic phenomena of interest as well as classical or advanced NLP features (e.g. bag-of-words features, term-frequency inverse document frequency (TF-IDF), n-grams, skip-grams or recurring n-grams, sparse matrices, or suffix arrays).

Semantic representations

Typical semantic features used in text mining are for example document keywords, distance or similarity measures, topics extracted via topic modelling or semantic analysis, distributional compositional semantics (Lenci, 2008), word embeddings (Mikolov et al., 2013), or features from sentiment analysis, named entity recognition or argument mining.

Linguistic features designed with specific research intention

Over the last years a high number of conventionalized linguistic features, also called indices or measures have been established to measure certain linguistic phenomena that are expected to be of value for the predictive models in text mining. Furthermore, various studies addressed the general transferability of these features, probing them on different tasks, corpora or genres (Cimino et al., 2013; Stamatatos, 2016; Zesch et al., 2015). These features include:

- stylometry features, e.g. word lists for function words that are proven to be relatively independent from topic but dependent on writing style (HaCohen-Kerner et al., 2010)
- specific dictionaries or word lists: e.g. slang words, emotional words, etc. (Özyirmidokuz, 2014; Rustagi et al., 2009)
- readability indices (Crossley et al., 2008; Hancke et al., 2012)
- indices of cohesion or linguistic complexity (Crossley et al., 2013; Weiß, 2017)
- register-specific features, such as evaluative patterns (Degaetano-Ortlieb, 2015)
- linguistic category model, measuring abstractness vs. concreteness by the ratio of adjectives, state verbs and action verbs (Krishnamoorthy, 2015)

Metadata features

Possible features for text classification can also be retrieved from text metadata like socio-demographic data of the author (e.g. age or first language) or other context-related characteristics of the text (e.g. time stamps for online texts, reaction counts, etc.). These features, although not of linguistic nature can significantly enhance the predictive model, by accounting for certain biases in the data (Hovy, 2015).

Multi-modal features

Latest studies also take multi-modal features into account, which allow to consider not only text, but also possible imagery or videos that accompany the text. For some recent approaches see for example Kiela (2016) or Rangel et al. (Rangel et al., 2018).

4.3.2 Dataset: Extracting features and preparing feature sets

Once the modelling task, target variable and predictor variables are defined, the variables need to be extracted from the corpus and prepared in order to suit the modelling task chosen. The first step is to retrieve text features through frequency lists, metadata annotations, NLP tools and is usually referred to as feature extraction. After that the bare extracted features often still need to be transformed (feature transformation) or filtered (feature selection) in order to be useful for the modelling task. All these steps together are usually summarized as *feature engineering* and can be iteratively refined to improve the prediction models (see e.g. Iria, 2012).

Below, methods and tools for feature extraction, feature transformation and feature selection are introduced briefly.

4.3.2.1 Feature extraction

The act of retrieving the predictor variables from the corpus is called *feature extraction*. After the predictor variables of interest (the so-called features) have been defined and possible operationalizations found, these features need to be extracted from the corpus or from related metadata. For most machine learning methods, the preparation of *feature vectors*, i.e. a list of observations of the predictor variables and the corresponding target variable in matrix format, is needed (compare also the concept of word vectors from section 4.3.1.3). The feature vectors can be hand-crafted using for example frequency lists for specific linguistic phenomena retrieved from corpus software and query tools, but also tools and methods from NLP or specific text analysis tools are frequently used to extract features. Some of the more advanced feature extraction tools need very little background knowledge and/or manual intervention. In the best case, these feature extraction tools can also provide ready-made feature matrixes that can be used directly with the software or tools used for training the models. Below, examples for frequently used methods and tools are listed, focusing on freely available/open source options.

Basic corpus linguistic tools: Keywords, frequencies and collocations

Commonly used corpus linguistic features like keywords, frequencies and collocations can typically be extracted via corpus linguistic software such as the free software tools AntConc (or related Ant packages[64]) and WordCruncher[65] or commercial options like SketchEngine[66] or WordSmithTools[67].

NLP tools: N-grams, word frequency features and semantic features

While word n-grams, also called lexical bundles, chains, wordgrams or clusters, can usually be extracted with classical corpus software as well[68], other NLP word frequency features like bag-of-words representations, TF-IDF features or character n-grams are usually extracted programmatically with NLP tools. Popular open source NLP tools and software frameworks for these purposes are for example the NLTK Python library (Loper & Bird, 2002), Stanford CoreNLP Java Toolkit (Manning et al., 2014) or Apache OpenNLP (Java)[69]. Furthermore, there is a number of newer, less general tools and toolkits that provide extraction methods for more advanced NLP features like skip-grams, i.e. n-grams where some words in between a pattern can be skipped in order to make them more general (e.g. with Colibri

[64] http://www.laurenceanthony.net/software.html
[65] http://www.wordcruncher.com/
[66] https://www.sketchengine.eu/
[67] https://www.lexically.net/wordsmith/
[68] E.g. with *AntConc*, *kfNgram* (http://www.kwicfinder.com/kfNgram/kfNgramHelp.html), or *SketchEngine*.
[69] https://opennlp.apache.org/

Core by Van Gompel & Van Den Bosch, 2016); simple distance or similarity measures (e.g. with SpaCy[70]), word embeddings, i.e. numeric vector representations that encode the semantic similarity of words in the text (e.g. word2vec by Mikolov et al., 2013; or GloVE by Pennington et al., 2014), topic modelling (e.g. with MALLET[71]) or sentiment analysis (e.g. with VADER by Hutto & Gilbert, 2014; or TextBlob by Loria et al., 2014).

However, although NLP tools or corpus software can automatically detect and annotate certain linguistic phenomena or text characteristics and count frequencies or output labels, often the data is not automatically transformed into a standardized feature matrix format that can be imported directly into other machine learning or data mining application. Recently dedicated feature extraction tools have emerged that combine the functionality of NLP tools and data conversion into interoperable formats for the use in subsequent machine learning tools in order to facilitate easier and faster feature extraction (e.g. EDISON by Sammons et al., 2016).

Text analysis tools: Automatic extraction of special-purpose linguistic features

Lastly, there are also text analysis tools that usually combine the functionality of NLP tools with linguistic theory and experience in text mining to automatically retrieve linguistic features for texts. These features can be straightforward, countable linguistic phenomena like text length or the type-token-ratio as well as highly abstract operationalizations for concepts like cognitive load (cf. Weiß, 2017) or readability (e.g. Flesh-index for English, or the Gulpease index for Italian). Popular tools for the extraction of such interest-driven linguistic features are for example the proprietary tool LIWC (Pennebaker et al., 2001), Coh-Metrix for the analysis of cohesion and coherence (McNamara & Graesser, 2012) or Kristopher Kyle's NLP tools for the social sciences containing tools for the analysis of lexical sophistication or diversity, cohesion, or lexical and syntactic complexity (cf. Crossley et al., 2016; Kyle & Crossley, 2015). However, most of these tools are made for the analysis of English and only a few tools exist for other languages.[72] But see Dell'Orletta et al. (2011) or Pilán et al. (2016) for tools in other languages.

4.3.2.2 Feature transformation

Apart from feature extraction, feature engineering also includes all types of transformations of variables that can help the modelling process (Khurana, Turaga, et al., 2016). By changing the representation of a particular variable (e.g. from categorical into binary or one-hot-encoded format, from count frequencies to log-frequencies or from individual variables to grouped components) the learning algorithm can benefit from the additional information and/or reduced noise, redundancy, or skewedness in distribution. By that, feature transformations can significantly affect model performance. Below, some examples of common techniques for feature transformations are given.

- *Aggregating or splitting features*
  Features with equivalent measurement units can be aggregated or split to represent data better, e.g. differentiate mixed contributions or join the combined informational value of various sparse features (i.e. features with little information).
- *One-hot-encoding for categorical values*

[70] https://spacy.io/
[71] http://mallet.cs.umass.edu/
[72] The Common Text Analysis Platform CTAP (Chen & Meurers, 2016) for example provides linguistic complexity and cohesion measures for English and German (Galasso, 2014; Weiß, 2017).

One-hot-encoding, i.e. to code every possible category of a categorical feature as a Boolean variable, can remove complexity from the dataset and furthermore allows to use methods that cannot naturally deal with categorical values.

- *Log-transformations*
  Transforming variables with log-functions or similar can smooth non-normally distributed data, which in turn helps most learning algorithms.
- *Discretization of continuous variables*
  Continuous variables can be simplified by discretizing them into integer or even binned numerical values.
- *Binning discrete variables*
  Besides the discretization of variables, it can sometimes be beneficial to bin numerical values into sensical groups to reduce noise and prevent overfitting.
- *Conflating variable categories*
  Categorical variables can often be reformulated in order to reduce the number of categories. This is especially useful if there are high imbalances between categories.
- *Normalization of variables*
  Known biases in the variables can be accounted for by using relative, normalized values (e.g. occurrences per sentence, etc.)
- *Standardization or rescaling*
  Standardization, i.e. transforming the values of variable in a way that all variables have the same scale (all variables have similar ranges, means and standard deviations, e.g. all values are between 0 and 1, -1 and 1 etc.), can help to make variables (and their variance) comparable.[73] This is useful for the performance of some algorithms as well as for the interpretation of variable importances.
- *Exclusion or capping of outlier values*
  A detailed univariate and bivariate outlier analysis can help to identify erroneous measurements or outliers that should be excluded, as well as extraordinarily high or low values that could possibly be capped to reduce data skewedness.
- *Recoding ordinal data*
  Sometimes, qualitative categorical variables can be ordered according to some inherent logic and thus presented as ordinal variable instead equal categories.
- *Dimensionality reduction methods*
  Dimensionality reduction methods like principal component analysis or singular value decomposition can combine features and reduce the number of features in the feature set.
- *Missing values treatment*
  The way missing values are treated in the process, i.e. whether they are ignored, excluded together with the whole observation or imputed (approximated) also can have substantial effect.

4.3.2.3 Feature selection

Another frequent strategy when preparing the variables for data mining is the preselection of variables that shall be used for model training. *Feature selection methods*, also called *attribute selection* or *variable selection methods*, automatically select a subset of features from a bigger feature set, based on criteria of assumed relevance for the modelling task. Feature selection is usually seen as a preprocessing step of machine learning in order to reduce noise and complexity. This in turn can improve

[73] Min-max scaling or z-score standardization are common techniques to achieve such comparable variables.

prediction performance and reduce both training time and the chance of overfitting (Guyon & Elisseeff, 2003). There are three general categories of feature selection methods.

- *Filter methods*
  Filter methods create a ranked list of features, of which a custom amount (or percentage) features can be chosen for training.
- *Wrapper methods*
  Wrapper methods systematically test possible subsets of the feature set for their performance in a predictive model. They then return the subset with the best performance on the prediction task.
- *Embedded methods*
  Embedded methods use filter or wrapper methods but embed them in the concrete modelling task.

Filter methods are often based on univariate association measures or information theoretic entropy or distance measures like correlation coefficients, chi-squared test statistics, information gain metric, (pointwise) mutual information, the *gini*-index or weights from the popular *ReliefF* algorithm. The used measures are usually fast to compute, making the method scalable for big data. Furthermore, lists of ranked features can give insights into relationships in the data, before the actual training of a predictive model is done. They thus provide a method for preliminary interpretation of variable importances, especially when the outcome of various feature selection methods is compared. However, many of these methods come with their individual shortcomings (e.g. correlation coefficients that are biased by confounders) including that combinatorial effects are ignored in such univariate methods. Furthermore, most of the measures naturally give similar weights to redundant features, which might result in the selection of those redundant features although they don't provide any new information for the classifier or regression model (Guyon & Elisseeff, 2003). The measures for feature selection also can be chosen independently from the algorithm that is used in the classification or regression model. However, there are measures that are better suited for some algorithms than for others.

For *wrapper methods* there are different approaches depending on how the subsets for testing are chosen. As an exhaustive search method that tests all possible subsets is not sustainable if a high number of features is involved (Guyon & Elisseeff, 2003)[74], most modern approaches use methods to reduce the number of subsets and reduce computing time. Although they can lead to misleading local optima, such simplifying procedures provide computationally feasible and still relatively good results. The main two approaches can be categorized into sequential selection methods and heuristic search algorithms (Chandrashekar & Sahin, 2014). Sequential selection methods iteratively choose and test subsets of features and compare the performance achieved with those features. In forward-selection, the algorithm starts with a null model without any features and iteratively adds features to find a final model. Each step of the iterative process identifies the variable with the highest increase in performance and adds it to the model. In backward selection, the procedure starts with a full model and removes variables step-by-step, excluding each time the variable that results in the highest model performance decrease. Additionally, adaptations of these two main approaches try to deal with known restrictions of the procedure. Sequential floating forward selection (Chandrashekar & Sahin, 2014) (see also recursive feature elimination), for example, combines both methods either adding or removing at every step depending on the resulting changes in model performance. Heuristic search algorithms on the other hand are algorithmic solutions for optimization problems. Examples for such

74

methods are genetic algorithms (inspired by natural selection), simulated annealing or particle swarm optimization.

Wrapper methods can find feature subsets that are specifically good for the algorithm of choice and also allow to account for combinatorial feature effects (contrary to bivariate measures for filter methods). They can therefore be more precise in their estimate of feature importance than independent association measures of filter methods can be. They can also identify and sort out redundant features that do not contribute to the model performance. However, the approach is computationally very expensive or – in case a specific procedure is used to narrow down the subset choices (e.g. backward or forward selection processes) – might only lead to local optima. Additionally, the identified best subset of features is only valid for the algorithm and setup tested that was used for feature selection (see embedded methods for wrappers that are integrated directly in the modelling task). The resulting feature importances might therefore not coincide with the ones reported for different setups, meaning that feature selection with wrapper methods is not necessarily generalizable to other scenarios. It might lead to overfitting on the specific task, data or algorithm and lead to non-generalizable conclusions when interpretation is done solely through the selected features.

In *embedded methods* the subset selection is not done preliminary before the actual model training (e.g. with a different algorithm or a non-model-specific association measure) but directly during model training by using of the exact algorithm and configurations for the task combined with some optimization approach. Common types of embedded feature selection are for example regularization (or penalization) methods for regression models like *LASSO, Elastic Net* or *Ridge Regression* (see section 3.5.1). The selected variables of embedded methods are strongly dependent on the learning algorithm used. However, the methods are usually less complex and computationally expensive as wrapper methods and can also account for combinatorial effects.

Most data mining and machine learning frameworks provide in-built options for feature selection. For WEKA a wrapper method (`WrapperSubsetEval`) as well as ranking (i.e. filter) methods based on correlation (`CfsSubsetEval` or `CorrelationAttributeEval`), information gain (`InfoGainAttributeEval`), gain ratio (`GainRatioAttributeEval`), etc. are provided (Witten et al., 2016). For R the popular *caret* package for machine learning provides *univariate filters* as well as different wrapper methods including *recursive feature elimination* (i.e. backward selection), *genetic algorithms* and *simulated annealing* (Kuhn, 2015). The common machine learning library scikit-learn for Python provides *univariate feature selection* (i.e. filter methods) based on correlation or chi-squared test statistics, mutual information or f test measures for classification performance increase, next to *recursive feature elimination* (backwards selection wrapper) and embedded methods like L1 LASSO regularization (`SelectFromModel`).

4.3.2.4 Automated feature engineering

Because of the proportionally high amount of time spent on feature engineering, the practical engineering skills needed and its tendency to introduce errors, feature engineering is one of the bottlenecks in data mining and machine learning (Khurana, Nargesian, et al., 2016). Recent approaches try to reduce time and effort spent on feature engineering by automating processes like database operations that retrieve feature matrices for different statistical units or the search for new or aggregated features (Kanter & Veeramachaneni, 2015; Katz et al., 2017; Khurana, Turaga, et al., 2016; Lam et al., 2017). The open source Python framework *Featuretools*[75] for automated feature engineering for

[75] https://www.featuretools.com/

example, promises to provide interpretable features while automatically searching through possible feature representations (cf. *Deep Feature Synthesis* by Kanter and Veeramachaneni, 2015).

4.3.3 Analysis: Model building and evaluation

When the target variable and feature sets for the modelling task are prepared the actual model building includes training one or various, systematically chosen classification, regression or clustering models that can be evaluated and interpreted subsequently. In most text mining studies it is common practice to train not only one type of model but to use a set of different learning algorithms and most often also different parameter configurations for these learning algorithms for each modelling task. This allows to get a more robust estimate of the learnability of a task and makes it possible to identify and account for the shortcomings of particular model setups. For that reason, typical choices for learning algorithms and parameters usually include different types of complementing methods that are known to vary in their strengths and weaknesses. Depending on the aim of the study, focus is set on either simpler, more interpretable methods or more complex methods that are known for their high predictive performance from other studies. While prediction-oriented studies in NLP and machine learning research tend to favour highly customized, complex, state-of-the-art methods (e.g. *deep neural network architectures*, *multi-task learning*, *generative adversarial networks*), most studies with a focus on interpretation make use of more comparable, standard methods for simpler, intrinsically interpretable models, or already established black box models (e.g. *SVMs*, *random forest models*). These so-called *vanilla* models are trained by utilizing ready-made implementations of learning algorithms with their standard configurations or with minimal changes and simulate the general learnability of the task instead of perfect prediction.

Below some popular software tools and libraries for data mining and text mining are presented, giving examples for tools with graphical user interfaces as well as libraries for scientific and statistical programming languages. Afterwards, the motive and methods of model evaluation including some strategies for model comparison are discussed.

4.3.3.1 Software and tools for data mining and text mining

Having named popular and free data science tools for feature extraction, feature engineering and feature selection in the previous sections, this section introduces tools for predictive modelling and machine learning, i.e. those tools that provide methods for the analysis[76] and evaluation step in data mining. It focuses on currently supported, frequently used tools that are freely available (for research purposes). Depending on their origin and focus, they differ in their usability, intuitiveness and power as well as in the methods and additional features that they implement. Some of the tools actually provide full data mining frameworks that also incorporate methods for feature extraction, feature engineering and feature selection, as well as visualization (and sometimes even interpretation methods)[77]. However, this section focuses on predictive modelling and machine learning and therefore also names popular packages, libraries and other software providing just machine learning implementations that can be used for data mining and text mining in the respective programming language of choice.

[76] Although tools for feature extraction, feature engineering and feature selection as well as for visualization and interpretation could also be (and are indeed often) named as (text) analysis or text mining tools, they actually present other steps of the data mining process and shall thus be excluded here. For other data science tools related to different steps compare the related sections.

[77] These are for example *WEKA*, *RapidMiner*, or *KNIME* that provide graphical user interfaces, but also the packages *caret* or *mlr* for the statistical programming language R or *scikit-learn* for Python.

In general, we can divide the possible options into tools with graphical user interfaces that can be used without any programming skills and tools that need at least limited knowledge in the programming language that they are based on.

Data mining software with graphical user interface

The data mining tools WEKA, RapidMiner, Orange or KNIME provide graphical user interfaces, and offer also other preprocessing and visualization methods besides the actual classification, regression or clustering methods provided. Moreover, there are additional plug-ins or add-ons for each of them that provide preprocessing methods, feature extraction, transformation and selection methods, specifically built for (raw) text data (i.e. text mining plug-ins).

WEKA

WEKA is probably the most widely known software for data mining and machine learning and is particularly popular in the scientific community. The acronym WEKA means "Waikato Environment for Knowledge Analysis" and refers to the University of Waikato in Hamilton New Zealand where it was developed. The accompanying introductory book on data mining and machine learning "Data Mining: Practical Machine Learning Tools and Techniques" (Witten et al., 2016) has had various editions over the last years (the last being released in 2016) and is one of the standard references used by many scholars. Next to the main graphical user interface "Explorer" that allows to load datasets, transform and select variables, investigate univariate and bivariate distributions and perform classification, regression, clustering or association mining and interactive visualization of the results, it also offers and "Experimenter" interface that allows to statistically compare results for different modelling attempts. Furthermore, the WEKA implementations can also be used without the graphical user interface directly via Java or one of the adaptations for python or R.

WEKA provides a multitude of classification, regression, clustering and feature selection methods off the shelf. Further, newer or less well-known methods can be added via plug-ins, provided by the originators or the interested open source community. Figure 15 shows the Explorer interface of WEKA with a simple dataset loaded.

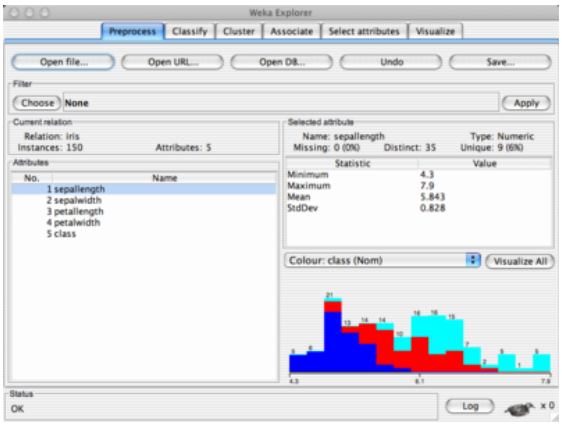


Figure 15: The WEKA Explorer interface.[78]

RapidMiner

Similar to WEKA, RapidMiner is a fully equipped data mining software. The software that originated in the rapid prototyping project YALE (Yet another learning environment) by Mierswa et al. (2006) is now

[78] Image Source: https://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html

widely used for data mining in commercial settings. However, there is still a free plan for university members to use the software in research settings[79].

The graphical user interface for RapidMiner is slightly less intuitive than WEKA from a research perspective, but also more flexible in terms of how the individual components are combined together. Interactive visualizations furthermore enhance and ease data exploration steps.

RapidMiner implements a vast amount of learning algorithms, data exploration methods, feature extraction, transformation and selections methods and other data mining related tasks. It even claims to have implemented all algorithms provided by WEKA. Figure 16 shows a screenshot of the graphical user interface ("*Visual Workflow Designer*") in RapidMiner.

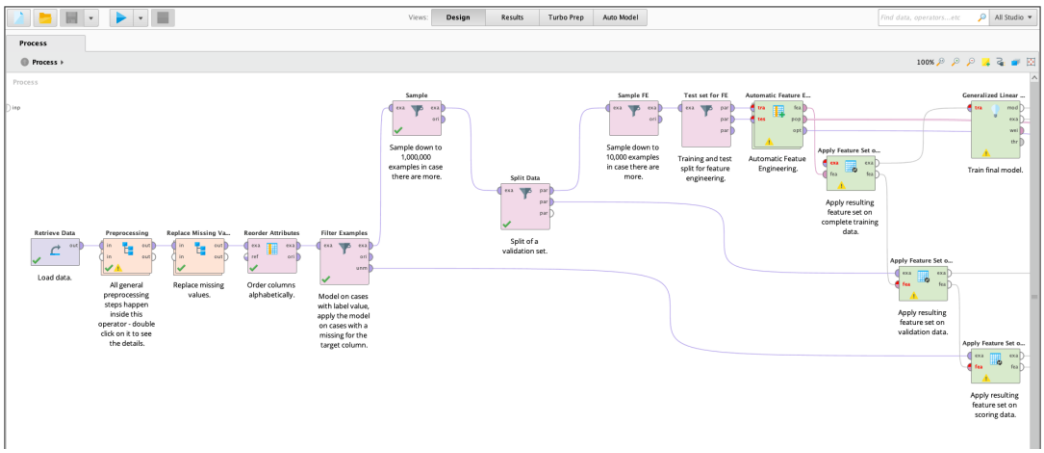


Figure 16: Combining individual tasks and processes in the RapidMiner Visual Workflow Designer.[80]

KNIME

The KNIME Analytics Platform[81], short for "*Konstanz Information Miner*", is especially known in the computer-science-driven knowledge discovery and data mining community. It is a powerful, comprehensive tool that also requires a certain amount of background knowledge and experience (next to substantial resources in terms of memory and computation power). The interface is similar to the one of RapidMiner and displays individual components that are plugged together to create a data mining pipeline (cf. Figure 17). The software provides all the main functions for any data mining task, including the most popular algorithms for classification, regression, clustering and outlier analysis.

[79] https://rapidminer.com/educational-program/

[80] Image Source: https://rapidminer.com/

[81] https://www.knime.com/knime-software/knime-analytics-platform

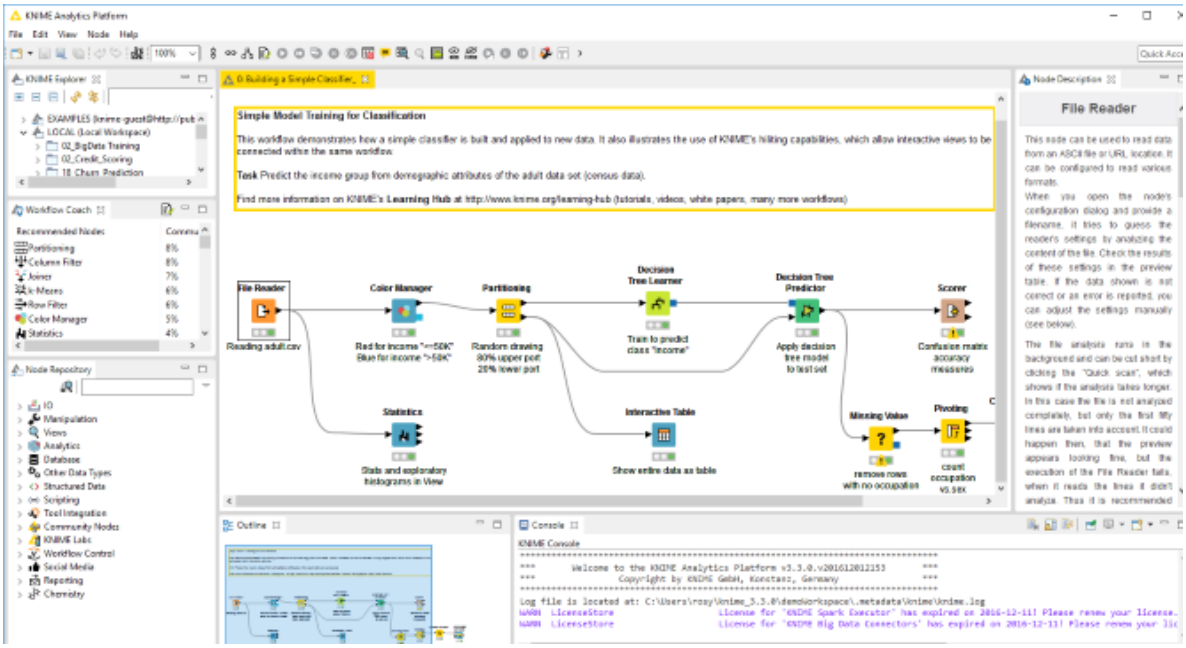


Figure 17: The KNIME Analytics Platform.[82]

Orange

The graphical user interface of the open source data mining software Orange[83] is visibly simpler compared to RapidMiner and KNIME (cf. Figure 18). Orange is based on Python and provides visual programming methods and interactive visualization tools for various data mining tasks including implementation for the main machine learning methods. However, it is not as frequently used in the scientific community as other tools like WEKA or RapidMiner.

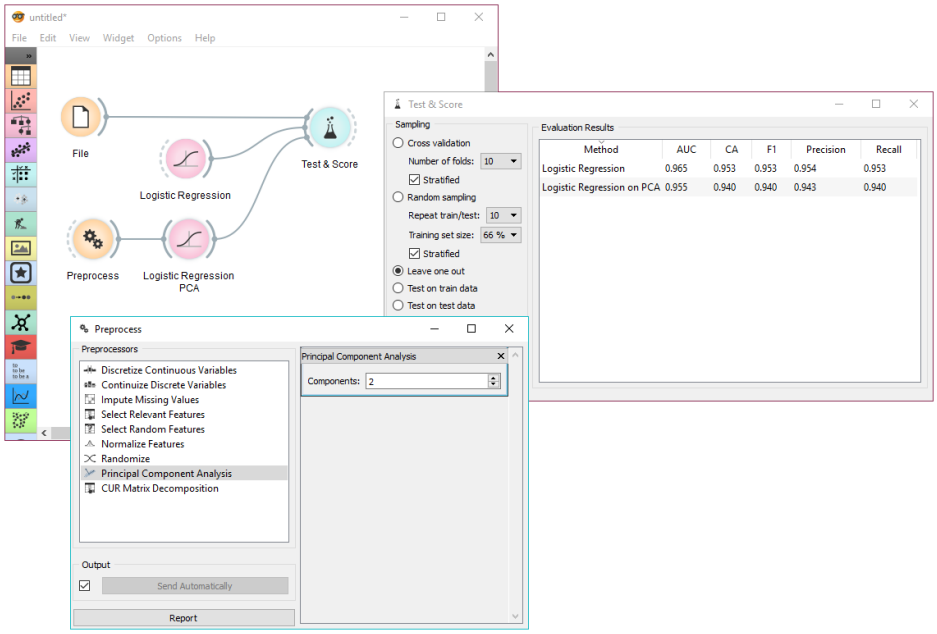


Figure 18: Graphical user interface for the data mining software Orange.

[82] Image Source: https://www.knime.com/knime-introductory-course/chapter1

[83] https://orange.biolab.si/

Machine learning and modelling packages and libraries for programming languages

Apart from software solutions that come with a graphical user interface, there are various software packages or libraries that provide single or various implementations for well-known learning algorithms for common programming languages.

These packages allow to use machine learning methods, without having to re-implement the algorithm. They wrap individual methods into more abstract objects (called estimator, model, classifier or similar), to unify the way individual methods are invoked. That way they offer convenient and easily understandable ways to use machine learning methods even with limited programming skills in the respective language.

Table 3 lists popular packages, modules or libraries for Python, other languages like *Java, Scala, C or C++,* and the statistical software or programming language *R*.

R	**Python**	**Other (various languages)**
caret, mlr, CORElearn, RWeka, rminer, lars, lasso2, glmnet, lme4, nlm, rpart, party, tree, ctree, randomForest, gbm, xgboost, nnet, deepnet, h2o, tensorflow	scikit-learn (sklearn), tensorflow, keras, Theano, PySpark, XGBoost, Statsmodels, Pytorch	WEKA, Apache Mahout, Apache SparkML, Apache, Singa, MALLET, Shogun, Microsoft distributed machine learning toolkit

Table 3: Machine learning packages and libraries for R, Python and other programming languages.

R

Rather than a full programming language R is an open source statistics software that comes with its own programming language. R is very popular for statistics in scientific research and bases on a solid user community that contributes with countless new and up-to-date packages for almost any task related to statistical analysis. Among these there are various R packages that implement methods for predictive modelling and machine learning. R packages like *caret*, *RWeka, mlr* and *CORElearn* provide sets of various standard machine learning methods for classification, regression and clustering. Other packages specialize on individual algorithms or types of models (e.g. *randomForest* or *h2o* for deep learning). Below a short list of current packages for machine learning, ordered by their focus.

- *caret*: implements a vast series of machine learning methods
- *CORElearn*: classification, regression and feature evaluation
- *mlr*: large number of classification and regression techniques
- *RWeka*: interface to use WEKA implementations in R
- *rminer*: interface that combines machine learning methods from several packages
- Regression models:
    - *textir*: inverse regression for text analysis
    - *lars*, *lasso2* and *glmnet*: all provide regularization methods for regression
    - *lme4*, *nlme*: mixed-effects modelling
    - *MASS*: stepAIC function for automatic stepwise model selection
- Decision trees: *rpart, party, tree, ctree*
- Tree ensembles: *randomForest, gbm, xgboost*
- Neural networks: *nnet, deepnet, h2o, tensorflow*

Python

Python slowly developed into one of the main programming languages for scientific programming in general and for data science and natural language processing in particular. Because of its intuitive use and smooth learning curve[84] and an large availability of software packages and resources, it is one of the easiest ways to use machine learning and data science without intensive computer science background.

Many well-known beginner-level and advanced machine learning libraries base on Python (e.g. *scikit-learn*, *tensorflow*, *keras*, *Theano*, *Pytorch, XGBoost*, etc.) and recent research in machine learning makes use of this programming language providing proof-of-concepts or study replication possibilities in open source repositories. Basic skills in Python thus enable one to use the most recent and powerful methods proposed to the research community. Moreover, interactive programming environments (cf. *iPython*[85]) and so-called notebooks (cf. *Jupyter Notebook*[86]) allow to compile code and document in research-appropriate manner.

One library that is particularly noteworthy is *scikit-learn* (or *sklearn*)[87]. It is probably the most well-known and well-integrated machine learning library and provides implementations for most machine learning methods[88], feature extraction and selection as well as basic visualization methods. Although it is does not include implementations for all the newest methods, many other data and text mining libraries (e.g. for interpretation) can integrate *scikit-learn* models.

4.3.3.2 Model evaluation

In general, models are evaluated on their predictive and/or explanatory power.

Predictive power

Predictive power, or generalizability, is usually measured in performance metrics (Japkowicz & Shah, 2011; Sokolova & Lapalme, 2009). Some of the most common performance metrics for predictive power are

- *accuracy*, i.e. the percentage of test instances that have been predicted correctly
- *recall*, i.e. the percentage of instances belonging to one class that were also predicted as such by the model
- *precision,* i.e. the percentage of instances that have been predicted to be of one class and did indeed belong to the class
- *F-score* (also *F1* or *F-measure*), i.e. the harmonic mean of precision and recall
- Specificity (also true negative rate), i.e. the proportion of instances not belonging to a class that were correctly predicted as such
- Sensitivity (also recall or true positive rate), i.e. the proportion of instances of a class that were successfully predicted as such
- *Area under the curve of the receiver operating characteristics curve* (AUC ROC curve), i.e. a measure of the space under the probability curve that displays the true positive rate and the

[84] supported by countless online tutorials
[85] https://ipython.org/
[86] https://jupyter.org/
[87] https://scikit-learn.org/
[88] Advanced deep learning methods, however, are not (yet) available for scikit-learn.

false positive rate on a two-dimensional graph, showing the performance of a classification model for varying thresholds.

In machine learning contexts the predictive performance of a model is usually established on previously unseen data like a held-out test set, or evaluation techniques like *cross-validation or bootstrapping*. This is necessary to ensure that the model does not only represent the training data but also generalizes to unseen data of the same kind.

The simplest approach is to exclude a certain amount (e.g. 20%) of the available data from the training set, in order to use it as a test set for evaluation (i.e. *train-test-set split*). However, this approach is very sensitive to the selected test set and results can vary depending on the distributions in both sets (i.e. a test set can be particularly easy or difficult for the model).

In *cross-validation (CV)*, one of the most common approaches in NLP, the test-train set split is therefore repeated a couple of times. The dataset is first split into a number of equal parts. Each of these parts is then used as a test set, for a model trained on the rest of the parts, so that a performance for each training instances can be extracted. The final performance for the model is then estimated by using the weighted average of the evaluation instances.

Bootstrapping approaches are similar to the cross-validation approach, in the sense that they train and evaluate on a series of different test sets. However, the amount of test runs and the size of the test set does not depend on the splitting criteria used. Instead, for each test run, a new sample of test cases is drawn from the training set, allowing observations to be chosen various times (sampling with replacement). Through this a higher number of test runs can be done, without diminishing test set size. The prediction performance is then estimated as out-of-bag error (OOB), i.e. the prediction error of test instances that were not in the training set. For random forests or boosted decision trees this estimate is already used in the bootstrap aggregation algorithm (bagging) while training the model. Because of the increased number of evaluation instances, bootstrapping methods can also provide confidence intervals that are usually not foreseen in CV.

Explanatory power

In traditional statistics explanatory power is usually measured by goodness-of-fit tests using for example the R-squared measure of a regression model. However, although this measure provides an estimate of how much of the variance of the dependent variable is explained by the model, it does not give any information on generalizability (as it evaluated classification accuracy on the train set instead of prediction accuracy on unseen data) nor does it allow to measure the amount of valuable insights derived from the model.

In machine learning research for explainable artificial intelligence and interpretable machine learning, scholars thus aim at providing measures to operationalize the interpretability and explanatory power of models using human-centred and/or model-based approaches (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Mohseni et al., 2018). *Human-centred evaluation approaches* establish interpretability through the usefulness of the model for an human audience (Lage, Chen, et al., 2018; Lage, Ross, et al., 2018; Narayanan et al., 2018). *Model-based evaluation approaches* try to quantify trade-off between predictive power and model complexity (Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Lipton, 2017; Molnar et al., 2019).

Model comparison

However, it is not enough to estimate the predictive or explanatory power of a model without comparing it to a reference model. In order to claim that a model is actually capable of producing non-random classifications it has to be compared against a baseline model that reflects the behaviour of random data or random predictions. Moreover, one has to test statistically whether the difference between the two models is due to chance[89].

There are different ways to formulate a baseline model for random behaviour. One approach is to produce completely random predictions for the test set and evaluate the model performance with this random data. The other approach is to always predict the output with the highest prediction probability (i.e. the majority class for classification problems or the mean for regression problems) for each instance in the test set and evaluate the resulting performance. This baseline emulates results for unrelated data, while the model is still aware of the base distribution of the dependent variable. This approach, henceforward called the *majority baseline*, is usually the preferred way to estimate a baseline, as it is easier to compute and also tends to give a higher, and thus stricter, value than the purely random baseline.

While comparing a model against a baseline model is necessary to establish its general validity, models are also compared against each other, in order to draw conclusions from the different model setups, identifying the model with the best *fit* (performance on the test set or rarely also on the train set) (Ben-David, 2008; Daelemans & Hoste, 2002; Demšar, 2006; Salzberg, 1997).

The difference between different trained models should in theory also be tested using statistical tests (Demšar, 2006). In order to find a trade-off between interpretability and predictive performance, the principle of *Occam's razor* is applied as selection criterion, choosing the simplest model that is not significantly worse than the others.

Model comparisons are prevalently conducted for different models that were trained on the same data. This guarantees comparability of the nature and distribution of the data that has been used (and thus of the comparability of the tasks). However, sometimes it is necessary to compare models do not base on the exact same data, e.g. when class distributions are different due to different operationalizations or different train-test set splits, but also when the theoretical comparability of two data sources has been established. In order to give standardized estimates for model performance, the complexity of the task has to be evaluated by accounting for differing baselines (e.g. using Cohen's Kappa as a measure of prediction accuracy, given the possible range of accuracy (cf. Ben-David, 2008).

4.3.4 Model interpretation

In order to test if there is any relationship between the variable of interest (the descriptor variable) and the predictor variables that are assumed to be relevant (e.g. because there are theoretical accounts for them). A simple comparison of a trained model with a baseline model for random prediction or majority class prediction can testify the existence of some sort of relationship between the whole feature set and the target variable (see section 4.3.3.2 above). However, the identification of relationships for individual variables as well as the investigation of the type, direction and magnitude of the variable effect is less straightforward and usually depends on the learning algorithm used.

There are a number of interpretation approaches used to interpret specific intrinsically interpretable learning algorithms as well as strategies for interpreting models as a black box (independently of

[89] This could be the case if test and train set split were particularly favourable or particularly difficult.

whether they can or cannot be interpreted intrinsically). Furthermore, post-hoc black box interpretation methods from the field of interpretable machine learning and explainable artificial intelligence (see section 3.6) supply model-agnostic and model-specific tools and metrics for in-depth analysis of model internals. Possible approaches for interpreting predictive modelling in text mining studies are sketched below.

- Model evaluation and error analysis
- Monofactorial analysis and feature selection
- Model comparison and feature engineering
- Interactive visual analysis
- Inspection of intrinsically interpretable models
- Post-hoc black box interpretation of complex models

4.3.4.1 Interpretation through model evaluation and error analysis

The most simple and standard approach is to interpret a machine learning model by evaluating its predictive performance and confirming that it has learned from the data, as well as by investigating the errors the model made. This approach is mainly used for confirmatory settings, where the presence or absence of relationships between features and class variables is the main focus of the analysis. However, model evaluation measures can also help to explore the data (and/or refine the model) by investigating errors.

In regression modelling, regression residuals and residual plots (e.g. the plots in Figure 19) allow to find observations with particularly high or low errors. They can give information about whether model assumptions were met, or the model is in general suited for the relationships present in the data (e.g. linear relationship, independent data points, normally distributed residuals).

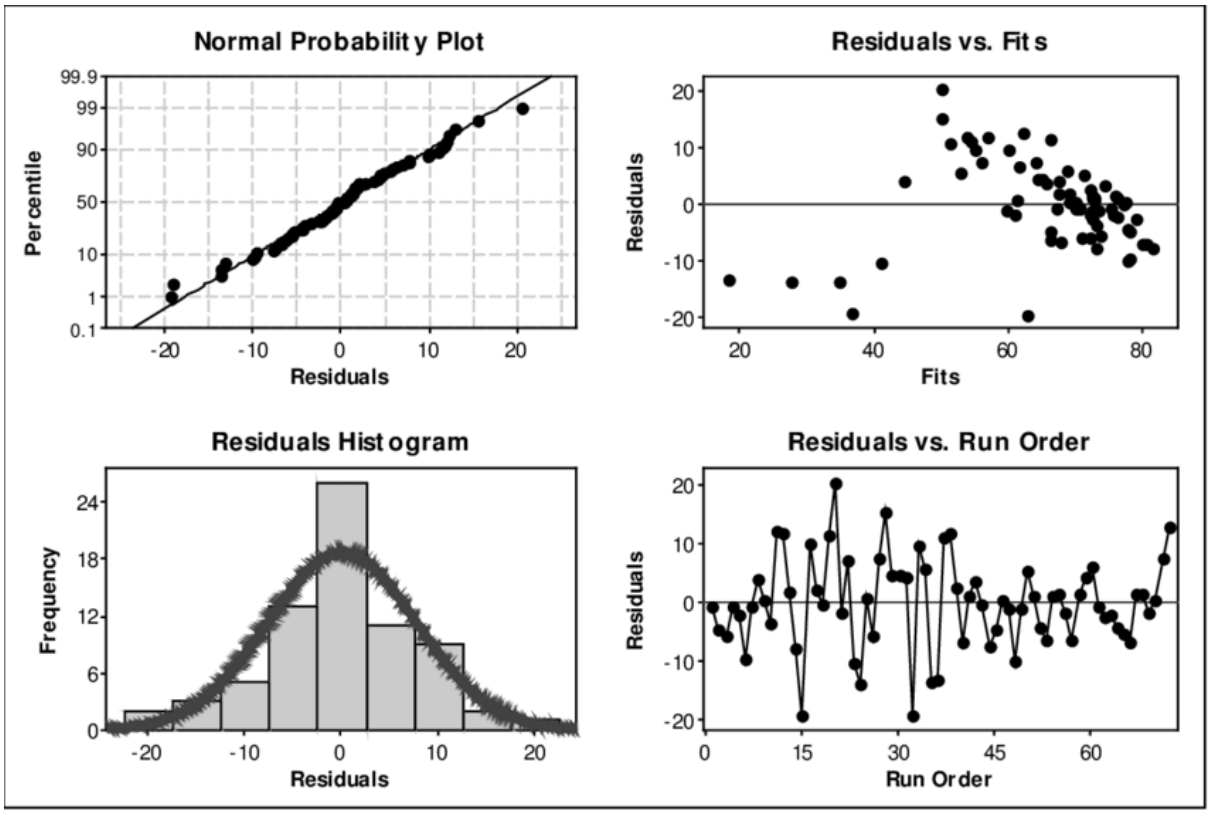


Figure 19:Examples for residuals plots for regression problems (Source: Salazar Aguilar et al., 2006).

For classification models, the confusion matrix shows the numbers (and possibly also instances) that have been classified correctly or incorrectly and indicate the type of misclassification (true label vs. predicted label, cf. Figure 20). This allows to reason over the types of misclassification made by the model as well as to refine the model by accounting for misclassifications with further features or better operationalizations of features or classes[90].

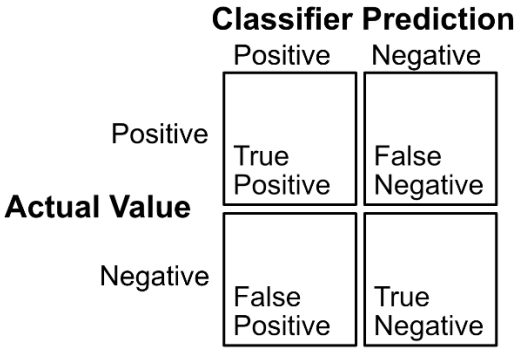


Figure 20: Confusion matrix for classification problems.

4.3.4.2 Interpretation through monofactorial analysis and feature selection (ranking) methods

Although no predictive model is needed in order to perform monofactorial analysis (e.g. correlation analysis or comparison of group means) or to perform feature selection or ranking methods in order to identify relevant features, some studies combine monofactorial methods with predictive modelling by inspecting correlations after a model has been trained, or by investigating the features chosen in feature selection (Flekova et al., 2016; Koppel et al., 2005; Park et al., 2014; Schler et al., 2006; Simaki et al., 2016b; Simaki, Simakis, et al., 2017; Teich et al., 2015; Vajjala, 2018; Volansky et al., 2015; Weiß, 2017). However, the results remain in most cases independent from each other and can only serve to complement our knowledge on a certain task.

4.3.4.3 Interpretation through feature engineering and model comparison

One of the most used interpretation methods in computational linguistics and NLP is to compare the prediction performance for different models (see section 4.3.3.2). Therefore, a series of classification or regression models is trained to compare the model performance for different algorithms, parameters, class representation or feature sets. In NLP, the most common case is to compare algorithms, parameter configurations and feature representations with the aim to improve the models for prediction. For text mining purposes it is, however, more relevant to systematically compare different sets of features (see Table 4) in order to find the feature set that is most informative for predicting the output variable (cf. feature engineering). Another common approach are so-called ablation studies that show the results for different subsets similar to model selection where the performance with the full feature set is compared to the performances of different combinations of subsets (see the example in Table 5). Although "brute-force" approaches for ablation studies that perform model selection based on individual features are possible, in practice such approaches only work well if comparisons and feature sets are guided by domain knowledge (Kotsiantis et al., 2006)

[90] Confusion matrices are frequently used e.g. in studies investigating learner language (Bykh, 2017; Peersman et al., 2011; Stemle & Onysko, 2015; Wong & Dras, 2009).

Systematic feature comparison	
Feature (set) A	X % Acc. / X F1
Feature (set) B	X % Acc. / X F1
Feature (set) C	X % Acc. / X F1
Baseline	*X % Acc. / X F1*

Table 4: Systematic model comparison for data mining.

Ablation study	
Feature (set) A + B + C	X % Acc. / X F1
Feature (set) A + B	X % Acc. / X F1
Feature (set) A + C	X % Acc. / X F1
Feature (set) B + C	*X % Acc. / X F1*
Feature (set) A	*X % Acc. / X F1*
….	*X % Acc. / X F1*
Baseline	*X % Acc. / X F1*

Table 5: Ablation study design.

Linguistic studies applying this approach are for example Hancke et al. (2012) for systematic comparison and Ilisei (2012), Pryzant et al. (2018) or Szatlóczki et al. (2016) for ablation studies (see also section 4.1).

4.3.4.4 Interpretation through interactive visual analysis

Another strategy is to use interactive visualization tools to explore the results of a predictive model, e.g. by blended reading strategies. Approaches of this category usually allow to explore individual texts according to their predicted values (e.g. Plaisant et al., 2006), or highlight relevant parts within the texts (e.g. Arras et al., 2017; Koch et al., 2014; Ming et al., 2018; Ribeiro & Guestrin, 2018). Figure 21 and Figure 22 show examples for both strategies.

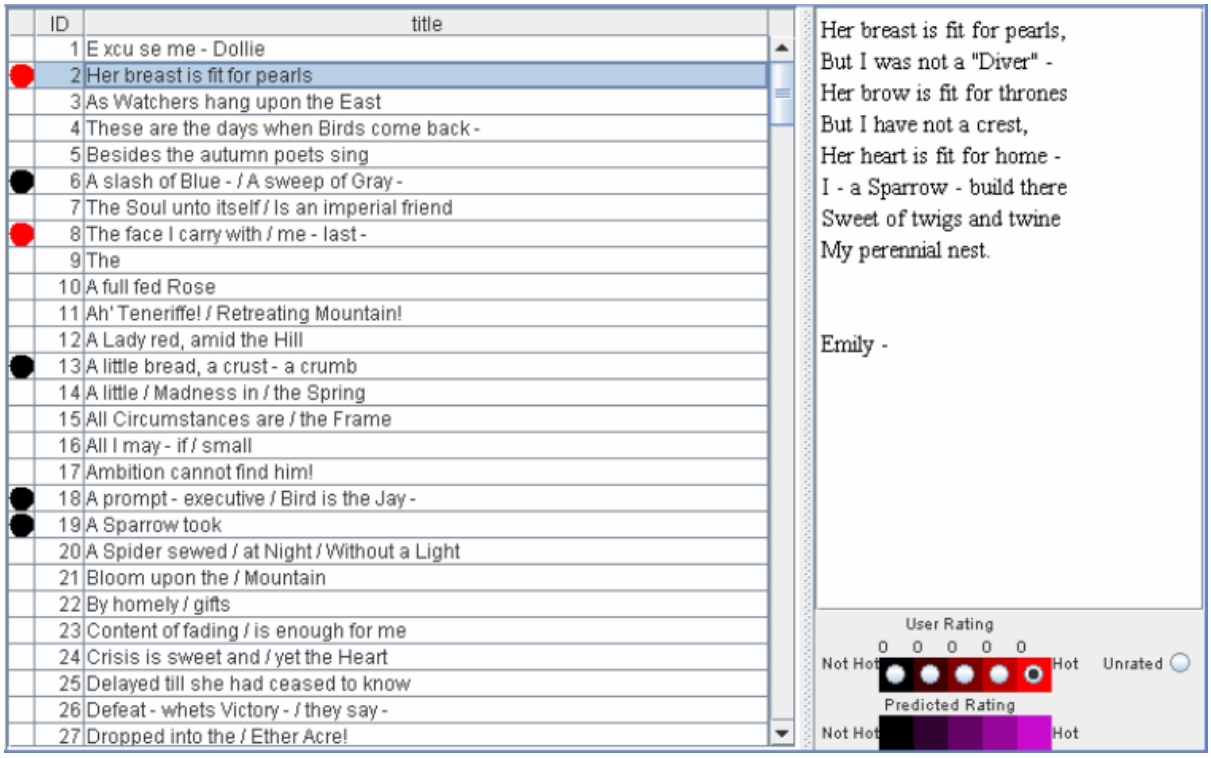


Figure 21: A blended reading scenario from Plaisant et al. (2006).

Figure 22: Relevancy of individual words with the Anchors methods (Ribeiro 2018) (left), or with the gradient-based methodology of Arras et al. (2017) (right).

4.3.4.5 Interpretation through inspection of intrinsically interpretable models

The inspection of intrinsically interpretable models is the most frequent case for statistically oriented quantitative linguistic studies. However, besides regression models, also decision trees, rule induction models, sparse Naïve Bayes classifiers count as intrinsically interpretable models and offer some sort of symbolic model representation that can be investigated (see section 3.4.8). In order to make sure that the model actually did find dependencies that are not due to randomness, the predictive

performance of the model has to be evaluated and a significant increase over the baseline ascertained before the model internals of a predictive model can be investigated.[91]

The most common type of model interpretation in quantitative corpus analysis based on modelling techniques is probably the inspection of variable significance and effect in regression models (cf. section 3.5.1). However, decision tree models are recently also used for interpreting model internals directly through the symbolic representation of the tree. Examples for the interpretation of decision three models in corpus linguistics are (Bernaisch et al., 2014; Deshors & Gries, Forthc.; Gries, Forthc.; Mairesse et al., 2007).

4.3.4.6 Interpretation through post-hoc black-box interpretation methods for complex models

If the model internals cannot be interpreted straight away, post-hoc interpretation methods are needed to make sense of the model. Recent research in explainable artificial intelligence and interpretable machine learning has made substantial advances in the last three years, providing methods that extract and calculate measures, visualizations and simplifications that allow to inspect feature importance, effects and interactions for otherwise uninterpretable complex models.

So far, only very few linguistically-motivated studies have made use of these newly developed strategies. One example is the study of Deshors and Gries (Forthc.) exploring *global surrogate models* for the analysis of grammatical preferences in world Englishes. Model-specific variable importance measures (e.g. marginal effects for regression models or mean impurity decrease for random forest models) have been used for example by (Gries, 2015d; Nesset, 2019; Szmrecsanyi, 2019; Szmrecsanyi et al., 2017). In the following, methods for model-specific and model-agnostic post-hoc interpretation are presented, focusing on the more flexible model-agnostic methods.

Methods for model-specific post-hoc interpretation

A significant number of model-specific interpretation techniques has been developed in the last years, in particular for (deep) neural networks as they are particularly hard to explain but widely used in decision-making systems and artificial intelligence. Established methods are for example *salience mapping*, *sensitivity analysis*, *layer-wise relevance propagation, maximum activation analysis* or *attention-based methods*[92]. For other black box algorithms such as bagging or boosting methods (i.e. gradient boosting algorithms and random forests) there are methods that allow for the identification of interactions, as well as for the extraction of importance measures based on aggregated values for minimal tree heights or entropy measures, e.g. *mean impurity decrease* (Breiman, 2001a; Louppe et al., 2013).

Methods for model-agnostic post-hoc interpretation

Although model-agnostic interpretation methods can have reduced robustness and reliability (Alvarez-Melis & Jaakkola, 2018) the trend in interpretability research goes towards using flexible and comparable techniques that do not depend on one specific type of model but can be applied to any machine learning model. Model-agnostic techniques comprise measure for feature importance, feature effect and interactions, surrogate models, counterfactuals and adversarial examples as well as model criticism and prototypes.

[91] This part, for example, is sometimes neglected with regression models in quantitative linguistic studies (Barth & Kapatsinski, 2018).

[92] For a more detailed description see for example Montavon et al. (2017).

Measures and visualizations of feature importance

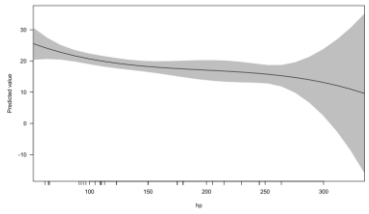
To evaluate the contribution of an individual feature to the predictive performance of the model and the overall ranking of features in terms of their importance independently from the type of algorithm used, two main approaches can be observed (Scholbeck et al., 2019). *Variance-based methods* like the *feature importance ranking measures* proposed by Zien et al. (2009) use the measured effects from *partial dependence plots* or *Shapley values* (see below) to control whether there is a high variance in the effect of a variable or whether the demonstrated curve is instead low. *Performance-based feature importance measur*es such as *permutation feature importance* (Casalicchio et al., 2018; Fisher et al., 2018), or the LOFO[93] method measure the importance by the loss in predictive performance when removing or purposefully permuting the feature values.

After the global *feature importance ranking measure* (Fisher et al., 2018) and the *permutation feature importance* (Greenwell et al., 2018), Caslicchio et al. (2018) proposed the *individual conditional importance*, *Shapley feature importance* and *partial importance curves* to analyse local feature importance.

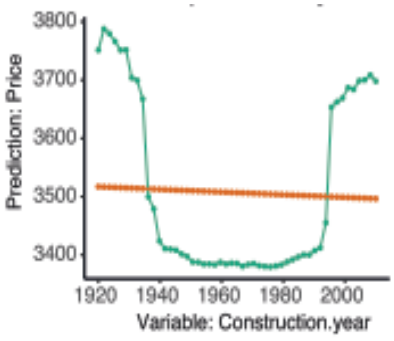
Measures and visualizations of feature effects

Feature effects "indicate the direction and magnitude of change in the predicted outcome when a feature value changes" (Scholbeck et al., 2019). In order to interpret the effect of a feature various model-agnostic methods have been proposed, differing mainly in the underlying mathematical functions used. Most of them strongly rely on two- or three- dimensional plots, in order to allow intuitive interpretation of effect direction, effect size as well as of the type of relationships (linear, non-linear) they illustrate.[94] Popular measures are *marginal effects*, *partial dependence, accumulated local effects, individual conditional expectation*, as well as *Shapley values* and *break down plots*.

Marginal effects, partial dependence (PD) and accumulated local effects (ALE)



Marginal effects is a frequently used measure in traditional statistics, often used for non-linear generalized linear models. However, it has been introduced as a model-agnostic technique to explore global model behaviour for any machine learning method by Leeper (2018) as well as under the name *input gradients* by Hechtlinger (2016). (Image source: Leeper, 2018)

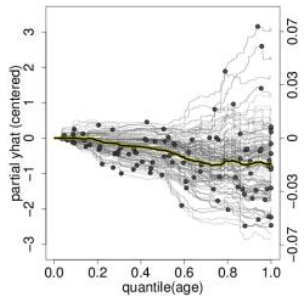


PD plots (Friedman, 2001) are one of the most well-known interpretation methods to explore global feature effects in machine learning models. They are very similar to marginal effects, but the internal calculations to compute them differ. Like marginal effects plots they can obfuscate relationships when there are interactions between features (Scholbeck et al., 2019), this is why they are often replaced by the more robust *accumulated local effects* method (Apley, 2016) that can deal with correlated features. (Image source: Biecek, 2018)

[93] Leave-one-feature-out, also known as LOCO: leave-one-covariate-out

[94] Model-specific alternatives are for example: the regression coefficient, and marginal effects plots for regression models or the variable importance measure provided for many tree-based algorithms).

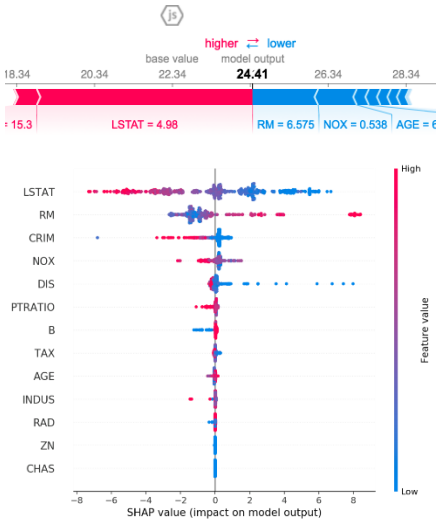
Individual conditional expectation (ICE)



Contrary to partial dependence plots, ICE plots (Goldstein et al., 2015) illustrate local effects by disaggregating partial dependences. They thus suffer from the same issues with correlated features as the PD plots. Although they seem more difficult to interpret intuitively, they show all the individual occurrences and therefore provide more detail than the previous methods.

(Image source: Goldstein et al., 2015)

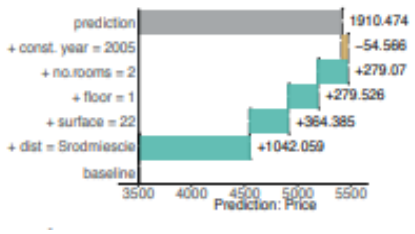
Shapley values



Shapley values display local effects for individual predictions. The methods that has been introduced for machine learning models by (Štrumbelj & Kononenko, 2014), has roots in game theory, where it was used to identify the effects of player entering a game. The extended application of Shapley values in the python library SHAP[95] (Lundberg & Lee, 2017) that allows for local and global interpretation of single features and interactions makes Shapley values one of the most used interpretation technique next to PD plots, even though they are computationally expensive, especially for complex models.

(Image source: https://github.com/slundberg/shap/blob/master/README.md)

Break down plots



Break down plots (Staniak & Biecek, 2018) were originally meant as fast approximations of Shapley values for local feature effects. However, there is also a model-agnostic implementation (Biecek, 2018; Staniak & Biecek, 2018).

(Image source: Staniak & Biecek, 2018)

Measures and visualizations for interaction effects

While interactions can be investigated with two-way partial dependence plots and its derivatives, accumulated local effects plots and individual conditional expectation plots as well as two-way plots for Shapley values, there are other strategies to identify interactions, e.g. Friedman and Popescu (2008). An overview of techniques and some new proposals for visualizations can be found in Britton (2019).

Surrogate models

A surrogate model is a second, usually simpler, intrinsically interpretable model that is trained on the predictions of another (more complex, but better performing) model[96]. If the surrogate model can

[95] https://github.com/slundberg/shap/

[96] Although the term can also be used to describe a more complex model that is used to extend the outcomes of a simple, interpretable model.

successfully explain the predictions of the more complex model, it can serve as a proxy for interpretation. Therefore, surrogate models are also often called proxy models (cf. Gilpin et al., 2018). Surrogate models differ in the type of intrinsically interpretable model used for explaining the black box model, as well as on whether they aim to explain global or local effects. A well-known method for local surrogate models based on a linear regression model is *Local Interpretable Model-agnostic Explanation* (LIME) by Ribeiro et al. (2016b). The model-agnostic (i.e. algorithm-independent) method was one of the first of its kind and gained a lot of popularity. However, later, surrogate models using rule induction methods (Guidotti et al., 2018; Ribeiro & Guestrin, 2018; Sushil et al., 2018) or decision trees (Thiagarajan et al., 2016; van der Waa et al., 2018) have also been used.

Counterfactuals and adversarial examples

The concept of counterfactuals or counterfactual conditional explanations in interpretation means to automatically identify the minimum of changes one needs to make to the observed values for one data point in order to get a different prediction from the model and can be assigned to Wachter et al. (2018).

The concept is as such related to adversarial examples, which focus more on finding close examples (through intentional permutation of observations) that might make the algorithm break[97](Molnar, 2018b). However, adversarial examples have high currency in machine learning research as they were found to enable mutually enhancing training cycles in *generative adversarial networks*. It can therefore be expected that adversarial examples will still gain importance in interpretation settings in future times (e.g. Jia & Liang, 2017).

Prototypes and criticisms

The idea of prototypes and criticism is similar to the investigation of confusion matrices. In order to observe the behaviour of the model and identify potential biases, prototypical data points (prototypes) are put next to exceptions (criticisms) that are similar to the prototypes but don't belong to the same class or belong to the same class but are not similar to the data points (Adadi & Berrada, 2018). Various automatic methods have been proposed over the last years, in order to provide interpretability of black box models via the detection of such prototypical data points (e.g. Bien & Tibshirani, 2011; Gurumoorthy et al., 2017). Kim et al. (2016) proposed a strategy and algorithm (MMD-critic) for both prototypes and criticisms and contributed with this substantially to the field of machine learning interpretability.

Tools and implementations for post-hoc model interpretation methods

There is a number of software tools that provide implementation for the previously introduced model-agnostic and model-specific interpretation methods. Most methods are available either as R packages or Python modules (for some of the methods implementations for both languages exist). Other frameworks or programming languages, however, do not yet offer these very recent methods.

Table 6 and Table 7 below give an overview on the most popular open source packages for R and Python, focusing on model-agnostic tools and only stating some exceptionally well-recognized model-specific options[98]. It gives references and lists the implemented local and global interpretation methods within the packages to give an orientation of the available open source resources for black box model interpretation. The list further focuses on post-hoc interpretation methods. Some machine learning packages sometimes provide model-specific variable importance measures and interaction

[97] Hence, the predictive performance of the model and not the interpretability is in focus here.

[98] Given the fast development of the domain, this list can of course not be complete but should give references for established tools that can be adapted easily to one's needs.

lists (e.g. *randomForest*, *gbm*, etc.). However, this list focuses on packages and libraries specifically developed for interpreting already built models.

R packages

Model-agnostic								
	global				*local*			
	Feature importance	**Feature effect**	**Interactions**	**Surrogate models**	**Feature importance**	**Feature effect**	**Interaction**	**Surrogate models**
iml (Molnar, 2018a)	Shapley importance	ALE, PDP, Shapley values	Rule-ensemble-based interaction detection	tree surrogate	Shapley importance	ALE, ICE, Shapley values	Two-way PDP	Local linear model (LIME)
DALEX (Biecek, 2018)	Permutation feature importance	PDP, ALE, ceteris paribus profiles, merging path plots,	-	-	Permutation feature importance	PDP, ALE, Merging Path plot, Break down plots, ceteris paribus plots	Two-way PDP	-
vip (Greenwell et al., 2018)	Model-specific importance measures, permutation-based importance measures, importance derived from PDP or ICE	PDP, ICE	Two-way marginal effects	-	-	-	-	-
*lime (*for R)	-	-	-	-	Permutation-based feature importance	-	-	Local linear model
Model-specific								
	global				local			
	Feature importance	**Feature effect**	**Interactions**	**Surrogate models**	**Feature importance**	**Feature effect**	**Interaction**	**Surrogate models**

NeuralNetTools	Garson's algorithm	Sensitivity analysis	-	-	-	-	-	-
RandomForestExplainer	Various model-specific and monofactorial association measures, compare different feature importance measures	-	Model-specific interaction detection	-	-	-	-	-
ML packages: *Gbm*, *randomForest*	Model-specific, marginal effect	-	Model-specific	-	-	-	-	-

Table 6: R packages for model interpretation.

Python modules

Model-agnostic								
	Global				local			
	Feature importance	Feature effect	Interactions	Surrogate models	Feature importance	Feature effect	Interaction	Surrogate models
LIME	-	-	-	-	Permutation-based feature importance	-	-	Linear model
Anchors	-	-	-	-	Permutation-based feature importance	-	-	Rule induction
SHAP	SHAP summary plots	SHAP summary plots	Two-way Shap value-based partial dependence plots		Shapley Values	Shapley Values	Two-way Shapley value-based partial dependence plots	-
ELI5	Permutation-based feature importance	-	-	-	Permutation-based feature importance	-	-	Linear model (LIME)
Skater	Model-specific or filter approach	PDP	Two-way PDP	-	-	-	-	Linear model (LIME)

What-if Tool	Association measures	PDP	-	-	-	-	-	-
InterpretML	-	PDP, Morris Sensitivity	-	EBM	Shap Values	Shap Values	-	EBM, LIME
Mode-specific								
TreeInterpreter (decision trees, random forests)	Model-based importance measures	-	-	-	-	-	-	-
Neural networks: Rationale	By visualizing important parts of text.	-	-	-	-	-	-	-

Table 7: Python modules for model interpretation.

5 Summary

In section 1, we defined data science as the aim to address complex research questions by using interdisciplinary computational methods

- to retrieve, prepare, and filter relevant variables,
- to analyse them with advanced statistical/machine learning models,
- to interpret the resulting models and
- to repeat and refine the previous steps

in order to get relevant insights from the data that confirm existing or generate new hypotheses.

Data science, as the application of an interdisciplinary toolset of methods, has been used for quantitative analysis of empirical, observational data in many fields including social sciences, (digital) humanities, economy and business intelligence or biomedicine. While corpus linguistics has used computational methods for the preparation and retrieval of language data or linguistic phenomena of interest ever since early studies in the 1960s (cf. Geoffrey Leech, 1996), the use of predictive modelling and machine learning methods for the analysis and interpretation of corpus data is a rather recent trend that has its motives in the inappropriateness of previously used monofactorial methods and the need for more complex research designs as well as the aim to repurpose time-consumingly created language corpora (see section 2). Predictive modelling and machine learning is, however, an integral part of most studies in text mining scenarios, where the aim is to find interesting new insights from existing language data. Contrary to its main use for processing and prediction tasks in NLP, text mining uses predictive modelling, i.e. machine learning, not only to predict but also to explain the data and generalize the relationships learned to a wider scope (see section 3). This is done by training a set of different prediction models, complementary in used learning algorithms, features and target variable operationalization, and interpreting the results by comparing the predictive performance (e.g.

accuracy) with a baseline or with the other models, inspecting errors through error analysis or predictions in interactive visual analysis approaches, as well as by interpreting the model internals for intrinsically interpretable model or through post-hoc black box interpretation method (see section 4). For all these steps we can draw from the experiences of recently increasing amount of studies in applied linguistics but also from the many methods, frameworks and tools developed in a broader, interdisciplinary field of data science and related disciplines (e.g. text mining, NLP, knowledge discovery, artificial intelligence).

Part II

Repurposing German language corpora with data science

Two case studies

Part I discussed the potential of computational methods of data science that involve various methods of automatization for the individual steps of quantitative analysis (e.g. feature extraction, transformation and selection, visualizations) as well as the methodologically promising approach of using predictive modelling for data-driven analysis in exploratory and confirmatory research designs.

However, it also illustrated how the successful building and interpretation of predictive models depends on theoretical background knowledge, technical skills as well as the modelling task and the size and nature of available data. While there are examples, methods, frameworks and software tools to leverage the skills and background knowledge needed, they are often still experimental, isolated in their research fields, not easily integrated with other technologies or frameworks and most often only provided for English language. All these factors can reduce the aforementioned potential of data science in corpus linguistics substantially.

The aim of this thesis is thus to evaluate the current applicability and relevance of data science methods and predictive modelling when used to repurpose existing corpora of non-English language. In the following, the feasibility, added value and shortcomings of using data science methods and in particular predictive modelling and machine learning for quantitative analysis on already existing German language corpora is evaluated on the basis of two empirical corpus studies.

Both corpus studies make use of previously compiled and annotated corpora that were originally created for other purposes. Furthermore, the studies are based entirely on the available corpus data and annotations. No additional hand-coded annotations have been added to the corpora, and automatic annotation was conducted solely with off-the-shelve methods, e.g. to retrieve indices for cohesion or linguistic complexity, n-grams or to aggregate or transform existing annotations as described in section 4 of part I. This approach allowed to focus on the analysis itself, while evaluating the methods used on linguistic resources that are made available to the community for research purposes[99]. While the studies present and discuss a variety of applicable data science methods and strategies, they function independently as case studies of corpus linguistic research that aim at extending the linguistic knowledge in both investigated fields. To give a broad spectrum of possibilities and applicable methods, the studies differ in their general approach.

The first study, '*Exploring holistic text quality ratings',* has a prevalently exploratory design, using pre-existing and automatically extracted annotations of an extensively annotated corpus of student essays to analyse aspects of text quality in argumentative student essays.

The second study, '*Investigating age-specific language in social media',* has a prevalently confirmatory design, investigating and elaborating upon existing linguistic theories in the realm of age-specific language use in computer-mediated communication on a corpus of German social media texts.

However, both corpus studies integrate in the linguistic study of their field while utilizing interdisciplinary methods for knowledge discovery and hypothesis testing.

[99] E.g. via research data repositories where researchers publish their corpora and annotations alongside articles, infrastructure initiatives like CLARIN (Hinrichs & Krauwer, 2014), or via enhanced corpus collection possibilities through digitalization processes or user-generated content in social media.

Empirical corpus study 1

Exploring holistic text quality ratings

6 Introduction

One reason to employ data science methods in corpus linguistics is to explore a corpus in a data-driven way, investigating a broad range of linguistic features that might be related to a certain (text) characteristic and contribute individually or in combinatorial way to explaining a given characteristic of the text. In this manner, the first empirical corpus study uses predictive modelling and machine learning to investigate and possibly explain manually annotated holistic judgments of text quality in German argumentative student essays, while controlling also for systematic biases and rater effects. The (linguistic) features analysed are of diverse nature, allowing to demonstrate potentials and difficulties with typical corpus linguistic data types.

The study starts with an overview on how text quality is investigated in other fields, to set the theoretical background of the investigation. Then the corpus used for the analysis is introduced in detail, giving information on the type and nature of available data. Following that, the study design and methodology is presented, describing the operationalization of text quality and feature sets used in the analyses and giving an overview of the subsequent experiments.

The experiments are designed to illustrate workflows and strategies for exploring an existing corpus with data mining methods. More precisely, the analysis

- uses data stored in annotations, in additional questionnaires or retrieved by NLP feature extraction methods;
- investigates the data by conflating, transforming, splitting, subsetting, or aggregating features, changing their data types or comparing features of different nature;
- attempts mono- and multifactorial analysis using traditional statistical as well as machine-learning-based methods for predictive modelling
- inspects the data with descriptive statistics, visualization methods for monofactorial analysis, as well as interpretation methods for intrinsically interpretable and complex black box models.
- uses methods and tools that can be utilized without strong background in programming and NLP.

The experiments are grouped into three sections, each of them containing a detailed description of the methodology, before the results are presented and discussed.

The first, section 10, introduces principle workflows and concepts while analysing the relationships and interactions between the holistic grades and categorical data from a text analysis questionnaire.

Section 11 then focuses on methodological difficulties of corpus data (distributional issues, e.g. biases, outlier, hierarchical feature sets, noise and redundancy) while analysing error annotations in the corpus and their relationship to the holistic grades of the essays.

Section 12, finally, is dedicated to the interpretation of text classification and regression models of different types, using intrinsically interpretable models and black box models trained for predicting the holistic grades on the basis of automatically extracted indices of linguistic complexity.

At last, the summary and conclusion discusses the value of this study for the field of German linguistics and address follow-up research desiderata that arose from this study.

7 The study of text quality

Text quality is studied in various fields including language testing, second language acquisition, writing research, translation studies, media studies and social media analysis, the digital humanities, and literary studies and, more recently, in recommender systems, business intelligence and marketing. Language testing, second language acquisition, writing research and also literary studies see it as one of their core interests to assess, relate and re-evaluate the concepts of text quality. Often these approaches arise as an answer to evolving possibilities and needs given the availability of new types and dimensions of potentially useful – or harmful[100]- data. In particular, the prominent fields of social media analysis and business intelligence have recently made big contributions to large-scale text analysis.

However, when we look at these different fields it is clear that there is no shared understanding of text quality. On the contrary, we deal with rather diverse perspectives, where text quality can mean well-judged by literature critics evaluating the literary value of a novel or some piece of poetry, or having the best combination of keywords and text structure to be easily retrievable via a search engines e.g. in blog articles, but also leading reliably l to higher sales numbers (e.g. for social media customer relations in Twitter feeds).[101] We only need to think of the often painfully well-written examples of hate speech one can find on social media to see that text quality highly depends on perspective. Moreover, even within one field we can find different definitions of text quality depending on the precise research question that is being investigated (Neumann, 2016). Consequently, the operationalizations used in order to define and analyse text quality also vary widely.

7.1 Concepts and operationalizations of text quality

Below, I introduce common ways to view, define and operationalize text quality in different fields.[102] The section starts with operationalizations from closely related fields which analyse text quality within

[100] E.g. in order to distinguish between good user-generated text and bad text like spam or discriminating or annoying texts written by trolls or people engaged in cyber bullying and hate speech.

[101] Even David Robinson's study on the authenticity of Donald Trump's tweets distinguishing self-authored tweets from those written by his public relations team might be seen as a sort of text quality prediction (see http://varianceexplained.org/r/trump-tweets/) (see also Preoţiuc-Pietro & Devlin Marier, 2019).

[102] For the sake of coherence, approaches from spam detection and social media forensics are not in the focus of this description, as these somewhat related but still conceptually different approaches would exceed the scope of this work. But see Davidson et al. (2017), Zhang et al. (2014), or Zheng et al. (2013) for approaches to such problems.

student writing (e.g. language testing, second language acquisition, and writing research) and then move on to views and respective operationalizations appearing more recently in other fields.

Holistic text quality judgments

Holistic text quality judgments are one of the main instruments used in the measurement of text quality in and outside of language testing scenarios. The aim of this kind of evaluation is to judge the overall quality of a text. The judgments are holistic in the sense that there is only one judgment scoring the whole of the text without considering individual parts or different aspects of quality individually, i.e. "holistic ratings refer to general impressions of the quality of a writing product as a whole, while it is in fact a mix of evaluations of different dimensions (e.g. content, language use, structure)" (Van Den Bergh et al., 2012). Judgements are often obtained from human annotators who give subjective ratings. While studies in language testing usually adhere to certain standards of reliability, e.g. a minimum inter-class correlation of 0.7 for inter-rater-reliability (cf. Graham et al., 2012; Wilmsmeier et al., 2016), other fields don't necessarily consider the consistency of raters, which is sometimes not even achievable[103]. Furthermore, there is evidence showing that even though holistic ratings might be highly subjective and generally result in a lower inter-rater-agreement than using analytical scores for example (see section below), they also tend to be less topic-dependent and generalize better across different genres (Van Den Bergh et al., 2012)[104].

Analytic text quality judgments

Analytic text quality judgments, on the other hand, aim to derive a text quality rating by summarizing (often mathematically) a number of ratings for subcriteria of text quality. "Each essay is scored on a similar set of characteristics, which are usually easy to assess" (Van Den Bergh et al., 2012). This method of scoring a text, although more time-consuming (Wilmsmeier et al., 2016), is known to be more reliable. However, studies have shown that this scoring technique does not generalize well over different writing tasks (Van Den Bergh et al., 2012).

The item set used for analytic text quality judgments is usually either extracted from sources in text linguistics or language assessment (e.g. Becker-Mrotzek & Böttcher, 2006; Brinker, 2010; Kruse et al., 2012; Nussbaumer & Sieber, 1994) or based on standardized tests and scoring rubrics (e.g. East, 2009; Neumann & Lehmann, 2008). A detailed rating manual and intensive rater training is needed to ensure reliable ratings (Wilmsmeier et al., 2016). Hence, another line of research investigates the agreement between analytical and holistic text quality judgments and the bias introduced by raters (Harsch & Martin, 2013).

Criterial text quality judgment (statistical factors)

In their comparison of different approaches to measuring text quality, Grabowski et al. (2014) point out another frequently used method of assessing text quality, which is to approximate the quality via known linguistic/non-linguistic factors that have already been proven to correlate with text quality in a narrower sense. Measures frequently used for this approach comprise the length of a text or for example its type token ratio, for example. Grabowski et al. show that the text length can indeed be used as an approximation of text quality, and can mitigate possible biases introduced by raters as text length can be measured objectively. Nevertheless, measuring text quality with only one text statistic always represents an approximation, and the identification of further factors related with measures such as text length does not ensure that these factors also correlate with text quality.

[103] Because of the subjectivity of the task or the unavailability of repeated judgments (Neumann, 2012).
[104] See also Olinghouse et al. (2012) for relativizing results.

Competency levels

In the fields concerned with language learning, be it research on academic literacy, second or first language acquisition or language testing, often text quality is often a proxy for the writer's language or text competence (or vice versa) (Neumann, 2016). Many scholars have worked on defining levels of language, text or writing competency. Different schools of thought have developed on language, national or international level depending on whether they are concerned with literacy development or second language acquisition. While second language acquisition nowadays often builds on international or European standards (cf. CEFR; Council of Europe), writing research for first language competences often refers to models and schemes that have been developed for a specific language and/or educational system (cf. Neumann & Lehmann, 2008). The generalizability of competency levels and text quality is one of the major concerns in this field (Schoonen, 2012; Van Den Bergh et al., 2012). Although many of the current models rely on the same sources[105], when applied in language testing and linguistic assessment the proposed models are diverse and controversial, with some scholars and practitioners in language teaching even denying the usefulness or validity of levelled competence schemes in general.[106] However, all of these attempts usually define a (linear) model of text competence that has well-defined characteristics. Annotators are then asked to classify texts on the basis of the model description and guiding material.

Readability

While in the previously introduced operationalizations text quality is usually interpreted as higher competence (featuring higher linguistic complexity, coherence, lexical diversity and longer texts), non-learner or student-oriented operationalizations might focus on transversal effects, defining text quality as high readability (hence featuring lower complexity of sentences, shorter sections to maintain attention, coherent terminology, and less lexical sophistication). Readability seems to particularly concern the field of language pedagogy and teaching, alongside public administration and public media, non-literary translation and scientific writing. Even web publishing is concerned nowadays with readability as a measure of text quality, offering text analysis tools that provide the author of a weblog or web text with indicators of its readability and help with search engine optimization[107]. As early on as 1969, McLaughlin defined readability as "the degree to which a given class of people find certain reading matter compelling and comprehensible" (Mc Laughlin, 1969), thereby indicating its dependence on the intended audience (cf. Schriver, 1989). Scholars were also aiming to measure readability as a proxy for text quality in a more objective way. Using linguistic features of the text, they proposed composed or analytic indices of readability, some of which gained very high popularity in the ensuing decades of writing research (e.g. *Flesch reading ease score* (Flesch, 1948)*, Gunning FOG index* (Gunning, 1968)*, SMOG grading* (Mc Laughlin, 2014)*, GulpEase* (Lucisano & Piemontese, 1988)*,* etc.) However, readability in the sense of ease of comprehension generally aims for less complex, simpler text, which is not always a main concern for text quality definitions in other disciplines.

Outreach, success and engagement

A recently evolving strategy to operationalize text quality is measuring the outreach, success or engagement of texts. Scholars and marketers have studied textual elements related to message propagation (e.g. Tan et al., 2014), reaction counts and business transactions with easily quantifiable quality measures like number of downloads (e.g. Ashok et al., 2013), likes, comments or shares within a

[105] E.g. writing models of Bereiter (1980), Scardamalia/Bereiter (1987), Hayes and Flower (1986).
[106] For example, see Harsch and Martin (2012) for a discussion on the reliability of CEFR ratings.
[107] For example the tools in https://yoast.com, www.textanalyse-tool.de/, www.wdfidf-tool.com.

certain time span (e.g. Arakawa et al., 2014; Frey et al., 2013), clicks or sales counts (e.g. Packard & Berger, 2017; Pryzant et al., 2017).

Relevance and helpfulness

Another concept of text quality used particularly in the study of digital or social media is to define it as the text's relevance or helpfulness (for a certain community). This concept originates in information or document retrieval, where the main aim is to design algorithms that find and rank the most relevant documents in a database given a certain search query (e.g. Mihalcea & Tarau, 2004). The database could be a traditional relational database but also a company's knowledge base, or the whole world wide web (e.g. in the sense of search engines). The measures used to create this rank of relevance might vary, ranging from the *number of incoming and outgoing links* as in one of the first implementations of Google's PageRank (Page et al., 1999), to highly complex algorithms for recommender systems (Lops et al., 2011) or *news feed* / *friend feed* generation that consider factors like the similarity to previously viewed contents of the reader or his network, or the multimodality of a given content (e.g. Bucher, 2012; DeVito, 2017). Ranks of relevance or helpfulness are not always calculated. Another common approach is to crowd-source reader endorsements (Cao et al., 2011; Krishnamoorthy, 2015; Kuan et al., 2015; Singh, Irani, et al., 2017) of texts like online reviews, or answers in Q&A (i.e. Question-and-Answer) platforms[108] (Dalip et al., 2013).

7.2 Hitherto investigated features and exemplary studies

As is to be expected, both the concept and operationalization of text quality changes depending on the scientific interest, but also the (linguistic) features that are being investigated when attempting to explain text quality. In the following, I describe trends and commonly used features for the different fields named above and present some examples from recent studies.

Within the context of student writings, there are numerous works in *automated essay scoring* that try to predict holistic and or standardized analytic scores of essays using computer programs (Shermis, 2014; Shermis & Burstein, 2003). In general, the field is mainly concerned with simplifying the process of grading for teachers and evaluators and does not necessarily aim at interpreting the learned relations between score and text afterwards. Hence, automated essay scoring often bases on numerous sets of (mostly stylistic) features, often extracted automatically by NLP technology (e.g. argument mining, discourse marker detection) or text analysis tools for word counts or readability and complexity measures like Coh-Metrix, TAACO, TAALES, LWIC (Crossley, Roscoe, et al., 2011; Yannakoudakis et al., 2011; Zesch et al., 2015).[109] More recently, there are also neural network-based approaches that make use of complex text representations (e.g. word-embeddings) or similar non-interpretable features (Farag et al., 2018; Nadeem et al., 2019; Taghipour & Ng, 2016). While these give better overall prediction results on tasks where large annotated training sets are available, a recent study of Nadeem et al. (2019) reports that tasks where only small training data is available benefit more from feature-based approaches. However, no matter if the models are based on large hand-crafted feature sets or on pre-trained text representations, automatic essay scoring usually aims at achieving best possible prediction results. The interpretation of these models is thus secondary, and the studies barely contribute to understanding the relationship between linguistic features of the text and the resulting grade.

[108] E.g. *quora.com* or *stackoverflow.com*.

[109] An overview of the field and approaches can be read in (Dikli, 2006; Zupanc & Bosnic, 2015).

Exceptions to such a prediction-focused approach are for example recent works in academic language proficiency for first language (L1) and second language acquisition for non-native writers (L2) by Scott Crossley and Danielle McNamara, who made extensive studies under the framework of automatic essay scoring but with the aim of analysing linguistic features and writing strategies for well-rated essays (Crossley, Roscoe, et al., 2014; e.g. Crossley & McNamara, 2011b, 2016; Jung et al., 2019; McNamara et al., 2010). Although their analyses are often based on a high number of not immediately interpretable features, they use interpretable models (e.g. regression models, features selection, association measures, structural equation modelling, etc.) in order to gain insights that might also have pedagogical relevance (e.g. Crossley & McNamara, 2016). While first studies tested the predictability of holistic and analytic judgments and their linguistic correlates in general as well as for different writing tasks, some of the newer works put a lot of attention on coherence and cohesion, multi-word-expressions (analysed with n-grams) and personality features of the writer. The findings for L1 writers show for example that overall text coherence is one of the best predictors of holistic text quality ratings (Crossley & McNamara, 2010) as is text elaboration measured in text length and lexical sophistication (i.e. academic vocabulary and longer words in general) (Crossley, Roscoe, et al., 2014), the use of less frequent vocabulary (Crossley, Weston, et al., 2011) and a higher lexical diversity (McNamara et al., 2010). Syntactic complexity, i.e. how long and complex the sentences are, was also found to be correlated with text quality next to a higher use of rhetorical structures (e.g. exemplifications, reported speech) (Crossley, Roscoe, et al., 2014). While errors in spelling and punctuation were only weakly related with text quality in higher grade student essays, there was, however, no correlation between grammar errors and holistic grades (Crossley, Kyle, Varner, et al., 2014). Additionally, although coherence seems to be an important factor of writing quality, the attempts to measure coherence based on (local) cohesive devices have mostly failed (Crossley & McNamara, 2010, 2011a). This is supported by other research reporting that advanced level writers and texts of higher text quality scores use less explicit cohesive devices (e.g. Crossley, Kyle, et al., 2014; Crossley, Weston, et al., 2011; McCutchen, 1996). Other, similar works for L2 are for example Taguchi et al. (2013) or Vajjala (2018), who also investigate linguistic features of writing quality. The features that correlate with or have been found to be good predictors of text quality in these studies on English as a second language essays are mostly the same as for native writers. However, the used vocabulary and lexical diversity seems less important in these texts: errors, especially those not concerning spelling are, contrary to the texts of native writers, particularly important to establish text quality.

There is a notable overlap in scholars and methods in the assessment of text quality of student essays and the analysis of text readability in teaching contexts. Features and approaches in these two areas differ only slightly favouring transfer between the two perspectives of writing quality. In general, research investigating readability as an approach to text quality is often rooted in educational settings (e.g. the evaluation of teaching material).

However, readability has also been evaluated as a measure of text quality in non-educational settings (cf. Pitler & Nenkova, 2008). In public administration, for instance, readability is used to measure quality in the sense of accessibility for all groups of the society. Venturi et al. (2015) for example analysed document quality of informed consent forms in health care. Under the premise that health-related texts “should be accessible to all members of the society” they define text quality as “suitable and comprehensible for patients and consumers in general” and investigate automatically extracted readability measures (e.g. word length, parse tree height, word frequency or word overlap between paragraphs) as features to classify texts as difficult vs. easy-to-read. They built an SVM classifier for a reference corpus that provides news articles written for an audience with reduced literacy skills or slight mental disability next to a classic newspaper corpus. Comparing central tendencies for both corpora, they found for example that easier texts had shorter sentences and words, more words from a base

vocabulary, more concrete nouns and verbs and fewer adjectives, shorter dependencies and fewer subordinate clauses (all measured in central tendencies per corpus). After they confirmed the discriminative power of their chosen readability measures on those two corpora, they used average values for the measures for both corpora — easy and difficult — as reference points to locate new texts (i.e. informed consent forms) on a scale of difficulty.

Research that can be classified under the methodological umbrella of stylometry and stylistics has also contributed to the analysis of text quality e.g. when analysing literary style, stylistic features of well-written journalism, successful novels or well-cited scientific papers. Common methods of analysis in this area are usually the comparison of classifiers, interpretation of monofactorial relations or generalized linear models (but see also multi-dimensional analysis by Biber (e.g. 2014), linear discriminant analysis (Brown, 1984), or structural equation modelling (e.g. Yang, 2009). Ashok et al. (2013), for example, try to predict the download counts of the Gutenberg catalogue as a measure of the degree of success of novels using SVM. They use stylistic features (lexical and syntactic) adapted and evolved from previous studies in stylometry, genre discrimination and authorship attribution, namely lexical choices (unigrams and bigrams), distribution of word categories (POS tags), grammar rules (CFG), constituents and sentiment/connotation. In their analysis they compare SVM classifier accuracies for 5-fold cross validation (reported prediction accuracy at 84%) and inspect differences in most frequent lexical items and mean distributions of features for subgroups. Their findings suggest that readability of the text is not proprietary to highly successful writing. Highly successful books do not necessarily need to be readable; they are characterised by verbs that describe thought-processing instead of concrete actions and expression of emotions and use more discourse connectives and prepositions but less sentiment-laden words. The authors furthermore point out clear differences in the style of successful texts between genres, hinting to the fact that the popularity (as measured by them) does not necessarily mean literary excellence.

The aspect of "literariness" has been taken up later in a similar study by van Cranenburgh and Bod (2017). The authors used a crowd-sourced holistic rating to operationalize quality as literariness on a Likert-scale of 1-7. The ratings were then predicted using word and character n-grams, stylometric features (e.g. sentence length, direct speech rate, vocabulary richness), topics extracted via LDA (Blei et al., 2003) and syntactic tree fragments (i.e. arbitrarily-sized fragments of parse trees) for support vector regression. They were able to explain 76% of the variation in the ratings. The comparison of 5-fold CV and an error analysis including the investigation of worst predictions and mean differences for subgroups let them conclude that literary language is usually richer and more varied, employing a larger set of syntactic constructions, but that there are subgroups of literary texts that behave very differently and are stylistically closer to non-fictional texts.

Tan et al. (2014) analysed the effect of wording on message propagation, i.e. the virality of a message measured in retweets on Twitter. They analysed, among others, shallow NLP features like n-grams and skip-grams, readability measures, sentiment and indices of informativeness and compared SVM classifier accuracies of 5-fold CV for the different feature sets and ranked lexical features on the basis of their coefficient. They control for known confounding factors in social media (namely the popularity of the author and the topic of the tweet) pairing alternative wordings of the same content from the same user and comparing the two. Their findings show that successful tweets were more informative (longer and more content words), resembled the style of headlines and contained conventionalised patterns and vocabulary that followed community norms and was true to the writer's own previous language use.

Pryzant et al. (2017) also control for known confounders, but contrary to Tan et al. (2014) they don't limit the dataset, but the set of features. In this business research-oriented study, they predict sales

on a Japanese e-commerce platform based on the language of the product descriptions. They use mixed-effects models to analyse the data and compare the generated models for different feature sets containing morphological and lexical features as well as non-linguistic (confounding) features. Additionally, they develop a neural network-based feature mining system to make sure that (for some of the feature sets) only those features known to be independent from the confounders brand loyalty and pricing strategy are considered. Furthermore, they consider different variable importance measures to interpret the created models and find meaningful relations between linguistic (lexical) features and the sales number achieved with a product description. A final clustering-based content analysis of variable importance rankings allowed to draw valuable insights for further marketing or research ambitions.

The helpfulness of online reviews is investigated in Krishnamoorthy (2015) and Singh et al. (2017) with a similar business-oriented perspective. Krishnamoorthy compares F-score and accuracy results for different classifiers (Naïve Bayes, SVM and random forests), different review types (positive and negative) and different product types. Singh et al. (2017) shows a similar study setup using tree ensembles for regression and reports also variable importance ranks for individual features. Both consider linguistic features e.g. stylometry and readability features, lexical features as well as non-linguistic metadata.

Summary

As we have seen, the concept of text quality depends on the text genre, the text function and scientific interest and might be measured in different ways. Text quality can be defined as a text's relevance, helpfulness, readability, success, its virality or even its author's underlying text competence. All these views have developed their own operationalizations, utilizing quantitative metadata, crowd-sourced reader opinions, holistic or standardized analytic rating rubrics to measure quality. While language testing and psycholinguistic research have very standardized and sophisticated scoring schemes, school and educational settings often refer to holistic text quality judgments. Marketing and Social media research have rather pragmatic, easily quantifiable views on text quality, while administration and public media focused on ensuring readability for their target audience. However, when investigating the (linguistic) features related to text quality, we can see considerable overlap between those disciplines. This overlap is nurtured especially by the recent availability of well-performing automatic feature extraction tools from natural language processing and computational linguistics. Although, disciplines often have a particular interest in specific features (e.g. error frequencies in writing research or sentiment in social media analysis), more general, transferable features from stylometry, indices of cohesion or linguistic complexity, n-grams and word-embeddings are making their way into most fields concerned with text quality. When it comes to explaining and exploring text quality on a linguistic level, recent analysis methods in all disciplines frequently use predictive modelling. However, interpretation techniques barely go beyond the comparison of classifier performances.

8 The KoKo corpus and its characteristics

This study makes use of the KoKo corpus created at the Institute for Applied Linguistics, Eurac Research Bolzano and first introduced in 2014 (Abel et al., 2014). The corpus was designed to describe language competences of L1 writers at the transition from high school to university and to conduct sociolinguistic research comparing texts with differing writer characteristics.

The authors thus collected German student essays from German native speakers at the end of their secondary education in combination with a questionnaire obtaining socio-demographic metadata from the writers. In total 1319 essays (around 716,000 tokens) written by German native speakers[110] from three different German speaking areas were obtained and transcribed. The texts have a mean length of 654 words (with a standard deviation of 268, see Figure 23).

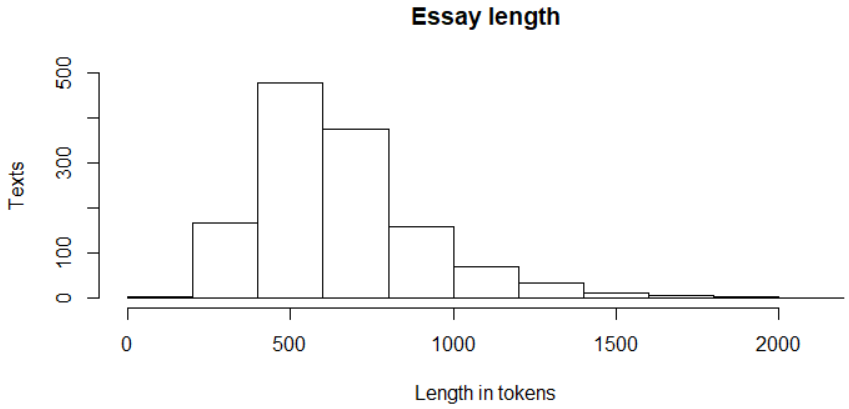


Figure 23: Histogram for essay length in tokens.

All the texts had been produced during school hours as a non-graded extracurricular activity. The students were asked to compose a hand-written argumentative essay ("Erörterung") answering the same input prompt. All students had 120 minutes of time to accomplish the task that was phrased as follows:

> Bitte schreiben Sie zum folgenden Thema eine Erörterung:
>
> Der deutsche Schriftsteller und Essayist Hans Magnus Enzensberger (*1929) hat in einem Interview vom 4. Mai 2001 mit der Wochenzeitung Die Zeit unter anderem Folgendes gesagt:
>
> "Aber wissen Sie, ich finde, die Jugend ist sowieso keine beneidenswerte Phase des Lebens. Ich verstehe gar nicht, warum die Leute so einen Kult damit treiben. Ein junger Mensch ist labil, unsicher, schwankend, hat keine Souveränität, macht jede Dummheit mit. Denken Sie nur an diese Klamottensucht, ein Leben in der Diskothek, schrecklich. Wenn der eine ein Motorrad hat, muss der andere auch eines haben. Das ist doch entsetzlich. Man muss froh sein, wenn man das überstanden hat."
>
> Setzen Sie sich mit diesem Zitat auseinander und nehmen Sie persönlich Stellung!
>
> Zum Bearbeiten dieser Aufgabe haben Sie 120 Minuten Zeit.

The students were in the penultimate year before their high school graduation and between 17 and 18 years old. Participating school classes were sampled randomly ensuring an equal number of

- school classes from Thuringia (Germany), North Tyrol (Austria) and South Tyrol (Italy);
- school classes located in small, medium and large cities;
- school classes of academic and vocational high schools[111].

[110] As whole school classes were involved in the data collection, the authors of the corpus also obtained essays from pupils with other first languages. These have been ignored in the corpus creation (Abel et al., 2014).
[111] Both types allow to enter university after graduation.

Apart from the transcribed texts, there are additional materials available for the KoKo corpus that are provided in the form of annotations (see Abel et al. 2014 for detailed description) and other metadata tables. These annotations and materials comprise: the data from the socio-demographic questionnaire filled out by the students themselves (for all 1319 texts), the frequency of errors and norm violations per text for orthography, grammar, punctuation and lexical errors and salient features for partly overlapping subsets of texts, the results from a text analysis questionnaire, filled out by linguistically trained annotators (for a subset of 584 texts).

Socio-demographic data

The available metadata for the students include socio-demographic data and information on the student's language background and communication habits. Table 8 gives a short overview of available student metadata. The data is partly saved as metadata annotation in the corpus and partly as a data table file.

general demographics	- gender - birth year - birth country/province
academic performance	- mark in German first language instruction of the year before the student's participation in the study
school type	- academic track vs. vocational track
first language background	- first language of student (German or free choice) - German monolingual/bilingual/other language monolingual language spoken with a) mother, b) father, c) sisters and brothers, d) other people at home, language prevalently spoken with friends, variety prevalently spoken a) at home b) with friends c) with teachers outside of school d) at public places (standard German, colloquial standard German, standard-oriented register with dialectal influence, dialect or no German) - frequency of use of any of these registers (scale of 4) - used registers for different communicational forms or genres (mail, blog, chat, diary, letters)
communication and media consumption habits:	- frequency of use of *communicational forms* or genres (mail, blog, chat, twitter, diary, letters) - frequency and duration of a) television consumption, b) radio consumption, c) newspaper, d) literary texts, e) factual (non-literary) texts, f) comic books, g) blogs, origin of a) television, b) radio channels consumed - number of books read in the last three months
attitudes towards dialectal and non-dialectal registers / language varieties:	- judgments of appropriateness of dialect/standard in school, public events, media, with friends or family, appropriateness of standard spoken in Germany, Austria, South Tyrol - do students like dialect/standard German in their region (7-point scale) - do students find dialect/standard German in their region a) useful, b) educated, c) acknowledged, d) attractive, e) sympathetic,

	f) pretty, g) close or distant, h) shameful, i) warm or cold, j) rich or poor, k) easy or difficult (7-point scale)
residence:	- country of residence: Germany (Thüringen), Austria (Tirol) or Italy (South Tyrol) - duration of residence, size of city of residence: countryside, small city, middle-sized city, big city - boarding school attendance - number of people at home - number of parents living at home
family and economic status:	- education and employment of mother/father - number of cars, estates, vacations - pocket money

Table 8: Overview of socio-demographic metadata available for the writers.

Error annotations and frequency tables

As the KoKo project (Abel & Glaznieks, 2017) aimed at analysing student's first language competences at the end of their school career, one important aspect of the project was to analyse formal correctness by looking at the frequency of errors in the essays.[112] Therefore, a detailed hierarchical error annotation scheme has been created and used to categorize errors and lexical misuse in the corpus. The error annotation scheme is based on German orthography and grammar and is divided on a basic level into a) orthography, b) punctuation, c) grammar and d) lexis[113], each of the categories subdivided by error type. Although the annotation scheme is highly hierarchical and split into many subcategories, providing fine-grained information on the occurring errors, not all subdivisions of errors have been used in the analysis conducted during the KoKo project for reasons of simplicity and feasibility (Abel & Glaznieks, 2017).

Orthography

The annotation scheme for orthographic errors is based on the rules and principles of German orthography (Duden, 2005, 2006; Fuhrhop, 2005). Orthography errors have been subcategorized into omissions, insertions, transpositions of graphemes as well as errors in capitalization or word splitting (Abel et al., 2014). These subcategories were further split into smaller groups of errors (e.g. omission errors with missing double consonants, split words that should not have been split, etc.) The category '*other*' subsumes error types that only occurred occasionally, i.e. hyphenation errors, errors with apostrophes, proper names, abbreviations, positioning of graphemes and all other errors of unknown type (see Table 9).

orth_lcp.cap	capitalization errors
orth_sep.tog	erroneously split or combined words, including errors with hyphenated words
orth_omissions	omissions of letters or morphemes
orth_insertions	insertions of letters or morphemes

[112] In (corpus-based) research in second-language acquisition, this is one way of measuring *linguistic accuracy* (cf. Polio 1997), which is, together with *linguistic complexity* and *fluency*, part of the triad of second language competence according to the popular framework of Housen and Kuiken (2009).

[113] For the lexis, also non-erroneous but salient lexical features were annotated during the KoKo project.

orth_transpositions	confusions of letters
orth_other	other errors (e.g. apostrophes, proper names, hyphenation, unknown errors)

Table 9: Orthography error types.

Punctuation

Abel & Glaznieks 2017 split punctuation errors into seven main analysis categories and a category 'other' for all remaining punctuation errors (see Table 10). They further provide subcategorizations for most groups, especially errors concerning the comma, which are split into a fine-grained categorization. However, subcategorizations have not been analysed in detail in order to keep the analysis simple and feasible.

punc_fehl_komma	missing comma
punc_fals_komma	wrong comma
punc_fehl_punkt	missing period
punc_fals_punkt	wrong period
punc_fehl_sz	missing punctuation mark (other)
punc_fehl_doppelp	missing colon
punc_fals_sz	wrong punctuation mark (other)
punc_other	other errors

Table 10: Punctuation error types.

Grammar

The grammar annotations are based on Duden (2005) and Zifonun et al. (1997) and were summarized into seven main categories according to Abel et al. (2014). The remaining error types have been categorized as '*other*' grammar errors (see Table 11). Subcategorizations regarding specific cases, word categories or `acc_spk` have been made for the individual categories.

gram_anak	anacoluthon (ungrammatical blending of phrases and clauses)
gram_corr	correspondence relations (erroneous selection of case, number, gender or person of a dependent word with respect to government and congruency)
gram_infl	inflective (incorrect inflected forms that are independent of a governing element, e.g. forms following the wrong inflection paradigm such as weak instead of strong verbal inflection)
gram_uncl	unclear (not categorizable grammatical error)
gram_inco	incompleteness (incomplete sentences and phrases as well as the incorrect use of ellipses)
gram_redu	redundant (erroneous repetitions of words and parts of sentences)
gram_woor	word order (violations of any kind of word order restrictions)
gram_other	all other grammar errors

Table 11: Grammar error types.

Lexis

On the lexical level there are annotations for errors as well as for other (neutral or positive) salient lexical features in the essays. The annotation scheme used for the analysis of lexical errors and particularities in KoKo is complex and comprises various dimensions. These are illustrated in Figure 24. They include the structure of the unit (single word vs. formulaic sequence), a subclassification for the unit (e.g. neologism, vs. argumentative adverbs and conjugations or referential, communicative, structural phrasemes), a semantic dimension (denotative or connotative correctness of use), a stylistic dimension (e.g. redundancy or repetition), a form dimension (formal correctness of the unit), and finally a metalinguistic dimension (whether units are correctly marked as metalinguistic or not) (Abel et al., 2016). Contrary to the other error categories, the error typology of the lexical category are thus not subcategories that can be combined to a total amount but transversal dimensions, where various grouping options are possible. Annotation types can be grouped on a formal dimension, correctness dimension, stylistic or semantic dimension as well as according to the size of the investigated unit. The following abbreviations are used for the categorization of lexical errors (see Table 12).

EW	single word
FS	formulaic sequence
semkonnot	semantic connotation
semdenot	semantic denotation
stil	stylistic dimension
form	formal errors
meta	meta-linguistic item or emphasis

Table 12: Abbreviations for lexical error types.

The annotations were made by linguistically trained annotators. For reasons of feasibility, not all texts have received error annotations for all four categories. Table 13 shows the annotation layers for errors and salient features and the number of annotated texts for each annotation layer. Because of only partly overlapping subsets, the amount of texts where all annotations are available amounts to 349.

Error annotation layer	**Texts**
Orthography	1319
Punctuation	1319
Lexical errors and salient lexical features	980
Grammar	596
Holistic evaluations (and grades)	584
All annotations	349

Table 13: Number of texts annotated for each error annotation layer.

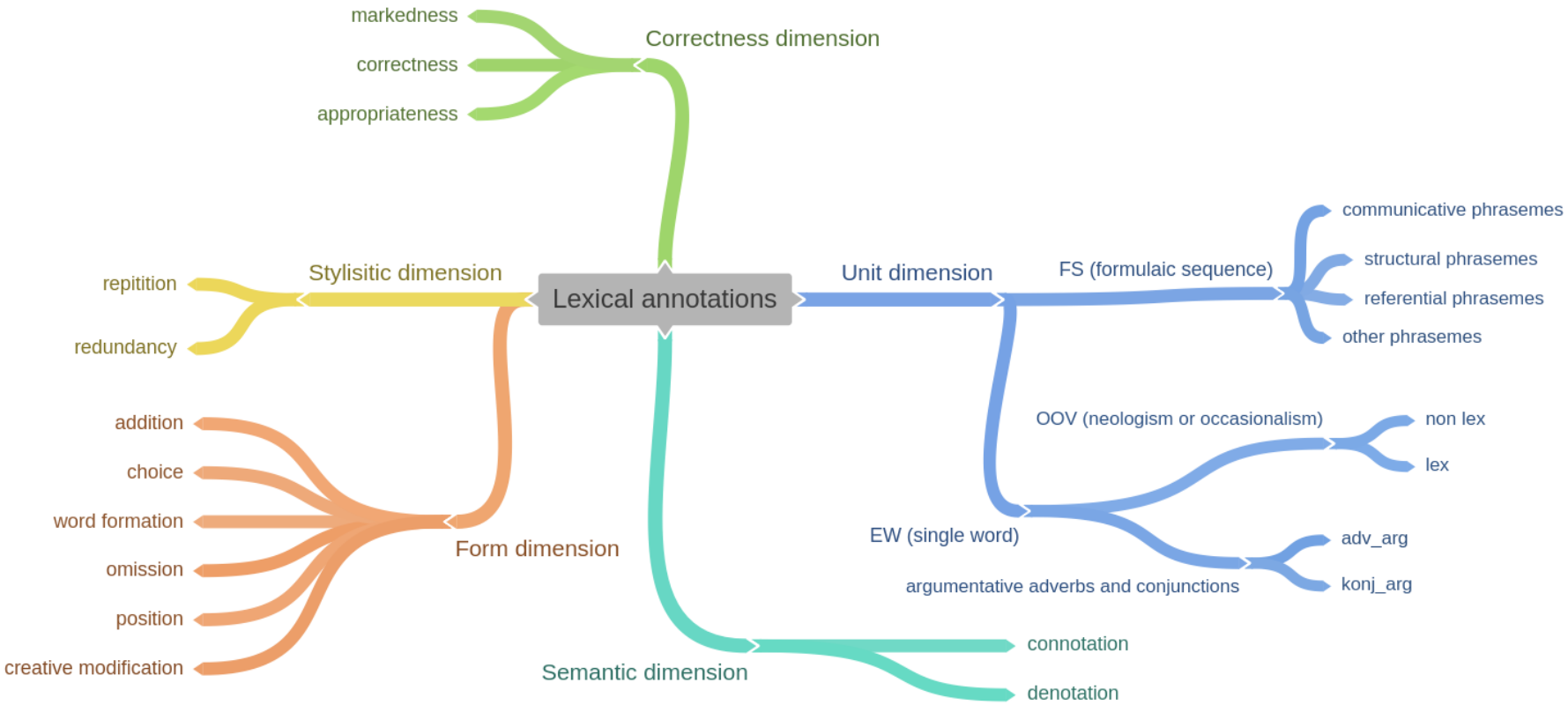


Figure 24: Multidimensional annotation scheme for lexical errors and salient features.

Text analysis questionnaire

During the KoKo project detailed manual qualitative text analysis has been conducted to evaluate text quality features like coherence or appropriateness of structure, topic and register (Abel & Glaznieks, 2017). The text analysis questionnaire was inspired by the *Zürcher Textanalyseraster* (Nussbaumer, 1996; Nussbaumer & Sieber, 1994), which is a popular text analysis rubric in German linguistics and writing research. However, some questions have been adapted in order to account for recent works in German writing research and text analysis (in particular Augst et al., 2007; Becker-Mrotzek & Böttcher, 2006; Brinker, 2010; Feilke, 2010; Jechle, 1992) and for the curricula of the participating regions (cf. Abel et al., 2016).

The text analysis questionnaire contains 65 items from four categories: (A) formal completeness, (B) content, (C) formal and linguistic means of text arrangement and (D) overall impression. Most items only allowed to choose one answer out of a set of often ordinally ranked categories, but there are also multiple choice and free text items in the questionnaire.

The topics addressed in the questionnaire can be summarized as follows.

(A) Formal completeness
- Elements of texts (introduction, main part, conclusion, ending)
- Presence of writer's opinion

(B) Content
- Reference to the task prompt
- Topic
- Genre patterns and register
- Modes of argumentation

(C) Formal and linguistic means of text arrangement
- Paragraph structure
- Textual discourse structure

(D) Overall impression
- Fulfilment of task
- Consistency of quality
- Coherence
- Appeal of text
- Other summarizing text characteristics

The analysis was carried out by three linguistically trained annotators[114]. Each of them annotated around 220 texts. 27 of these texts were annotated by all three annotators, to evaluate inter-annotator reliability for the items. Additionally, a consensus annotation was conducted for three texts. In total there are 646 filled questionnaires for 584 texts.

All corpus data, annotations and materials have been made available for the purpose of this study by the Institute for Applied Linguistics at Eurac Research Bolzano, Italy. The corpus itself is available for research purposes and can be downloaded as XML files via CLARIN[115] or queried at https://commul.eurac.edu/annis/koko. Further information and description of the corpus and the KoKo project can be read in Abel and Glaznieks (2015, 2017) or Abel et al. (2014, 2016).

[114] In the following the annotators are referred to as Annotator A, B and C.
[115] http://hdl.handle.net/20.500.12124/12

9 Study design and methodology

This corpus study on text quality in student essays provides a systematic overview on current data mining methods that can extend the core corpus linguistic toolkit when analysing a corpus in an explorative manner.

As an example, the study identifies and interprets relations, possible confounding factors and interactions as well as individual and maybe outstanding observations of text quality in the argumentative student essays of the KoKo corpus. The KoKo corpus, as we saw above (section 8), provides holistic judgments of text quality (see section 7.1), i.e. the dependent variable analysed in this study, as well as a number of readily available feature sets, concerning analytical text evaluations, error frequencies and measures of linguistic complexity. It therefore represents a ready-made test base, where one can explore text quality in a data-driven manner.

Apart from the methodological investigation, the study offers an in-detail description and prediction of the holistic text quality judgments of the corpus and might thereby contribute

- to decompose human judgments of text quality and link them to observable characteristics of the text
- to automate grading and the annotation of text quality for other/larger datasets, which in turn supports research on a broader empirical basis and can serve as a foundation for practical applications.

Below, I present the distribution and character of the dependent variable, the used feature sets and the experiments that have been conducted in this study and discuss the adopted methodology, as well as the tools and methods I used.

9.1 Holistic text quality judgments

The holistic grades for text quality in the KoKo corpus have been obtained via the text analysis questionnaire described in section 8. They are represented as a vote of overall text quality on a 5-point Lickert scale[116]. While the extreme values 1 and 5 were labelled with *'Mangelhaft'* (insufficient) and '*Ausgezeichnet*' (excellent) respectively, the values in the middle did not have any labels in the questionnaire. This was done to ensure that annotators would assume levels of equal distance. Figure 25 shows the distribution of grade labels of all 646 evaluations.

[116] The concrete questionnaire item was phrased as such: "Welche Note gibt am besten die Textqualität wider?" ("Which grade best reflects the text quality?")

Figure 25: Distribution of 646 holistic grades obtained through the text analysis questionnaire.

As we can see in Figure 26, the distributions of assigned grades differ visibly from annotator to annotator and suggest varying means variances for grade labels. A Kruskal Wallis rank sum test was used to check group means, showing that the mean grade label of annotators was not significantly different from each other (df = 3, alpha = 0.5, p-value = 0.7106). However, the Fligner-Killeen test of homogeneity of variances showed significant differences between Annotator A, B and C (df = 3, alpha = 0.5, p-value = 0.0124) and suggests that Annotator A has a lower variance and tends to assign average grades (grade '*3*').

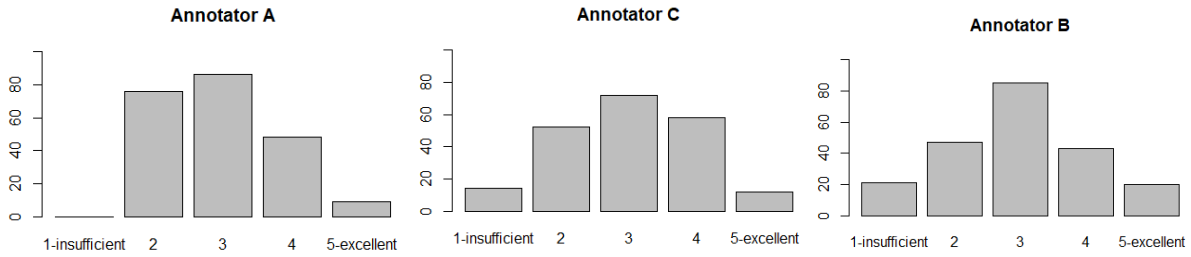


Figure 26: Distribution of holistic grades for each annotator.

Table 14 shows the standardized residuals for the chi-squared test for independence for comparing annotators and grade distributions. The values show the difference between what would be expected for annotators if they all had the same distribution of assigned holistic grades in a standardized way. Here, too, we can see that Annotator A assigns '*1-insufficient*' much less frequently than the other annotators.

	1 (insufficient)	**2**	**3**	**4**	**5 (excellent)**
Annotator A	-4.36	3.12	0.50	-0.56	-1.67
Annotator B	3.43	-2.16	0.53	-1.41	2.15
Annotator C	1.02	-0.82	-1.19	1.94	-0.41
Annotator Gold	-0.42	-1.06	1.03	0.42	-0.45

Table 14: Standardized residuals for chi-squared test for independence for annotators and holistic grades.

Moreover, for testing the reliability of the grades, 27 texts have been annotated by more than one annotator, resulting in a total of 584 annotated texts. The inter-rater reliability for the grades was rather low (0.285 ICC, 0.29 Krippendorff's alpha), indicating that the holistic grades represent a subjective rating of each annotator. Given the fact that such holistic rating is the main evaluation

technique in school settings, it is, however, still interesting to analyse and decompose the annotators' notion of text quality.

The experiments in this study thus consider the effect of the rater wherever possible, including the annotator as feature in classification approaches and as random effect in regression models as well as investigating differences in bivariate associations and marginal effects, when interpreting the data.

9.2 Feature Sets

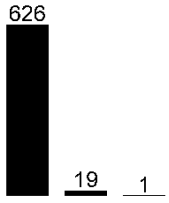
This study uses three feature sets composed of different feature typologies that are frequently used to analyse text quality in other studies.

1. Qualitative evaluations on individual items of a text analysis scheme in the form of text metadata (already available for the corpus in the form of the text analysis questionnaire)
2. Relative and binarized error frequencies, in the form of frequency lists of manual linguistic annotations (available for the corpus in the form of raw frequency lists)
3. Measures of linguistic complexity, as automatically extracted NLP features

The use of these three different feature sets allows to demonstrate affordances and difficulties that arise from different data types typically used in corpus linguistics on a methodological level.

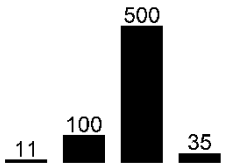
9.2.1 Metadata: Text analysis questionnaire

One of the feature sets used in this study contains the items of the text analysis questionnaire that has been collected for a subset of 584 texts of the KoKo corpus[117]. For this study the subset of single-answer items from the original questionnaire is used. Items with free-text answers and questions with the possibility to choose multiple options are excluded for the sake of reducing noise. Additionally, I excluded Question 20 that was asked twice in the questionnaire (cf. Question 57) and is therefore not expected to provide any additional information[118]. In total, 50 items of the questionnaire have been used in this study. While most questionnaire items show a positive evaluation for the majority of texts (e.g. most texts have been evaluated as fulfilling the proclaimed text functions, having a meaningful paragraph structure and provide a conclusion to the text), some of the items showed that argumentative structure and argumentative power of texts, for instance, cannot necessarily be expected from the students. Figure 27 shows examples of mainly positively evaluated questionnaire items and items with a rather balanced distribution of positive and negative evaluations. The full list of used items in this feature set can be seen in appendix A. The list also shows the answer possibilities and the distribution of answers for each question as well as conflated versions of the items used in some of the analyses.



Q31_explizite_Adressatenorientierung
Is there an explicit addressing of a fictitious or actual recipient?
- No
- Yes
- Not determinable (NA)

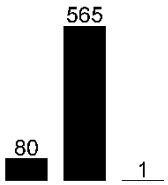
[117] The questionnaire is described in detail in section 8.
[118] Indeed, the answers given for both questions where almost the same with just a few exceptions.



Q39_Umbrueche_gelungen

Are the line breaks between introduction, main part and ending appropriate?

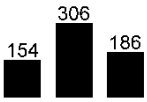
- No
- Not all
- Yes
- Not determinable (NA)



Q47_Qualitaet_konsistent

Is the text quality fluctuating or constant?

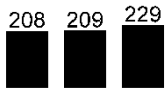
- Fluctuating
- Constant
- Not determinable (NA)



Q36_Argumentationsgang_deutlich_logisch_nachvollziehbar

Course of argumentation: Which characteristic applies to the text?

- There is no clearly recognizable and logically comprehensible course of argumentation that leads to a certain result or conclusion.
- There is a clearly recognizable and logically comprehensible line of argumentation that leads to a certain result or conclusion.
- Not determinable (NA)



Q30_konzessive_Argumentation

Is the argumentation concessive?

- No
- Yes
- Not determinable (NA)

Figure 27: Examples for questionnaire items from the text analysis questionnaire.

9.2.2 Annotations: Error frequencies

The second feature set is derived from the error annotations that have been made in order to analyse orthography, punctuation, grammar and lexis errors during the KoKo project[119] (c.f. section 8), using frequency tables that contain the number of errors per error type, per essay. To represent the hierarchical annotation scheme designed by the authors of the corpus, aggregate values for each level of hierarchy have been created.

As the number and selection of annotated texts for each annotation layer differ slightly, only 349 texts have been annotated for all the annotation layers and the holistic grade. The dataset for all analysis that include the error frequency feature set, therefore refers to this subset of 349 texts, so that results for different error categories can be compared.

Furthermore, I expect errors to be dependent on the text length of the essays, as shown by previous studies (cf. Dikli, 2006; Grabowski et al., 2014; Perelman, 2014 for discussion). The average text length for the subset of 349 texts used here is 642 words, with a standard deviation of 275.1. Because of the high variance in text length, it seems reasonable to normalize error frequencies for text length, calculating the number of errors per 1000 words.

[119] For study results see Abel and Glaznieks (2015, 2017).

Table 15 shows descriptive statistics for the four main error categories. For visual comparison, Figure 28 illustrates boxplots of absolute frequencies and Figure 29 boxplots for the normalized values.

Error category	**Total errors in corpus**	**# of texts with error type**	**Mean rel. error frequency**	**SD rel. error frequency**
Orthography errors	2899	331	13.7	14.4
Punctuation errors	3884	343	18.1	12.7
Grammar errors	1619	321	7.6	6.0
Lexis errors	4218	338	19.0	10.4
All errors	1262	*349*	46.4	27.8

Table 15: Distribution of error categories.

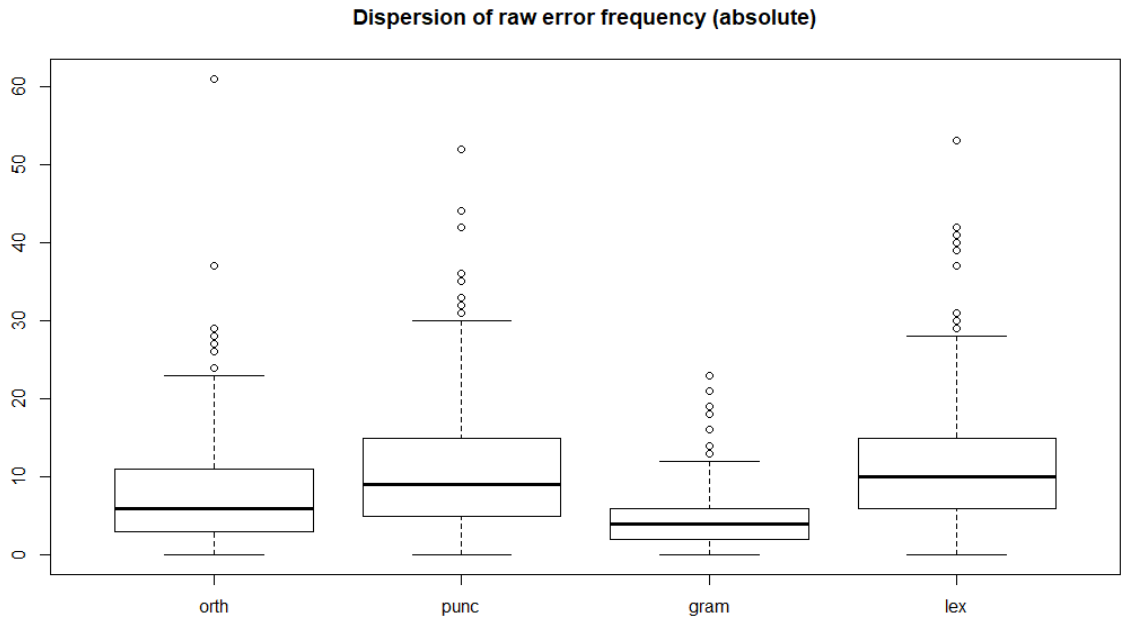


Figure 28: Boxplots for distribution of error types per text.

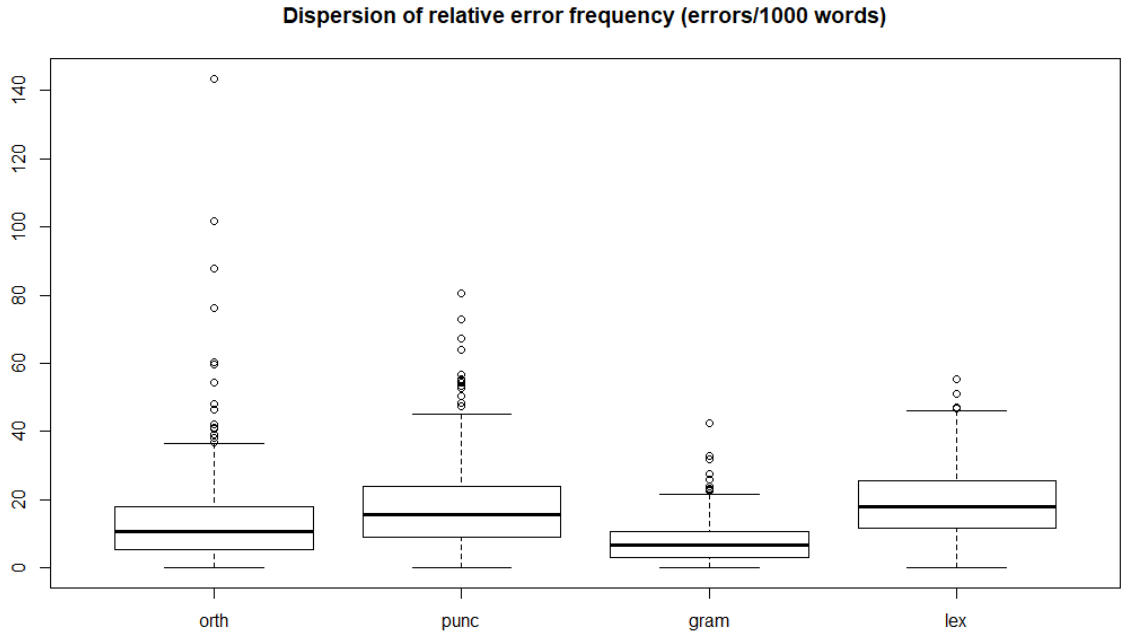


Figure 29: Boxplots for relative error frequencies.

With respect to the subcategories of orthography, punctuation, grammar and lexis errors, as they have been categorized by Abel et al. (2014) (see Table 16 to Table 19), we can see that distribution of many error types is skewed. The majority of texts has only few errors of each category while some texts contain many errors. Below, I show distributions (histograms) and descriptive statistics for the subcategories of error types for the four categories orthography, punctuation, grammar and lexis.

9.2.2.1 Orthography errors

Error type	Total errors	Texts with error type	Mean rel. error freq.	SD rel. error freq.
orth lcp/cap	1043	275	5.0	6.3
orth sep/tog	622	246	2.9	3.1
orth omissions	496	179	2.4	5.0
orth insertions	236	133	1.1	1.9
orth transpositions	270	147	1.3	2.4
orth other	232	128	1.1	1.9
All orthography errors	*2899*	*331*	*13.7*	*14.4*

Table 16: Distribution of orthography error types.

Capitalisation errors are the most frequent type of error in the KoKo corpus, they are also the most diffused among the texts, with almost 79% of texts having at least one capitalization error. However, although capitalization errors occur up to five times as often as other orthographic errors, word compounding or splitting errors are frequent followed by omissions of characters.

9.2.2.2 Punctuation errors

Error type	Total errors	Texts with error type	Mean rel. error freq.	SD rel. error freq.
punc fehl komma	3069	339	14.6	12.3
punc fals komma	580	217	2.5	2.9
punc fehl punkt	45	35	0.2	0.7
punc fals punkt	74	57	0.3	0.8
punc fehl sz	27	26	0.1	0.5
punc fehl doppelp	52	38	0.2	0.8
punc fals sz	22	20	0.1	0.5
punc other	15	13	0.1	0.4
All punctuation errors	*3884*	*343*	*18.1*	*12.7*

Table 17: Distribution of punctuation error types.

The vast majority of punctuation errors are missing commas, they make for almost 80% of all punctuation errors and occur in 339 of 349 essays. Wrongly set commas are also common. All other punctuation errors, however, are almost neglectable and usually occur only occasionally and/or are clustered within a few texts. As the subcategories of punctuation errors established in the KoKo project have very different ranges, I decided to aggregate errors related to commas, as well as all non-comma-related errors.

9.2.2.3 Grammar errors

Grammar errors are very unequally distributed over the corpus (see Table 18). Some error types only occur in some of the texts (where they then tend to occur more often). The most frequent grammar error category are correspondence errors that make about 60% of all the grammar-related errors and occur in 80% of the texts. Other error categories are much less frequent.

Error type	Total errors	Texts with error type	Mean rel. error freq.	SD rel. error freq.
gram anak	70	64	0.3	0.7
gram corr	951	277	4.4	4.1
gram infl	127	96	0.6	1.1
gram uncl	66	57	0.3	0.8
gram inco	200	136	1.0	1.2
gram redu	20	19	0.1	0.7
gram woor	52	45	0.2	4.5
gram others	133	102	0.6	1.1
All grammar errors	*1619*	*321*	*7.6*	*6.0*

Table 18: Distribution of grammar errors

9.2.2.4 Lexical errors

Lexical errors comprise the incorrect use of words and idioms (formulaic sequences). Around a third of this error category was related to formulaic sequences (acronym FS), where the most common error was to mix up the words that are used in the idiom (lex_FS_form e.g. “Stress auf Arbeit haben” instead of “Stress auf der Arbeit haben”). The rest of the errors were incorrect uses of single lexical items (acronym EW for *Einzelwort*). Here form errors (e.g. “zwangsvoll” instead of “zwanghaft”) were less important and wrong semantic denotation (semdenot) or inadequate semantic connotation (semkonnot) were the most frequent error types (48% and 37% of all single word errors).

Error type	Total errors	Texts with error type	Mean rel. error freq.	SD rel. error freq.
EW.semdenot	1289	302	5.7	4.5
EW.semkonnot	1010	266	4.5	4.6
EW.stil	116	82	0.5	1.1
EW.form	82	63	0.4	1.1
EW.meta	197	117	1.0	2.0
FS.semdenot	252	146	1.1	1.8
FS.semkonnot	291	164	1.3	1.9
FS.stil	30	28	0.1	0.6
FS.form	895	278	4.1	3.8
FS.meta	50	44	0.2	07
EW.total error	2694	333	12.0	7.4
FS.total error	1518	312	7.0	5.3
All lexical errors	4218	338	19.0	10.4

Table 19: Distribution of lexical error types.

9.2.3 NLP and computational linguistics: Linguistic complexity measures

The third and last feature set is composed of linguistic features designed to measure linguistic complexity, cohesion and readability automatically extracted via NLP tools. In order to extract the features, I referred to the linguistic complexity measures provided by Weiß et al. (2017; 2019; 2018).

Their feature extraction pipeline for German linguistic complexity measures implements a vast amount of linguistic complexity measures, including indices of cohesion and coherence; of lexical, morphological and syntactical complexity; as well as measures of cognitive complexity and language use. It provides German language implementations of indices known from other studies on English and combines them with additional features specifically tailored to German (e.g. features on the topological field model). Furthermore, the indices have already successfully been used in other studies analysing L2 competence levels, readability and cohesion in learner texts (Galasso, 2014; Hancke & Meurers, 2013; Weiß, 2015, 2017), and therefore represent a valid standard for automatic feature extraction for the KoKo corpus.

Through the feature extraction pipeline, it was possible to extract 284 features (or measures) that were potentially relevant for the 584 texts of the KoKo corpus with holistic grades (see appendix B). Weiß (2017) grouped the measures into four major categories:

- Measures of language use
  (e.g. measures of mean age of acquisition of lexical items, word frequencies in representative language samples, phraseological sophistication)
- Measures of discourse and encoding of meaning
  (e.g. textual cohesion, lexical concreteness;)
- Measures of human language processing
  (e.g. cognitive load, integration costs)
- Measures of the linguistic system
  (e.g. lexical complexity, syntactic complexity, morphological complexity)

The measures investigate form and meaning-oriented features concerning syntax, lexis, morphology, semantics, phonology or pragmatics of the text that have been designed to operationalize elaborateness, complexity and variability of the used language. They thus provide quantifiable variables that can be used to compare texts and corpora for aspects such as concreteness of the vocabulary or use of cohesive devices.

Below, I give a short overview of groups of measures that are potentially interesting for the analysis at hand and give examples of how to read them in context. For each group only a few example measures are given to illustrate how the measures are operationalized. The tables below the description of the groups show the internal variable name, a verbal explanation and descriptive statistics for the distribution of the feature in the KoKo corpus. For most groups, there are, however, different versions available in the feature set (e.g. comparison with different reference corpora, measures based on different representations such as lemma or original token or normalized on different linguistic units such as per token, per sentence, per clause)[120].

9.2.3.1 Measures of text length

Measures of text length have a long tradition as operationalizations for text quality and are known to be good predictors of holistic text quality ratings (Grabowski et al., 2014; McNamara et al., 2010).

Examples:

nParagraphs	The number of paragraphs in the text.	mean: 8.03 sd: 5.03
nSentences	The number of sentences in the text	mean: 32.6 sd: 15.5
nTokens	The number of Tokens in the text	mean: 548.7 sd: 234.8

9.2.3.2 Word frequency measures calculated on representative language samples

This category of measures operationalises the elaborateness of the vocabulary used by giving an estimate of how common/rare the lexical items in the text are. To achieve this, the vocabulary used in the text is compared with word frequency lists of different reference corpora. The corpora used are texts from Google Books from the year 2000 to 2009 (Brysbaert et al., 2011) abbreviated as *Google00*, the *DlexDB* reference corpus for German prose and news articles from the 20th century (Heister et al.,

[120] A full list of measures used in this study is provided in the appendix. For further description of the measures and their implementation please refer to Galasso (2014), Hancke et al. (2012), as well as Weiß (2015, 2017; 2019; 2019).

2011) , the *Subtlex-DE* corpus of movie and TV subtitles (Brysbaert et al., 2011) or the *KCT Karlsruhe Child Text* corpus that collects written vocabulary for children of different ages (Berkling et al., 2014).

Examples:

typeFreqsPerTypeFoundInDlex	The type ratio for DLEX, i.e. the mean frequency of types found in the reference corpus of German language.	mean: 33910 sd: 6047
logAnnotatedTypeFreqBand6PerTypeFoundInSubtlex	The types of the text that were found in the most infrequent words of the Subtlex corpus of television subtitles, log transformed for dependency on text length.	mean: 0.017 sd: 0.01
lemmaFreqsPerTypeFoundIn KCT	The average frequency of lemmas for each type found in the Karlsruhe Child Text corpus (KCT)	meand: 0.57 sd: 0.104

9.2.3.3 Other lexical features concerning diversity, word length, relatedness and concreteness of used words

Further lexical features in the feature set regard lexical diversity (also called *lexical richness,* i.e. the variedness of the used vocabulary), lexical relatedness (i.e. if the vocabulary is semantically close) or concreteness (i.e. whether the used words are rather abstract or concrete) as well as word length measures (e.g. number of characters per word) or part-of-speech ratios such as non-auxiliary verbs per token.

Examples:

MTLD	The measure of textual lexical diversity (McCarthy & Jarvis, 2010), known to be less dependent on text length than other lexical diversity measures.	mean: 135.2 sd: 29.2
nonAuxVerbsPerToken	The number of non-auxiliary verbs per token in the text.	mean: 0.12 sd: 0.017
syllablesPerToken	The number of syllables per word.	mean: 1.76 sd: 0.087
synsetPerTypeFoundInGnet	The number of synonyms in the text according to a lookup of all words of the text in GermaNet.	mean: 2.05 sd: 0.35

9.2.3.4 Measures of cognitive complexity

This type of measures is designed to estimate the cognitive load needed to process a text. Calculating distances in between positions in the topological field as well as measuring integration cost according to the Dependency Locality Theory by Gibson (2000), Weiß (2017) implemented a number of features to capture the cognitive complexity of processing a given text.

Example:

oMaxTotalIntegrationCosts PerFiniteVerb	The maximum total integration cost per finite (cf. Weiß, 2017).	mean: 1.4 sd: 0.35
syllablesInMdleFieldPerMdle Field	The number of syllables in the middle field per middle field in the text.	mean: 7.7 sd: 1.7

9.2.3.5 Measures related to cohesion

Various features have been proposed to measure global (text level) and local (paragraph level) cohesion in texts (cf. Crossley et al., 2016; Crossley & McNamara, 2010; Galasso, 2014). The measures for textual cohesion used in this study regard the transition of grammatical roles according to the *Entity Grid* by Barzilay and Lapata (2008), the use of pronouns as a co-reference instruments as well as the number of connectives in the text. The measures have been implemented for German by Galasso (2014).

Examples:

probSubObjsPerTransition	The probability that the subject of a sentence occurs as an object in the subsequent sentence.	mean: 0.000369 sd: 0.000959
pronounsPerNoun	The ratio of pronouns per noun.	mean: 0.77 sd: 0.2
Breindl_all_ConnectivePerSentence	The ratio of connectives listed in the *Handbuch der deutschen Konnektoren* (Breindl et al., 2014) normalised per sentence.	mean: 0.25 sd: 0.13

9.2.3.6 Syntactic and grammatical features

Syntactic and grammatical features give estimates on the organisation, length, sophistication and complexity of sentences, clauses, t-units or other grammatical constructs. The measure e.g. the parse tree height as a measure of clause complexity, how prepositional phrases the sentences contain in average or how many subordinated or co-ordinated clauses a sentence has.

Examples:

sumParseTreeHeightsPerClause	The average parse tree height per clause.	mean: 2.87 sd: 0.42
coordinatedPhrasesPerClause	The ratio of coordinate phrases per clause.	mean: 0.125 sd: 0.062
WordsPerFiniteClause	The average length of a finite clause measured in words.	mean: 8.37 sd: 1.02
PPsPerSentence	The ratio of prepositional phrases per sentence.	mean: 1.47 sd: 0.45

9.2.3.7 Measures of morphological complexity

The morphological complexity of the text is measured by features of derivation, inflection, composition or word length. Measures of morphological complexity show how abstract and sophisticated the used vocabulary is, by measuring e.g. the length of compounds or the number of derived word forms.

Examples:

compoundDepthsPerCompoundNoun	The Depth of compounds per compound in the text.	mean: 2.01 sd: 0.2
derivedNounsPerNoun	The number of derivate forms per noun in the text.	mean: 0.145 sd: 0.046
charactersPerToken	The average length of words measured in characters.	mean: 5.37 sd: 0.24

9.2.3.8 Measures related to academic language

Finally, Weiß (2017) specifically designed a number of features to measure academic language use in German. Academic language is commonly known to be characterised by nominal style i.e. a higher ratio of nouns to verbs as well as a higher use of periphrastic tenses, passives and non-subject pre-fields (Weiß, 2015).

Examples:

coverageDeagentivationPatterns	A number indicating how many different deagentivation patterns have been used in the text.	mean: 0.695 sd: 0.159
nonSubjectPrefieldsPerPrefield	The ratio of non-subject pre-fields.	mean: 0.524 sd: 0.175
prenominalModifiersPerNp	The mean number of prenominal modifiers in noun phrases of the text.	mean: 0.174 sd: 0.049
muVerbClusterSize	The mean verb cluster size for the text.	mean: 2.31 sd: 0.57

9.3 Study setup

The experiments in this study shall give an overview of possible methods when exploring an existing dataset for new, interesting insights valuable for linguists. They are designed in order to demonstrate general workflows and popular methods and approaches but also to illustrate challenges in analysis design, issues of data skewness and other biasing elements as well as issues generated by the use of complex data or complex methods when analysing or interpreting the data linguistically. The study splits in three sections with increasing methodological complexity, starting from naïve approaches to data mining with off-the-shelf data mining tools that do not need any programming skills to the training and interpretation of complex black box models with machine learning libraries and interpretation modules for Python. Each section focuses on one of the feature sets above.

Section 10 (First explorations & basics of data science) illustrates a basic workflow for machine-learning-based data-driven analyses using text classification that can be followed without further

programming knowledge with the WEKA data mining software. It starts with a naïve approach, without consideration of the characteristics and distributions in the data, it evolves step-by-step refining the analysis approach with the available functions of the software and shows built-in interpretation methods. By showing the effect of different algorithms, parameters, evaluation and feature selection techniques it gives metalevel information on the used methods that can help to design similar analysis.

Section 11 (Dealing with issues of observational data) deals with some challenges in the analysis that arise from non-normal distributions, outliers and complex or noisy datasets. It shows constraints and limits of the standard analysis methods previously presented and hints towards follow-up problems of interpretation.

Section 12 (Interpretation of complex feature sets) focuses on interpretation challenges when working with complex datasets and complex model architectures. Interpretation techniques for intrinsically interpretable as well as for black-box models are discussed, testing their capacity to get insights on the notion of text quality encoded in the holistic grades of the corpus.

10 First explorations & basics of data science

Feature set 1: Text analysis questionnaire

In this section a basic workflow for using predictive modelling for linguistic inquiry is illustrated. All experiments base on feature set 1 containing the various (categorical/ordinal) items of the text analysis questionnaire presented in section 9.2.1. Text classification as well as regression models are used to model relations between the analytical evaluations given for a text and the corresponding holistic grade.

The section starts with a naïve approach using all variables in their categorical representation, without consideration of the characteristics and distributions in the data. The analysis then evolves step-by-step by refining the analysis approach with the available functions of the WEKA data mining software and by transferring the data into ordinal variables. By showing the effect of different algorithms, parameters, evaluation and feature selection techniques it gives metalevel information on the used methods that can help to design similar analysis.

10.1 Methodology

The first experiments show different methods and strategies for building text classification models on categorical data. The classification approaches are performed using standard learning algorithms and default configurations available off-the-shelf via the WEKA data mining software that can be used by corpus linguists and researchers that are not familiar with programming or with other more complex machine learning toolkits like KNIME or RapidMiner. Seven different learning algorithms were chosen in order to compare their predictive performance and interpretability.

1. Naïve Bayes Classifier
2. Logistic Classifier
3. PART Classifier for rule induction
4. J48 Decision Tree Classifier
5. Multilayer Perceptron
6. SMO with polynomial kernel (support vector machine)

7. Random Forest Classifier

The first four algorithms fall under the group of allegedly simple and intrinsically interpretable models as their architecture is less complex and they work with data representations that can often be interpreted without further interpretation methods. The last three algorithms on the other hand are known as black-box algorithms that apply complex transformations to the input variables (MLP, SMO) or base on a high number of submodels (random forest) which usually results in high predictive performance, allowing to model non-linear relationships, but also makes the resulting model complex and difficult to interpret. All algorithms have been used with their default settings.

After training the classifiers, their validity is evaluated by comparing their predictive performance against the majority baseline. The prediction performance of the built models is compared primarily using estimates for prediction accuracy in 10-fold CV[121]. Although the use of 10-fold CV (i.e. the percentage of correctly classified observations when splitting the dataset into ten parts and then training and evaluating ten different models each evaluated on one of the ten parts and trained with the remaining nine parts of the data) allows to account for different test-train-sample splits by repeatedly evaluating over different parts of the dataset, there is still a random element in the way the data is split.[122] Therefore, it is good practice to evaluate the statistical significance of every accuracy increase. This can be done with the WEKA Experimenter, which allows to compare model performances statistically, returning not only average performance measures for CV but also calculated paired *t*-tests to compare CV results of one classifier against the results of another chosen baseline model and indicates if the performance increase is significant.

The approaches using an ordinal representation of the data consist of a bivariate analysis of Spearman rank correlations and a stepwise backwards selection approach using mixed-effects regression modelling, both performed with R using the packages `lmer` for the regression model and `effects` for the visualisations of variable effects.

A potential bias introduced by the different raters is considered by including the rater in the feature set in the classification experiments as well as by interpreting the differences in bivariate correlation analysis when repeating the analysis for each rater individually. In the regression models the rater is added as a random effect.

Below I describe the experiments conducted in this section.

10.1.1 A data mining approach with WEKA – Comparing algorithms and task definitions for a naïve predictive model

In this first approach to text quality prediction we pretend to have no knowledge of the data and predict holistic grades, as well as identify and describe relationships between the holistic grade labels and the other variables retrieved via the text analysis questionnaire.

All completed questionnaire forms that received a holistic grade label (646) were used in the experiments of this section. The assigned holistic grade label represents the class variable or dependent variable that will be predicted in the subsequent steps. The variable presents a 5-point scale of text quality. The remaining non-free-text variables from the text analysis questionnaire serve as predictor

[121] Although other evaluation metrics like the F-score are preferred in NLP (as it accounts for differences in precision and recall cf. section 4.3.3.2), the accuracy metric is simpler and more interpretable. As we are mainly interested in establishing that the classification model did indeed learn something from the data and represents meaningful patterns, it should suffice for the purpose of this study.

[122] The initial splitting of the ten parts is done randomly (although without replacement).

variables. They are originally coded as categorical variables that can have between two (binary) and five levels. In the most straightforward approach, these are used without any transformation or selection.

Without prior knowledge or inspection of the data we have the following possibilities to predict and inspect the quality of the texts:

(1) To train, evaluate and interpret a text classifier that distinguishes between all five grade levels

    This approach needs no further transformations of the data or the task and is the most direct interpretation of the research question. However, the different levels of the holistic grade are treated as equivalent categorical values and the classification model will not assign any similarity between the class labels '*4'* and *'5-excellent'* or *'2*' and *'1-insufficient'* but treat them all equally. Hence, the implicit information of the order of grade levels will be ignored completely. Furthermore, multi-class classification is not a trivial task in machine learning and many classifiers need certain workarounds to perform multi-class classifications (e.g. by conducting a series of binary classifications and returning the class with the highest probability). This affects the performance and interpretability of classification models, as will be visible in the results. The number of classes also affects the general task complexity, having negative effects on performance and interpretability (the researcher has to interpret not only why some texts are better than others, but also why texts have received a good, average or insufficient grade).

(2) To train, evaluate and interpret a text classifier that distinguishes between three classes by conflating '*4*' and '*5-excellent*' grades to '*good*', and the grade levels '1-*insufficient'* and '*2*' to '*weak*' grades.

    By grouping the classes that have less instances, we can decrease class imbalance as well as complexity of the classification task, which in turn leverages prediction performance and interpretability. However, this approach also ignores the rank information encoded in the scale-variable of holistic grades. Furthermore, the conflation of grade levels could introduce a bias in the data in case '*4*' or '*2*' grades are actually closer to the average grade than to the extremes of the grading scale.

(3) To train, evaluate and interpret a text classifier that identifies good/insufficient texts and reformulate the task into a binary classification task.

    A third way to reformulate the task is to only predict one grade level in a binary scheme (e.g. is text graded as '*5-excellent'* or not). This naturally minimizes the complexity of the task and might enhance classification performance and interpretability substantially. However, this might also create very unbalanced training sets, especially if the aim is to identify classes with few instances. For illustration I try to predict two different binary classifications:

    a) insufficient grades vs. all other grades
    b) grade levels *'5-excellent'* and *'4'* vs. all other grades

To illustrate the effects of the different possibilities of task definition, estimates for the predictive performance of the resulting models for these three approaches are compared in the following sections.

By comparing the trained classifiers not only against the baseline but also against each other, it is possible to observe and leverage the strengths and weaknesses of the different algorithms as well as

the difference in complexity and interpretability between different task formulations. Through this methodology we will get a first impression of the complexity of the problem of predicting text quality from answers to the text analysis questionnaire and will be able to observe differences in classification performances depending on task complexity and used algorithm.

10.1.2 Improving the model with feature engineering

Using a dataset for prediction without any prior knowledge of the data, does certainly not suffice to explore or analyse a dataset well. In the next set of experiments, I therefore illustrate methods to remove noise and refine the feature set that is used for a prediction task.

When we look at the data from the text analysis questionnaire (e.g. using mosaic plots, grouped bar plots or contingency tables), we notice several problematic characteristics of this dataset.

- Some of the variables are conceptually similar to other variables. E.g. Q48_koherent ("Does the text appear coherent overall?") is phrased very similarly to Q50_konzeptionell_zusammenh ("Is the text conceptually coherent?") and is indeed distributed similarly in the dataset.
- Some of the variables have a high number of different levels that are not always very distinct from each other. E.g. Q44_Textgliederung_inhaltlich_graphisch_unterstuetzt („Does the formal and textual structure support the reception of the text?") has the factor levels *'(Mainly) no'*, '*Partly'*, *'(Mainly) yes'* and *'not determinable'*.
- Distributions of factor levels are highly skewed for some of the variables. E.g. for Q56_Konnektorenprobleme ("Does the use of connectives cause problems?") 92% of the values were "The use of connectives doesn't cause any problems".
- Some variables contain a factor level *'nicht bestimmbar'* (not determinable) that might be considered as an extra level or as a missing value. Treating them as extra class might be conceptually controversial but treating them as missing values might impede the analysis substantially. E.g. treating *'nicht bestimmbar'* as missing value and excluding the respective observations for the variable Q42_Angekuend_Textfunktion_eingehalten ("Does the text adhere to the announced text functions?") results in a predictor variable with practically no variance and therefore potentially no contribution to the overall model.

In the subsequent experiments, I therefore explore four different strategies to refine the feature set, removing unwanted noise from the data and reducing the complexity of the feature set.

(1) Testing bivariate association of variables with the text quality using the chi-squared test for independence and removing insignificant variables

In this first refinement attempt I investigate if reducing the noise in the feature set by removing variables that do not show monofactorial relations with the holistic grade changes the prediction performance and hence the quality of the language model. I test the bivariate association between the dependent variable and all predictor variables individually using the chi-squared test for independence. This test, as all statistical tests, is generally not suited for multiple testing, as repeated tests might eventually lead to false significant results. However, in this experiment I am using this approach primarily to narrow down the analysis and exclude irrelevant information for further steps and not in order to tell something about the relations itself. I therefore accept a possible error induced by multiple testing. Nevertheless, I use a slightly lower significance level (p-value < 0.01) for this test. I repeat the classifier training for predicting five grade levels from the previous experiment and compare the results for the refined and unchanged feature set.

(2) Using the built-in feature selection methods of WEKA to restrict the number of features to the 10/25 best-ranked features

The chi-squared test for independence is just one way to identify or pre-evaluate the importance of a single variable for the dependent variable. Other measures like information gain, gain ratio, gini index or Pearson's product-moment correlation are common metrics that are used for (pre-classification) feature selection (see section 4.3.2.3). The WEKA data mining software implements several feature selection and ranking methods that preliminarily remove irrelevant variables (e.g. in the case of the `cfsSubsetEval`, a wrapper method by Hall (1998) that evaluates the predictive ability of individual features and of feature subsets while minimizing intercorrelation between the features) or rank the variables in terms of importance so that the feature set can be reduced to a certain number or percentage according to this rating (e.g. `InfoGainAttributeEval`).

(3) Trying different treatments for missing values (e.g. imputation, replacement or exclusion)

The annotators in the text analysis questionnaire had the possibility to choose '*not determinable'* in many of the questionnaire items. This answer possibility introduces values that could be seen either as a separate category, and thus carrying important information, or as missing values, and thus be excluded, treated differently or even filled with replacement values (i.e. imputed) when training a classification model.

(4) Conflating feature levels conceptually to reduce the variance and skewedness in the predictor variables

As a last experiment on refining feature sets, I look at the variable set more carefully, and reduce noise by conflating variable levels that can be conceptually grouped together. Whenever possible, the variables were conflated to two different levels, excluding the category 'not determinable'. In that way I attempt to remove variance and skewedness in the predictor variables and might therefore improve the classification models. The list of questionnaire items in the appendix shows simplifications made for this experiment. The variables marked with the suffix *'.rec'* are conflated, simplified versions of the original items. The models trained for this experiment use either the conflated version of an item (if a simplified version was made) or the original version or both.

By removing unnecessary or uninformative variables or variable categories from the training data and refining the feature set with the strategies described above, the predictive performance could increase. Hence, I repeat the analysis from section 10.2.1.1 with the refined feature sets and compare the classification performances.

The experiments use the WEKA data mining software to train different classifiers and systematically compare the prediction performance. The results of these experiments primarily allow an interpretation on a metalevel, showing which algorithms, task definitions and variable sets produce the most accurate model of the data. However, as we want to know more about the relationships in the data, we also want to interpret the models on a data level.

10.1.3 Basic interpretation methods for text classification – Comparing feature sets, feature importances, and interpreting classifier outputs

This section shows three different strategies on how to interpret the data. All three strategies can be done directly within WEKA.

(1) Systematic classifier comparison

One way to do this is to use the same comparison strategy as before, comparing the performance of classifiers trained on two or more hand-picked features or feature sets. As example I compare completeness related features (section A), text structure related features (section C) and content-related features (section B) from the text analysis questionnaire with each other and inspect the difference in classification performance.

(2) Variable importance through feature selection

Another ready-to-use interpretation method is to inspect and compare feature importance measures, i.e. feature weights provided in classification reports or via feature selection methods. I compare the top ten features chosen by the built-in WEKA feature selection methods to identify important variables.

(3) Interpretable models

At last, I inspect the model internals for intrinsically interpretable classification methods like decision trees and rule learners, using their WEKA outputs and built-in visualizations methods to interpret the structure of the data.

10.1.4 Holistic grade as a numerical (ordinal) value – Correlations and mixed-effects regression modelling

The data in feature set 1 was originally coded as categorical variables, describing the type of certain text characteristics, but in some cases also their degree or existence in general. As can be seen in the detailed description of questionnaire items in section 9.2.1, many variables are thus conceptually binary or scale variables that could also be represented in an ordinal or numeric form instead of categorically (e.g. `Q50_konzeptionell_zusammenh` with its answer options '*conceptually incoherent'*, *'partially conceptually coherent'* and *'conceptually coherent'* is an inherently ordinal item that represents a scale of three levels). Treating such rank variables as ordinal (numerical) data allows to use other analysis methods that can be more appropriate for modelling ordinal holistic grades (correlation analysis, regression modelling, etc.) and whose interpretation is also more straightforward. I therefore recode the conceptually ordinal or binary questionnaire items of feature set 1 into numerical variables, indicating a scale of better (higher) or worse (lower) classifications for the individual evaluations. I used the conflated variables from experiment 10.2.2.1 to turn categorical variables into binary variables, when a rank order was not obvious. Moreover, I excluded `Q19_Thema_Gesamttext` as it was conceptually not interpretable in a ranked order nor transformable into a binary variable with one preferable category. Given this new data representation, one can make use of methods that are better suited to analyse ordinal/numeric dependent variables.

(1) Correlation analysis

I calculate bivariate Spearman rank correlations for the resulting numerical/binary variables and the holistic grade and test their statistical significance. The obtained p-values and correlation coefficients allow us to identify which items of the text analysis questionnaire are significantly correlated with the overall grades (i.e. are relevant) and how strong this correlation is. Contrary to the feature ranking approaches used in section 10.2.3.2 where I investigated just the variable importance, this approach allows us to investigate also the direction and magnitude of the relationships (i.e. variable effect).

I assume that higher values in the text analysis evaluations result in higher values for the holistic grades (one-tailed test design) and use a confidence level of 0.99 to judge significance, meaning that I set the probability of wrongly rejecting the null hypothesis to an alpha level of 1%.

A deeper analysis of the bias introduced by the annotators observes rater effects by calculating bivariate correlations for each subset of texts annotated by one rater as well as by visualizing differences in grade means for each annotator with interaction plots.

(2) Mixed-effects regression modelling with stepwise model selection

In order to model the ordinal dataset in a multifactorial analysis, I build a regression model with the ranked ordinal variables. A random effects model structure is used to account for the fact that evaluations have been drawn from different annotators (i.e. repeated measures per annotator) and that the data are thus not independent[123]. To narrow the scope of the model, I consider only those variables that have a correlation coefficient higher than 0.2 in the preliminary correlation analysis as predictor variables. I test the mixed-effects model against a model that does not account for random effects and use a) automatic stepwise regression modelling with backward model selection and b) manual stepwise regression modelling with forward model selection and compare the results.

The `lmerTest` package of R (function `step`) is used for the automatic stepwise regression with random effects. The function automatically tests all possible factors for dependence and iteratively eliminates the factor with the highest p-value from the scope of independent variables, first for random effects (log-likelihood ratio test), then for fixed effects (F-test) in the model. It then returns a summary of the final (best) model that only contains factors that significantly contribute to explaining the dependent variable. A secondary manual stepwise regression analysis was performed with the same scope of independent variables using the `lme4` package of R (mixed-effects structure) to check the automatic approach. I used the reverse method of adding iteratively the factor with the highest AIC and comparing the model with the previous smaller model using F-test, to see if the new factor significantly improves the model.

(3) Interactions and multicollinearity

Finally, possible interactions and multicollinearity in the data are investigated using the results and visualization of a correlation matrix and controlling the values for the variance inflation factor for the regression model built in the section before.

10.2 Results

10.2.1 A data mining approach with WEKA – Comparing algorithms and task definitions for a naïve predictive model

In the following I compare the prediction performance of different text classifiers, trained to predict text quality on the basis of the text analysis questionnaire items of feature set 1 (cf. section 9.2.1).

As the goal is to find structure in the data, I compare the classifiers to the accuracy that would be achieved if the data would be random and the classifier would only know about the base distribution of the classes of the dependent variable. As a first step, I therefore define a prediction baseline for the

[123] See for example the suggestions by Gries (2013).

following experiments, against which all classifier performances are evaluated (see section 4.3.3.2). I define the baseline accuracy as the percentage of the majority class (henceforward called *majority baseline*), i.e. the accuracy that would be achieved if the classifier would not learn anything from the data but would consistently assign the most common label[124]. Although it is common practice in NLP to consider also alternative baselines like the performance achieved with a standard model or even with the latest state-of-the-art model published for the data, this is approach is not reasonable in situations where one wants to interpret the relationships in the data rather than beat the state-of-the-art.

Given the grade level distribution for the text quality evaluations, the majority baseline for the dataset of 646 text evaluations is 37.9%, according to the percentage of grade level 3.

10.2.1.1 Prediction of five grade levels

Next, I train a naïve text classification model to distinguish between the five levels of the holistic grade, using all the items of the text analysis questionnaire for training the model. Table 20 shows 10-fold CV results for classifier performance (accuracy, F-score, precision, and recall; see section 4.3.3.2) for different classification algorithms. The values printed in bold are the best performing classifiers for the respective performance measure. The values marked with an asterisk are significant according to a paired *t*-test performed on the ten runs of the 10-fold CV with a significance level of 99.9%. All classifiers performed significantly better than the baseline (paired *t*-test, p-value < 0.001)

Classifier	Accuracy	F-Score	Precision	Recall
Naïve Bayes	55.83*	0.56*	0.58*	0.56*
Logistic Regression	51.31*	0.51*	0.53*	0.51*
PART Rule learner	49.44*	0.49*	0.50*	0.49*
J48 Decision Tree	53.90*	0.53*	0.55*	0.54*
Multilayer Perceptron	56.87*	0.56*	0.58*	0.57*
SMO (polynomial kernel)	56.88*	0.57*	0.58*	0.57*
Random Forest	**59.64***	**0.58***	**0.60***	**0.60***
Baseline	*37.93*	*0.21*	*0.14*	*0.38*

Table 20: Comparison of different algorithms and different evaluation methods for a naïve classification approach predicting five grade levels.

10.2.1.2 Prediction of three grade levels

Next, I conflate the five original grade levels to three levels by combining the grade levels at the poles ('*4'* and '*5-excellent'* become '*good'*; '*1-insufficient'* and '*2*' becomes '*weak'* and '*3'* becomes '*average'*) and repeat the experiment.

[124] The ZeroR classifier in WEKA implements a classifier that imitates such a classifier and can be used to get accuracy, F-score or other evaluation metrics for a majority baseline.

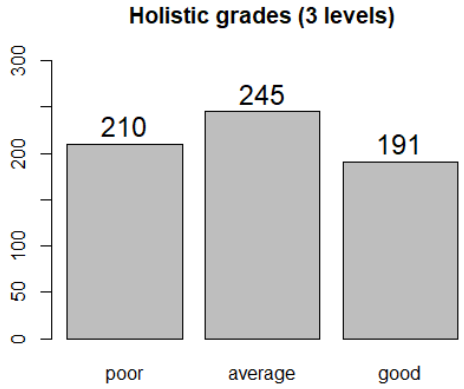


Figure 30: Class distribution for holistic grades with three grade levels.

Figure 30 shows the new class distribution that is now almost equally distributed. The baseline for the task is still the same as the grade '*3*' is still the most frequent class. Table 21 shows the results of 10-fold CV for the prediction task with reduced complexity and reduced class imbalance.

Classifier	Accuracy	F-Score	Precision	Recall
Naïve Bayes	66.41*	0.66*	0.67*	0.66*
Logistic Regression	61.81*	0.62*	0.62*	0.62*
Multilayer Perceptron	65.27*	0.65*	0.66*	0.65*
SMO (polynomial kernel)	65.85*	0.66*	0.66*	0.66*
PART Rule learner	60.80*	0.61*	0.61*	0.61*
J48 Decision Tree	64.38*	0.64*	0.65*	0.64*
Random Forest	**68.66***	**0.69***	**0.70***	**0.69***
Baseline	*37.92*	*0.21*	*0.14*	*0.38*

Table 21: Comparison of different algorithms and different evaluation methods for a naïve classification approach predicting three levels of holistic grades.

As expected, the evaluation results were significantly better for this task (paired *t*-test, p-value < 0.001), with an increase of around 10% in accuracy and 0.10 for F-score values. Again, the random forest classifier outperformed the other algorithms.

10.2.1.3 Binary classification for '4' and '5-excellent' vs. all other grades and for insufficient vs. all other grades

Finally, I transform the class labels one more time, to see how reducing the complexity to a binary classification changes the results. The resulting distributions can be seen in Figure 31.

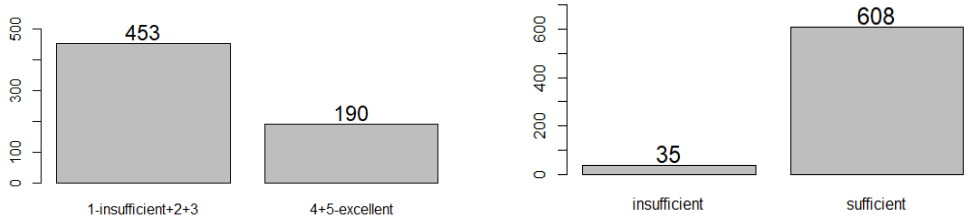


Figure 31: Distributions for two different versions for binary classification.

Table 22 shows prediction performances (weighted average for 10-fold CV) for excellent and good texts vs. all other grades (class distribution: 191 '*5-excellent*' or '*4*', 455 not '*5-excellent*' or '*4*'). Table 23 shows the results for insufficient vs. all other texts (class distribution: 35 '*1-insufficient*', 611 other). The baseline for both classification tasks changed substantially, as the most frequent class label makes up ca. 70% in the grouping ('*4*' and '*5-excellent*' vs. other) and ca. 95% in the second (insufficient vs. other). We now have once a rather imbalanced and second a very imbalanced class distribution. While for the classification of excellent and good grades we still achieve evaluations that are significantly better than the baseline, the task of identifying negative (i.e. insufficient grades) shows accuracies, precision, recall and F-score values above the baseline, but none of these results was significant (paired *t*-test, p-value < 0.05), i.e. the increase of prediction performance over the baseline majority is not high enough to reject the hypothesis that the classifier performs equal to the majority class classifier.

Classifier	**Accuracy**	**F-Score**	**Precision**	**Recall**
Naïve Bayes	79.16*	0.80*	0.83*	0.79*
Logistic Regression	78.99*	0.79*	0.79*	0.79*
PART Rule learner	78.45*	0.78*	0.79*	0.78*
J48 Decision Tree	80.19*	0.80*	0.81*	0.80*
Multilayer Perceptron	81.63*	0.82*	0.82*	0.82*
SMO (polynomial kernel)	81.06*	0.81*	0.82*	0.81*
Random Forest	**84.52***	**0.84***	**0.85***	**0.85***
Baseline	*70.43*	*0.58*	*0.50*	*0.70*

Table 22: Excellent and good grades vs. all other.

Classifier	**Accuracy**	**F-Score**	**Precision**	**Recall**
Naïve Bayes	89.52	0.92	**0.95**	0.90
Logistic Regression	91.63	0.92	0.94	0.92
PART Rule learner	94.25	**0.94**	0.94	0.94
J48 Decision Tree	**95.28**	**0.94**	0.93	**0.95**
Multilayer Perceptron	94.45	**0.94**	0.94	0.94
SMO (polynomial kernel)	94.06	**0.94**	0.94	0.94
Random Forest	95.03	0.93	0.92	**0.95**
Baseline	*94.59*	*0.92*	*0.89*	***0.95***

Table 23: Insufficient vs. all other grades.

These experiments point out that the baseline is actually crucial for establishing if a classifier learned to predict the holistic grade from the data or not. Although the overall accuracy for the last experiment (distinguishing insufficient graded texts from the others is much higher than the one for the other experiments, the models do not necessarily encode valuable information other than the base distribution of the classes. A model that would always predict the majority class apart from a few random cases where it correctly predicts the other class might get to the same results. Besides, these results are not directly comparable to the previous experiments. Having differing baselines changes also the difficulty for the classifier to pass the baseline. A lower baseline is easier to pass than a higher one. However, most learning algorithms always reach the majority baseline, as they are designed to learn at least the class distribution in the data.

There are some strategies to compare differing baselines in classification experiments. One method is for example to use the Kappa statistic (Cohen's Kappa) that calculates the increase over the baseline in relation to the space left between the baseline and 100% (Ben-David, 2008).

Calculating Cohen's Kappa above we can compare the prediction performance of the four different task definitions (Table 24).

Classifier	5 levels	3 levels	good-excellent	insufficient
Naïve Bayes	0.40	0.50	0.56	**0.43**
Logistic Regression	0.34	0.42	0.50	0.35
PART Rule learner	0.40	0.48	0.56	0.40
J48 Decision Tree	0.40	0.49	0.55	0.41
Multilayer Perceptron	0.30	0.41	0.48	0.33
SMO (polynomial kernel)	0.36	0.47	0.53	0.29
Random Forest	**0.43**	**0.53**	**0.62**	0.19

Table 24: Comparing predictive performance or different task definitions.

Again, the random forest classifier has the best values in most of the tasks. However, in the case of the very unbalanced identification of insufficient grades, it performed much worse than simpler algorithms like the Naïve Bayes classifier. In general, using the items of the text analysis questionnaire, it is easier to identify good or excellent grades (prediction of two class levels, with not too imbalanced distribution) than to predict three grade levels with almost balanced distribution or two class levels with very imbalanced distribution. Predicting five grade levels is, next to the identification of imbalanced groups the most difficult task in this scenario.

Summary
The results show that the prediction of holistic grades based on the items of the text analysis questionnaire is possible and gives performances that significantly exceed the majority baseline (p-value < 0.001). The accuracy of the best performing classifier that considers all five grade levels, i.e. the most complex task, is clearly above the baseline (59.64% with the random forest classifier in WEKA compared to 37.92% ZeroR classifier). However, reducing the complexity of the prediction task by reducing the number of class levels to three levels (good, average and weak grades) so that levels with low frequency are grouped together or by reformulating the class labels into a binary distinction ('*1-insufficient*' vs. other grades, '*4*' and *'5-excellent'* vs. other grades) clearly enhances prediction results (e.g. 68,6% for the prediction of three grade levels, and 84,% when predicting the best two grade levels vs. the others and 95,28% when predicting insufficient grades) but might also introduce further class imbalance that lowers the performance of most classification algorithms. When investigating Cohen's Kappa for the four prediction tasks, we could therefore find the best prediction performance for the binary classification identifying good (*'4'*) and excellent grades and discriminating them from the other grades.

10.2.2 Improving the model with feature engineering

In the first experiment, we could get a general understanding of the effect of different algorithms, different evaluation metrics and different task complexities when using text classification as a method to understand the data. We used a naïve approach, taking feature set 1 (text analysis questionnaire) almost without any preliminary investigation and modification. In the next step, I elaborate on this approach, refining the variable set used for classification considering the nature, distribution and monofactorial relations of the variables with the holistic grade.

10.2.2.1 Conflating feature levels conceptually to reduce the variance and skewedness in the predictor variables

As a first experiment on refining the feature set, I look at the variable set more carefully, reducing noise by conflating the multinomial variable levels in the text analysis questionnaire. For this low-frequency variable categories that can be grouped together conceptually (e.g. grouping categories such as '*the essay was rather not coherent*' and '*the essay was not coherent*' to oppose '*the essay was coherent*') have been conflated (cf. section A in the appendix).

However, comparing classifier results for the 5-class prediction of holistic grades shows the following results (see Table 25). Neither replacing the non-conflated variables with their simplified versions, nor adding the simplified versions to the original feature set gave better CV estimates for accuracy. However, their prediction performance also did not drop significantly. While the prediction performance does not benefit from removing category levels, the interpretability of the model might still increase, as the models get significantly simpler for a human to interpret. It is therefore an option to still consider the simpler model (conflated) although the absolute prediction performance is lower.

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
conflated_unconflated	37.93	53.74	51.73	49.07	51.23	55.20	55.22	57.35
conflated	37.93	55.39	52.20	48.12	53.18	55.48	55.23	58.94
original	*37.93*	*55.83*	*51.31*	*49.44*	*53.90*	*56.88*	*56.87*	***59.64***

Table 25: Comparison of classifier accuracies for simplified feature sets.

10.2.2.2 Trying different treatments for missing values (e.g. imputation, replacement or exclusion)

Next, I repeat the 5-class classification experiment from above using three different approaches for dealing with possible missing values introduced by the questionnaire option *'not determinable'*. When loading the dataset into WEKA I defined them once as a separate class, once as missing values, and once used a simple imputation technique provided by WEKA to fill such entries with the mode (most frequent value) for each variable. However, as Table 26 shows, there were no significant performance differences between the different methods. Prediction accuracy hints to slightly better results when not determinable values are treated as extra category, but the values were not significant at alpha 5% (paired *t*-test).

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
extra_cat *(original)*	37.93	55.83	51.31	49.44	53.90	56.88	56.87	**59.64**
missing	37.93	56.32	52.13	49.78	53.67	55.82	56.06	**58.04**
missing_imputed	37.93	54.74	52.27	47.93	51.32	55.60	53.32	**57.21**

Table 26: Prediction performance for different treatments of missing values.

Considering reasons of interpretability, one might still consider treating the not determinable item levels as missing values, as it renders the model considerably simpler.

10.2.2.3 Testing bivariate association of variables with the text quality using the chi-squared test for independence and removing insignificant variables

In a preliminary test for bivariate associations between the holistic grades and the predictor variables using the chi-square test for independence, the following variables of feature set 1 (see section 9.2.1) did not show any significant dependence at alpha 0.01:

- Q07_weitere_Textteile
- Q08_pers_Stellungnahme
- Q11_Stellungnahme_vs_Enzensberger
- Q13_expliziter_Bezug_Input
- Q23_weitere_Form_Themenentfaltung_vorhanden
- Q29_Argumentationsstrategien
- Q31_explizite_Adressatenorientierung
- Q35_abwiegen_und_enschraenken
- Q39_Umbrueche_gelungen
- Q40_weitere_Umbrueche
- Q41_Anfang_Textfunktion_angekuendigt
- Q47_Qualitaet_konsistent
- Q53_Fazit_zu_Input
- Q56_Konnektorenprobleme
- Q65_humorvoll_ironisch

This information is used in order to test the effect of refining the feature set for classifier training by removing the insignificant variables from the feature set. Again, I compare the classifier performances for 10-fold CV for the full feature set and the reduced feature set (Table 27). Although some of the classification algorithms (e.g. logistic regression) had small improvements with the reduced variable set, none of these differences was significant. We can therefore assume that preselecting associated variables is not particularly relevant for standard classification approaches in this field. However, it might increase interpretability by making the model simpler.

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
full set (*original*)	37.93	55.83	51.31	49.44	53.90	56.88	56.87	**59.64**
reduced set	37.93	56.51	53.69	51.73	52.95	57.66	56.24	59.53

Table 27: Prediction performance for the original model and a model with only correlated features.

10.2.2.4 Using the built-in feature selection methods of WEKA to restrict the number of features to the n-best ranked features

The WEKA Explorer interface provides feature selection methods to rank and/or select features that are important according to a particular usually bivariate association measure. Of-the-shelf the software offers a selection using the wrapper method `CfsSubsetEval` (for a description of wrapper methods see section 4.3.2.3), rankings on the bivariate association measures correlation, gain ratio and information gain as well as the feature selection algorithms `OneR`, `ReliefF`, and `SymmetricalUncertAttributeEval`). I use these methods in order to extract reduced feature sets for feature set 1. In case of `CfsSubsetEval` all the selected variable are chosen. In case of the other feature ranking methods, I test classifiers based on the top ten (20% of the feature set) and the top 25 features (50% of the feature set).

In particular I test[125]

- CfsSubsetEval
- Top 10 features for `CorrelationAttributeEval`
- Top 25 features for `CorrelationAttributeEval`
- Top 10 features for `GainRatioAttributeEval`
- Top 25 features for `GainRatioAttributeEval`

[125] The WEKA feature selection methods `Principal Component Analysis` and `WrapperSubsetEval` have been excluded as they are not suited for this task.

- Top 10 features for InfoGainAttributeEval
- Top 25 features for InfoGainAttributeEval
- Top 10 features for OneRAttributeEval
- Top 25 features for OneRAttributeEval
- Top 10 features for ReliefFAttributeEval
- Top 25 features for ReliefFAttributeEval
- Top 10 features for SymmetricalUncertAttributeEval
- Top 25 features for SymmetricalUncertAttributeEval

In the following, I show average predictive performance calculated with 10-fold CV, for each of the 13 reduced feature sets in comparison to the original full feature set for the prediction of five grade levels (Table 28).

Dataset	**Baseline**	**NB**	**LogReg**	**PART**	**J48**	**SMO**	**MLP**	**RandF**
cfsSubset	37.9	57.42	55.66	52.52	54.78	58.01	53.55	57.10
corr10	37.9	55.40	54.40	52.90	54.30	54.50	51.80	53.00
corr25	37.9	54.45	52.34	48.86	52.40	55.89	51.23	53.80
gainratio10	37.9	57.45	55.59	52.89	54.03	56.06	52.64	53.99
gainratio25	37.9	58.79	54.37	50.60	53.00	58.09	54.95	57.98
infogain10	37.9	55.41	54.20	52.41	53.30	55.22	52.26	53.70
infogain25	37.9	56.94	53.07	50.56	53.59	58.29	53.87	58.00
oneR10	37.9	51.88	54.87	47.66	49.74	53.79	46.89	50.06
oneR25	37.9	55.15	49.97	49.75	51.70	56.39	50.46	53.39
relief10	37.9	56.05	**60.18***	54.56	56.10	58.83	55.34	56.10
relief25	37.9	58.43	55.84	52.35	54.98	59.51	56.54	**61.09**
sym10	37.9	55.41	54.20	52.43	53.30	55.24	51.38	53.49
sym25	37.9	58.24	55.18	50.57	53.50	57.83	54.37	58.58
all	*37.9*	*55.83*	*51.31*	*49.44*	*53.90*	*56.88*	*56.87*	*59.64*

Table 28: Comparison of accuracies for built-in feature selection methods.

We can see that not all of the algorithms actually improved when reducing the feature set with feature selection methods before classification. In general, the classification method that improved most, was the classification with logistic regression. Using the RefliefFAttributeEval feature ranking and reducing the training features to 20% of the actual number of features (10 features) improved prediction performance in 10-fold cross validation significantly (p-value < 0.01, paired *t*-test) when compared to the previous results for the whole feature sets (60.18%). However, this result is not significantly better than using a random forest classifier with the full feature set.

Furthermore, when comparing the different feature selection methods against each other, we find that the ReliefFAttributeEval and thus the recently frequently used ReliefF algorithm consistently achieves high results even when different algorithms are used.

We can conclude that reducing the feature set for classifier training with preliminary feature ranking via the built-in methods provided in WEKA only helps certain categories of algorithm. The well-performing random forest algorithm that also showed the best prediction results in almost all the previous experiments (besides the prediction of the very unbalanced dataset for insufficient vs. all other grades) usually exceeds the improvements that are made by reducing the variable set but does itself not profit significantly from a reduced feature set.

However, although classification performance did not increase significantly by reducing the feature set, the comparison of different feature ranking results for different variable importance measures like information gain, gain ratio, etc. gives insights on monofactorial relations in the data. The rankings of the used feature selection methods show relatively homogeneous choices. Consistently high-ranked variables can be identified as particularly relevant for the holistic grade of an essay (see section 10.2.3.2).

Summary

In this section various methods for refining the used feature set and reducing its complexity have been applied to the data and tested in a 5-class prediction scenario. While some of the learning algorithms seemed to profit from the reduced complexity in the feature set, the best predictive performance was in general the achieved with the random forest classifier and the original full feature set. Only the best 25 features selected by the ReliefF algorithm yielded, with the random forest classifier, a better prediction performance than the original feature set with random forest. However, the increase in accuracy was not statistically significant (alpha = 0.01, paired *t*-test). While no significant increases could be found for any learning algorithm by changing the treatment of missing values, conflating variable levels or reducing the feature set to monofactorially related features, the ReliefF feature selection (Kononenko et al., 1997), significantly improved classification results for the logistic regression classifier. In general, the method performed better than other feature selection methods provided by WEKA and showed good results for all of the classification algorithms used.

10.2.3 Basic interpretation methods for classification algorithms

Section 10.2.1 and 10.2.2 showed various strategies for improving prediction models by changing algorithms, task definitions, and feature sets. Although refining the variable set by feature selection or variable level conflation did not produce better performing models (in these experiments), reducing the complexity of the data might still be relevant for model interpretability (see section 3.6). Identifying relations and interactions in the data is more straightforward, if we can rely on clearly defined and interpretable variables and feature ranks. Thus, I explore some basic interpretation methods available within WEKA.

10.2.3.1 Systematic model comparison for different feature sets

As a first interpretation method, I use the same approach as in the experiments before and compare the prediction performance of differently trained classification models. However, instead of methodologically motivated comparisons, I select linguistically motivated feature sets, comparing the first three groups of items in the questionnaire with each other. The feature sets used for comparison in this experiment are thus composed of a) features related to the formal completeness of a text, b) features related to the content of a text and c) feature related to the structure of a text[126].

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
completeness	37.93	**43.37***	44.64*	43.00*	44.50*	42.67	42.72*	42.60
content	37.93	39.50	**48.49***	**45.12***	**44.94***	**49.42***	**44.07***	**46.59***
structure	37.93	37.42	36.89	39.11	38.17	36.59	38.73	39.55
all	*37.93*	*55.83*	*51.31*	*49.44*	*53.90*	*56.88*	*56.87*	*59.64*

Table 29: Comparison of classifier accuracies for feature sets regarding textual completeness, content and structure of the text.

[126] Feature of group d) regarding the overall impression of the annotator are not considered in this comparison as they combine different aspects that already occur in the first three feature sets.

Table 29 shows the classification performance measured in 10-fold CV accuracy for the three feature sets. Results with asterisk are significantly better than the baseline and thus show that the classifier learned from the structure in the data. We see that there was no classifier that achieved a significant performance increase over the baseline when trained only on the questionnaire items regarding the structure of the text (group C of section 9.2.1). However, the variables regarding the completeness of the text and regarding its content were relevant for the prediction of the holistic grades. The best performing model was build based on the content-related variables using an SMO classifier.

10.2.3.2 Variable importance via feature ranking

Next, I investigate and compare the feature ranks produced with the built-in feature selection methods of WEKA for section 10.2.2.4. The generated lists of ranked features show the strength of relation or importance for prediction of individual metadata variables for the holistic grades. Combining the outcomes of various ranking methods can thus give an overview of monofactorial relations in the data. There are seven items of the text analysis questionnaire that are ranked among the top ten variables in each of the feature selection approaches provided off-the-shelf by the WEKA software.

- Q48_koherent
- Q49_Gesamtthema_nachvollziehbar
- Q50_konzeptionell_zusammenh
- Q51_inhaltlich_klarer_Aufbau
- Q58_Roter_Faden
- Q61_hat_gefallen
- Q62_inhaltlich_nachvollziehbar_klar

Furthermore, `Q60_ueberzeugende_Argumentation` and `Q63_interessant_langweilig` were chosen in six of the seven methods. `ReliefFAttributeEval` feature selection was the only one to choose annotator within first ten features. This information already gives us first insights about this data and our notion of text quality.

However, when using these seven features for training text classification models to predict five grade levels, the models report lower accuracies in 10-fold CV than using the full feature set (Table 30). In the case of the random forest and the multilayer perceptron the results are significantly worse (p-value < 0.05, paired *t*-test).

Dataset	**Baseline**	**NB**	**LogReg**	**PART**	**J48**	**SMO**	**MLP**	**RandF**
best monofactorial	37.93	53.09	54.62	51.89	51.49	52.24	49.64	50.54
ReliefF best 5	*37.93*	*54.40*	*55.14*	*54.77*	*55.64*	*52.11*	*53.70*	*55.00*
all	*37.93*	*55.83*	*51.31*	*49.44*	*53.90*	*56.88*	*56.87*	*59.64*

Table 30: Classification with features that were consistently ranked within the top ten features for all feature selection methods.

The monofactorially best features are not necessarily the features that complement each other best, when predicting holistic grades on student essays. The seven features tested here yield worse results than the best five features indicated by the ReliefF algorithm (see also section 4.3.2.3).

10.2.3.3 Interpreting model internals

Another way to interpret the data is thus to interpret the model internals itself. Given the model performs better than the baseline and thus learned something from the data, the internal structures of the model can be most informative for the interpretation of the task. However, the possibilities and means for interpretation of model internals differ from algorithm to algorithm and from one machine

learning or data mining software to the other. Within the WEKA Explorer interface, there are a few possibilities to inspect the structures that have been learned by the model without having to compare just monofactorial relations or model performances. The possibilities are, however, mostly restricted to the output that is provided after training the model. This output reports for any chosen algorithm the specifics of the classification task (classifier and configurations, number of observations, number and names of used features, and the evaluation strategy or test mode) some standard evaluation measures (like accuracy, training time in seconds, precision and recall, F-score, etc.) as well as a confusion matrix.

Some of the classifiers of WEKA additionally show other internal values, such as:

- variable weights for the SMO (SVM) classifier
- regression coefficients or odds ratios for the Logistic classifier
- rules for the PART classifier
- conditional probabilities for the Naïve Bayes classifiers
- visualizable tree structure for decision tree classifiers like J48.

This section illustrates the interpretation of a simple 5-class prediction model for holistic grades using the outputs provided by the WEKA Explorer interface. For the interpretation examples, a simple model is built with the following model setup:

1. Only the five best features according to the ReliefF feature ranking algorithm are used in the model
    - Q48_koherent
    - Q51_inhaltlich_klarer_Aufbau
    - Q60_ueberzeugende_Argumentation
    - Q61_hat_gefallen
    - annotator
2. Instances evaluated with '*not determinable*' were treated as missing values in this experiment.
3. The underrepresented categories '*not coherent*' and '*rather not coherent*' have been conflated, to remove irrelevant noise.

Although the model is not the best model built so far, the prediction accuracy is still high given the simplicity of the model (cf. Table 31).

.**Dataset**	**Baseline**	**NB**	**LogReg**	**PART**	**J48**	**SMO**	**MLP**	**RandF**
ReliefF 5	37.93	54.02*	55.42*	55.11*	55.73*	52.48*	55.88*	55.26*

Table 31: 10-fold CV accuracy results for a simple, interpretable model using the best 5 features according to ReliefF.

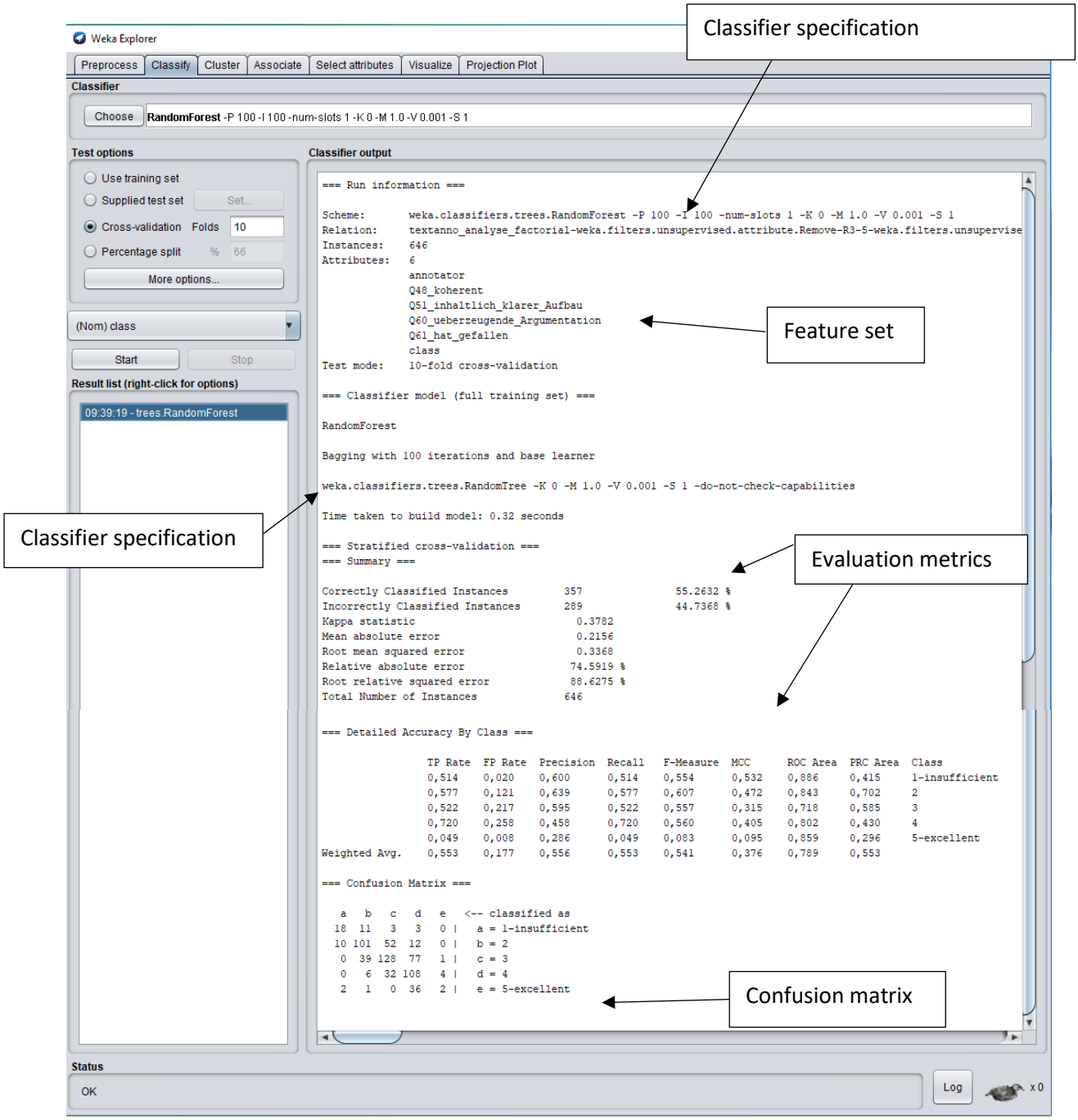


Figure 32: The output of a random forest classifier in the WEKA Explorer Interface.

Figure 32 shows an example output for WEKA's Random Forest classifier trained to predict the holistic grade based on the five features listed above. Once we know that a classifier found some generalizable structure in the data (surpassed the baseline), we can interpret this output in order to say something about the relationships and structures that the classifier learned, evaluating the various metrics as well as the confusion matrix at the bottom of the output.

While the confusion matrix is often used in other projects concerned with the interpretation of machine learning models where classes are not naturally related between each other, it is not very informative in this example as the grades actually represent ordinal values. In the confusion matrix above we see that most classification errors have been made between similar ratings (e.g. '*3'* grades were confused with *'2*' or '*4*' grades and no '*1-insufficient'* grades have been confused with *'5-excellent'* grades). This was to be expected and just confirms that there is an ordinal relationship between the holistic grade levels that can be retraced by the items of the text analysis questionnaire.

NaïveBayes Classifier

More information can be extracted from the conditional probabilities of the Naïve Bayes classifier.[127] Model output 1 shows the output for the NaiveBayes classifier in the WEKA Explorer. The model output gives a summary of the data, listing calculated conditional probabilities for each variable level of each predictor variable for each class. In this output we can see, for example that the prediction probabilities for the five grade levels are almost normally distributed for Annotator B and Annotator C, while skewed towards lower and average grades for Annotator A. Despite the probability that the model predicts the lowest grade (*1-insufficient*) is very low for Annotator A. Moreover, the output shows that there is a relatively strong difference in prediction probabilities of the text appeal variable (Q61_hat_gefallen). The probability of predicting an average grade level of '*3'* for texts that appealed to the annotator is about six times higher than the probability of predicting a grade of '*2'* one level lower. In general, one can observe that higher grades have higher prediction probabilities for positive evaluations (e.g. coherent, appealing texts, with clear structure) and vice versa. However, one has to consider that the extreme grade levels '*1-insufficient'* and *'5-excellent'* will always have lower probabilities as there are only few instances in the data. The interpretation of values in this table (especially horizontally) is therefore limited, although the model is rather simple. The interpretation can easily become tedious and bothersome when adding more variables or variables with many category levels. Moreover, there is no information on possible interactions in the data (i.e. whether one variable means something else depending on the value for another variable).

[127] For a more detailed description see section 3.5.5.

Naive Bayes Classifier

Attribute	Class 1-insufficient (0.06)	2 (0.27)	3 (0.38)	4 (0.23)	5-excellent (0.06)
annotator					
Annotator A	1.0	77.0	87.0	49.0	10.0
Annotator C	15.0	53.0	73.0	59.0	13.0
Annotator B	22.0	48.0	86.0	44.0	21.0
[total]	38.0	178.0	246.0	152.0	44.0
Q48_koherent					
Eher ja	10.0	79.0	136.0	53.0	5.0
Ja	7.0	29.0	100.0	99.0	36.0
Nein	21.0	69.0	12.0	1.0	2.0
[total]	38.0	177.0	248.0	153.0	43.0
Q51_inhaltlich_klarer_Aufbau					
Der Text hat einen inhaltlich klaren Aufbau.	4.0	60.0	137.0	123.0	39.0
Der Text hat keinen inhaltlich klaren Aufbau.	18.0	26.0	6.0	1.0	3.0
Der Text hat einen inhaltlich teilweise klaren Aufbau.	15.0	92.0	105.0	29.0	2.0
[total]	37.0	178.0	248.0	153.0	44.0
Q60_ueberzeugende_Argumentation					
wirkt Ã¼berzeugend in Bezug auf die Argumentation.	6.0	28.0	117.0	114.0	34.0
wirkt nicht Ã¼berzeugend in Bezug auf die Argumentation.	8.0	84.0	68.0	11.0	3.0
[total]	14.0	112.0	185.0	125.0	37.0
Q61_hat_gefallen					
hat gefallen.	6.0	24.0	142.0	139.0	39.0
hat nicht gefallen.	30.0	144.0	97.0	11.0	4.0
[total]	36.0	168.0	239.0	150.0	43.0

Model output 1: Naïve Bayes classification output for interpreting a simple prediction model.

Decision Tree Classifier (J48)

Multidimensional relations and interactions can be (up to a certain point) investigated with the output of decision tree models[128]. In WEKA, decision trees are usually reported in textual format but can also be visualized with the built-in decision tree visualizer or other visualization plugins for decision trees.

In Model output 2, we see the textual decision tree output for the decision tree model.

```
J48 pruned tree
------------------

Q61_hat_gefallen = hat gefallen.
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich klaren Aufbau.
|   |   Q60_ueberzeugende_Argumentation = wirkt überzeugend in Bezug auf die
        Argumentation.: 4 (231.14/122.47)
|   |   Q60_ueberzeugende_Argumentation = wirkt nicht überzeugend in Bezug auf die
        Argumentation.: 3 (22.17/8.54)
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat keinen inhaltlich klaren Aufbau.:
    3 (5.0/1.0)
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich teilweise klaren
    Aufbau : 3 (97.72/39.86)
Q61_hat_gefallen = hat nicht gefallen.
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich klaren Aufbau.
|   |   annotator = Annotator A
|   |   |   Q60_ueberzeugende_Argumentation = wirkt überzeugend in Bezug auf die
            Argumentation.: 3 (19.61/6.3)
|   |   |   Q60_ueberzeugende_Argumentation = wirkt nicht überzeugend in Bezug auf
            die Argumentation.: 2 (43.84/15.69)
|   |   annotator = Annotator C
|   |   |   Q48_koherent = Eher ja: 2 (7.0/3.0)
|   |   |   Q48_koherent = Ja: 3 (4.0)
|   |   |   Q48 koherent = Nein: 2 (1.0)
|   |   annotator = Annotator B: 3 (29.61/12.16)
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat keinen inhaltlich klaren Aufbau.
|   |   annotator = Annotator A: 2 (17.0/1.0)
|   |   annotator = Annotator C: 1-insufficient (7.0/4.0)
|   |   annotator = Annotator B: 1-insufficient (20.15/6.0)
```

[128] See section 3.5.3 for description.

```
|   Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich teilweise klaren
    Aufbau
|   |   annotator = Annotator A
|   |   |   Q60_ueberzeugende_Argumentation = wirkt überzeugend in Bezug auf die
            Argumentation.: 3 (8.35/2.35)
|   |   |   Q60_ueberzeugende_Argumentation = wirkt nicht überzeugend in Bezug auf
            die Argumentation.: 2 (19.09)
|   |   annotator = Annotator C: 2 (48.45/20.0)
|   |   annotator = Annotator B
|   |   |   Q48_koherent = Eher ja: 3 (36.19/12.49)
|   |   |   Q48_koherent = Ja: 2 (6.09/4.0)
|   |   |   Q48_koherent = Nein: 2 (22.59/7.45)

Number of Leaves: 19

Size of the tree: 30
```

Model output 2: Classification output for J48 decision tree classifier in WEKA Explorer.

The tree has a maximum depth of 4, meaning that at the utmost it takes the combined values of four different variables to decide on a grade level. The minimum depth is 2 and can be seen e.g. for texts that did appeal (Q61_hat_gefallen) and had no clear or just a partially clear content structure (Q51_inhaltlich_klarer_Aufbau). In both cases the tree would predict an average grade level of *'3'*.

The numbers in the parenthesis behind the leaf nodes (i.e. the end of the tree that resolves into a decision for one grade) show how many entries have been classified as such and how many of the classified entries have been classified incorrectly[129].

In this report we can also observe systematic biases of the classifier, e.g. the decision tree will never assign the *grade '5-excellent'* and '*1-insufficient'* will only ever be considered for texts not graded by annotator A. This is due to the simplistic model that does not allow to learn any more detailed predictions from the data.

However, building a better performing decision tree classification model, e.g. with the '*relief10'* feature set of experiment 10.2.2.4 that had an overall prediction accuracy of 56.1% also the complexity of the tree might increase. The decision tree obtained for the J48 model of experiment 10.2.2.4 is substantially more complex. Instead of a size of 30 with 19 leaf nodes for the simplified task it has a size of 93 and 68 leaf nodes. Moreover, the depth for the tree with ten features is seven, meaning that seven variables need to be combined in order to come to a decision. This depth, the high number of branches and leaf nodes, and the high variance in the leaf groups makes the interpretation of the bigger tree difficult for humans and further interpretation techniques need to be applied that might help to investigate the tree not only textually but also visually or even interactively.

The WEKA plugin Prefuse tree for example allows to investigate the tree interactively following the branches that one is interested in. Figure 33 shows an example branch investigation with Prefuse.

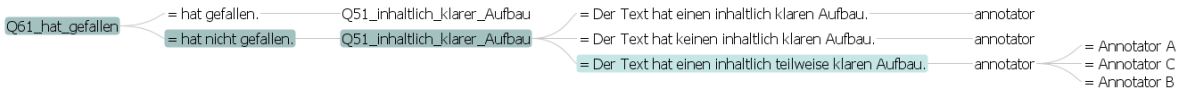


Figure 33: Example of an interactive exploration of a trained decision tree model with the Prefuse plugin for the WEKA Explorer interface.

[129] The floating-point numbers in this particular decision tree result from the fact that '*not determinable'* values have been treated as missing values which in turn have been distributed according to their probabilities across the dataset.

The most important variable for the tree to decide the holistic grade is whether the text appealed to the annotator or not (Q61_hat_gefallen). If it did not appeal it then depends on whether the content had a clear structure or not (Q51_inhaltlich_klarer_Aufbau). However, no matter which evaluation was given for the content structure it would still depend on the annotator variable, which grade the classifier would predict for a given essay. In this example we can already see that the holistic grades are biased by the annotator.

Rule Induction (PART)

The rule learner PART gives a similar output to the decision tree output. Instead of the binary tree representation it creates additive rules (see Model output 3). I use the same simplified example from before to generate a set of 20 rules that the classifier uses to predict the holistic grade. The values in the parenthesis after each rule report again the number of instances classified with this rule and the number of incorrectly classified instances.

```
PART decision list
------------------

Q61 hat gefallen = hat nicht gefallen. AND Q51 inhaltlich klarer Aufbau = Der Text
hat keinen inhaltlich klaren Aufbau. AND annotator = Annotator B: 1-insufficient (
20.15/6.0)

Q61_hat_gefallen = hat nicht gefallen. AND annotator = Annotator A AND Q48_koheren
t = Nein: 2 (33.0/2.0)

Q61_hat_gefallen = hat nicht gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text
hat einen inhaltlich klaren Aufbau. AND annotator = Annotator A AND Q60_ueberzeuge
nde_Argumentation = wirkt nicht überzeugend in Bezug auf die Argumentation.: 2 (36
.66/15.67)

Q61_hat_gefallen = hat nicht gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text
hat einen inhaltlich klaren Aufbau. AND annotator = Annotator A: 3 (17.87/5.54)

Q61_hat_gefallen = hat nicht gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text
hat einen inhaltlich klaren Aufbau. AND Q48_koherent = Eher ja AND annotator = Ann
otator B: 3 (17.53/6.53)

Q61_hat_gefallen = hat nicht gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text
hat einen inhaltlich teilweise klaren Aufbau. AND annotator = Annotator A AND Q60_
ueberzeugende_Argumentation = wirkt nicht überzeugend in Bezug auf die Argumentati
on.: 2 (7.71)

Q61_hat_gefallen = hat nicht gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text
hat einen inhaltlich teilweise klaren Aufbau. AND annotator = Annotator B AND Q48_
koherent = Eher ja: 3 (35.17/12.36)

Q61_hat_gefallen = hat gefallen. AND Q51_inhaltlich_klarer_Aufbau = Der Text hat e
inen inhaltlich teilweise klaren Aufbau. AND Q48_koherent = Eher ja: 3 (71.03/23.8
9)

Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich klaren Aufbau. AND Q4
8_koherent = Ja AND annotator = Annotator A AND Q60_ueberzeugende_Argumentation =
wirkt überzeugend in Bezug auf die Argumentation.: 4 (76.58/40.63)

Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich teilweise klaren Aufb
au. AND Q61_hat_gefallen = hat nicht gefallen. AND annotator = Annotator C: 2 (48.
19/20.0)

Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich teilweise klaren Aufb
au. AND Q48_koherent = Nein AND annotator = Annotator B: 2 (26.2/9.0)

Q51_inhaltlich_klarer_Aufbau = Der Text hat keinen inhaltlich klaren Aufbau. AND a
nnotator = Annotator A AND Q61_hat_gefallen = hat nicht gefallen.: 2 (6.0/1.0)

Q51_inhaltlich_klarer_Aufbau = Der Text hat einen inhaltlich teilweise klaren Aufb
au.: 3 (37.23/20.17)
```

```
Q48_koherent = Nein: 2 (10.05/5.05)

Q48_koherent = Eher ja AND Q60_ueberzeugende_Argumentation = wirkt nicht überzeuge
nd in Bezug auf die Argumentation. AND annotator = Annotator C AND Q61_hat_gefalle
n = hat gefallen: 3 (6.57/2.57)

annotator = Annotator B AND Q48_koherent = Ja AND Q60_ueberzeugende_Argumentation
= wirkt überzeugend in Bezug auf die Argumentation. AND Q61_hat_gefallen = hat gef
allen: 4 (47.24/28.1)

annotator = Annotator B: 4 (36.35/18.28)

Q61_hat_gefallen = hat nicht gefallen. AND Q48_koherent = Eher ja: 2 (7.61/3.61)

annotator = Annotator C AND Q61_hat_gefallen = hat gefallen: 4 (84.99/42.45)

: 3 (19.88/4.63)

Number of Rules:        20
```

Model output 3: The PART rule induction model with the learned rules.

The interpretability of PART classification models changes dramatically when the number of features increases. The classification model for PART for the *'relief10'* feature set of experiment 10.2.2.4 reports 59 rules that go up to a depth of 8 combined predictor variables.

Logistic Regression in WEKA

```
=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                                                                                          Class
Variable                                                                          1-insufficient            2            3            4
=======================================================================================================================================
annotator=Annotator A                                                                   -37.6087       1.1308       0.9039       0.5927
annotator=Annotator C                                                                    19.9653        0.046      -0.3061      -0.0341
annotator=Annotator B                                                                    18.5306       -1.191      -0.6157      -0.5672
Q48_koherent=Eher ja                                                                      -0.047       0.7055       0.9678       4.8992
Q48_koherent=Ja                                                                          -0.6246      -1.5854      -0.8808       3.5425
Q48_koherent=Nein                                                                         1.2445       1.6109      -0.1859     -15.7536
Q51_inhaltlich_klarer_Aufbau=Der Text hat einen inhaltlich klaren Aufbau.                -1.7127      -1.0919      -0.8747       1.2907
Q51_inhaltlich_klarer_Aufbau=Der Text hat keinen inhaltlich klaren Aufbau.                0.2463       -1.977      -2.9361     -14.7289
Q51_inhaltlich_klarer_Aufbau=Der Text hat einen inhaltlich teilweise klaren Aufbau.       1.7431       1.7541       1.8126       3.0677
Q60_ueberzeugende_Argumentation=wirkt nicht Ã¼berzeugend in Bezug auf die Argumentation.  0.1151       1.0351        1.183       0.3147
Q61_hat_gefallen=hat nicht gefallen.                                                      3.1653       3.0928       1.4333       -0.158
Intercept                                                                               -19.7049       1.0597       2.2638      -3.8404


Odds Ratios...
                                                                                          Class
Variable                                                                          1-insufficient            2            3            4
=======================================================================================================================================
annotator=Annotator A                                                                          0        3.098       2.4691       1.8089
annotator=Annotator C                                                             468603201.2521        1.047       0.7363       0.9664
annotator=Annotator B                                                             111613423.8803       0.3039       0.5403       0.5671
Q48_koherent=Eher ja                                                                      0.9541       2.0248       2.6322     134.1787
Q48_koherent=Ja                                                                           0.5355       0.2049       0.4145      34.5549
Q48_koherent=Nein                                                                         3.4713       5.0072       0.8303            0
Q51_inhaltlich_klarer_Aufbau=Der Text hat einen inhaltlich klaren Aufbau.                 0.1804       0.3356        0.417       3.6352
Q51_inhaltlich_klarer_Aufbau=Der Text hat keinen inhaltlich klaren Aufbau.                1.2793       0.1385       0.0531            0
Q51_inhaltlich_klarer_Aufbau=Der Text hat einen inhaltlich teilweise klaren Aufbau.       5.7149       5.7784       6.1263      21.4918
Q60_ueberzeugende_Argumentation=wirkt nicht Ã¼berzeugend in Bezug auf die Argumentation.   1.122       2.8153        3.264       1.3699
Q61_hat_gefallen=hat nicht gefallen.                                                     23.6953      22.0397       4.1923       0.8538
```

Model output 4: Regression coefficients and odds ratios reported by logistic regression classifier of WEKA.

Similar to the NaiveBayes classifier output, the logistic regression output shows coefficients and odds ratios (Model output 4) for each grade level, for all the variables and all their categories in the model. Although logistic regression models are in general able to take variable interactions into account, the WEKA default implementation does not account for interactions and therefore cannot be used for multidimensional analysis.

The coefficients show how the binarized log of the variable category changes when the grade level changes from '*5-excellent*' to the respective other class. Because of the log values the numbers

reported for the variable categories are not immediately interpretable to humans. More concrete are the odds ratios reported below. The odds ratio shows the non-log-transformed (exponential) value of the regression coefficient and can be read as the conditional probability (odds) of the outcome when the input variable in question is increased by one unit.

However, both are only meaningful when compared to other values of the table. Considering that the default value is one of the classes (per default the alphabetically first class), the interpretation of the odds ratios is still far from intuitive, especially when many, categorical factors are involved. Visualization techniques for variable effects in logistic regression models as for instance marginal effects plots or partial dependency plots are currently not provided by WEKA.

Support vector machine classification in WEKA (SMO)

Lastly, it is also possible to inspect the output of the SMO support vector machine classifier in WEKA. However, SVM multi-class classifiers usually calculate the prediction using a workaround by building a binary classifier for each possible combination of two class labels and then choosing the one with the highest weight. This makes the interpretation of the classifier outputs more difficult as weights for all possible combinations must be considered and compared.

Summary

A simple comparison of classification performance for subsets of the text analysis feature set showed that content-related features have the best prediction performance, and therefore give the most information to the models in order to predict the grade level of the texts. Features regarding the completeness of the genre-relevant text parts also predicted the holistic grades with results that were significantly above a majority baseline. The features regarding the text structure, however, did not yield a predictive performance that is high enough to conclude that the algorithm can learn the holistic grade from them.

Using different measures and algorithms for feature ranking, a list of seven features was identified as among the top ten features in all of the ranking methods. These were

- Q48_koherent (coherence)
- Q49_Gesamtthema_nachvollziehbar (comprehensible topic)
- Q50_konzeptionell_zusammenh (conceptual coherence)
- Q51_inhaltlich_klarer_Aufbau (clear content structure)
- Q58_Roter_Faden (golden thread)
- Q61_hat_gefallen (text appeal)
- Q62_inhaltlich_nachvollziehbar_klar (comprehensible and clear content)

However, except for logistic regression and for the rule induction algorithm PART (the two algorithms that suffer most when redundant noisy data is used, classification results drop significantly for classifiers trained with these features. The random forest classifier with the full feature set performed significantly better than this feature set. Also the results for a feature set only including the five best features (annotator, coherence, clear content structure, convincing argumentation and whether the text appealed to the reader) selected with ReliefF performed better, indicating that the above most probably intercorrelated features probably only capture one reduced aspect of text quality while other aspects are transported by features are less important according to monofactorial analysis.

Interpreting the model internals for this set of five features (according to the ReliefF algorithm) we can observe the following. The probability of predicting '*4'* or '*5-excellent'* is, in the Naïve Bayes classifier at least ten times higher, when the text has a convincing argumentation, a clear content

structure, appealed to the reader and was not incoherent. While the probability of predicting '*3*' was still twice as high, when the argumentation was convincing compared to when the argumentation is not convincing, the difference between appealing and not appealing texts as well as the difference between only partially clear content structure and a clear content structure is not so big in this category, but still tendentially higher. The other variable levels are instead a sign for lower grade levels.

The decision tree algorithm splits the dataset first on the variable for text appeal and then on the variable for clear content structure. While for appealing and clearly structured texts the persuasiveness of the argumentation is responsible to decide between grade level '*3'* and '*4'*, texts without clear structure don't achieve the grade level '*4'*. For not appealing texts, the content structure as well as the annotator decide whether grade level '*1-insufficient*', '*2'* or '*3'* is assigned. Grade *'5-excellent'* is never assigned by the model.

Most of the rules in the rule induction model (PART) also decide based on a combination of the text appeal and the content structure of the text. While many of the rules also make use of the annotator, the persuasiveness of the argumentation is less present in the rules. As the previous model this model never assigns '*5-excellent'* grades.

While the odds ratio for predicting lower grades ('*1-insufficient'* or '*2'*) in the logistic regression model is much higher for texts that did not appeal to the reader or were rated as incoherent, the differences between the lower levels and the higher ones were not that high for other categories. However, the odds ratio for predicting a better than average grade (*'4'*) was much higher than for the other grade levels for coherent or rather coherent texts and for clearly structured or at least partially clearly structured content.

The SMO, MLP and random forest model internals were not straightforwardly interpretable through the WEKA model output. More generally, the confusion matrix showed that grade levels that are close to each other are, as expected for ordinal data, more likely to be confused. The bias introduced by the raters was visible in each of the models, although it did not show up in the feature ranks of the ten most important features according to the (mainly monofactorial) feature selection methods (e.g. on correlation, information gain or gain ratio). However, it was a main criterium for deciding grades in the predictive models.

10.2.4 Holistic grade as a numerical (ordinal) value – Correlation analysis and mixed-effects regression modelling

Another way to identify (linear) relations and structure in the data is to calculate correlations and use linear regression modelling on numerical data. As many of the variables in feature set 1 were inherently ordinal, as can also be seen in the interpretation attempts of the previous section, it is therefore reasonable to transform the data into numerical values using monofactorial correlation and multifactorial linear regression models for analysis and interpretation.

10.2.4.1 Correlation analysis

As a first approach to the numerical data, I calculated Spearman rank correlation for the inherently ranked questionnaire items of feature set 1. Figure 34 shows questionnaire items that are significantly related with the holistic grades (p-value < 0.01) and show a moderate effect size of at least 0.2 rho.

The variable with the highest correlation with the holistic grade is the rather subjective evaluation if the text appealed to the annotator or not. The next six highest correlated variables all address matters of coherence and topic development (i.e. Q48_koherent, Q62_inhaltlich_nachvollziehbar_klar, Q49_Gesamtthema_nachvollziehbar, Q51_inhaltlich_klarer_Aufbau, Q58_Roter_Faden,

Q50_konzeptionell_zusammenh). Apart from the highly subjective question on the appeal of the text, also the rather subjective questions on whether the text was interesting (Q63_interessant_langweilig) or entertaining (Q64_unterhaltsam) showed moderate correlations with the holistic grade. Other still moderately correlated variables address the structuring of the text, its argumentative power, if the text contains a clear opinion statement of the writer or not and if it fulfils the expectations of genre and the announced topic and the task.

However, the observed differences between the three different raters in terms of the variance of grade labels in section 9.1, encourage to explore the correlations between text evaluations and holistic grades for each annotator individually.

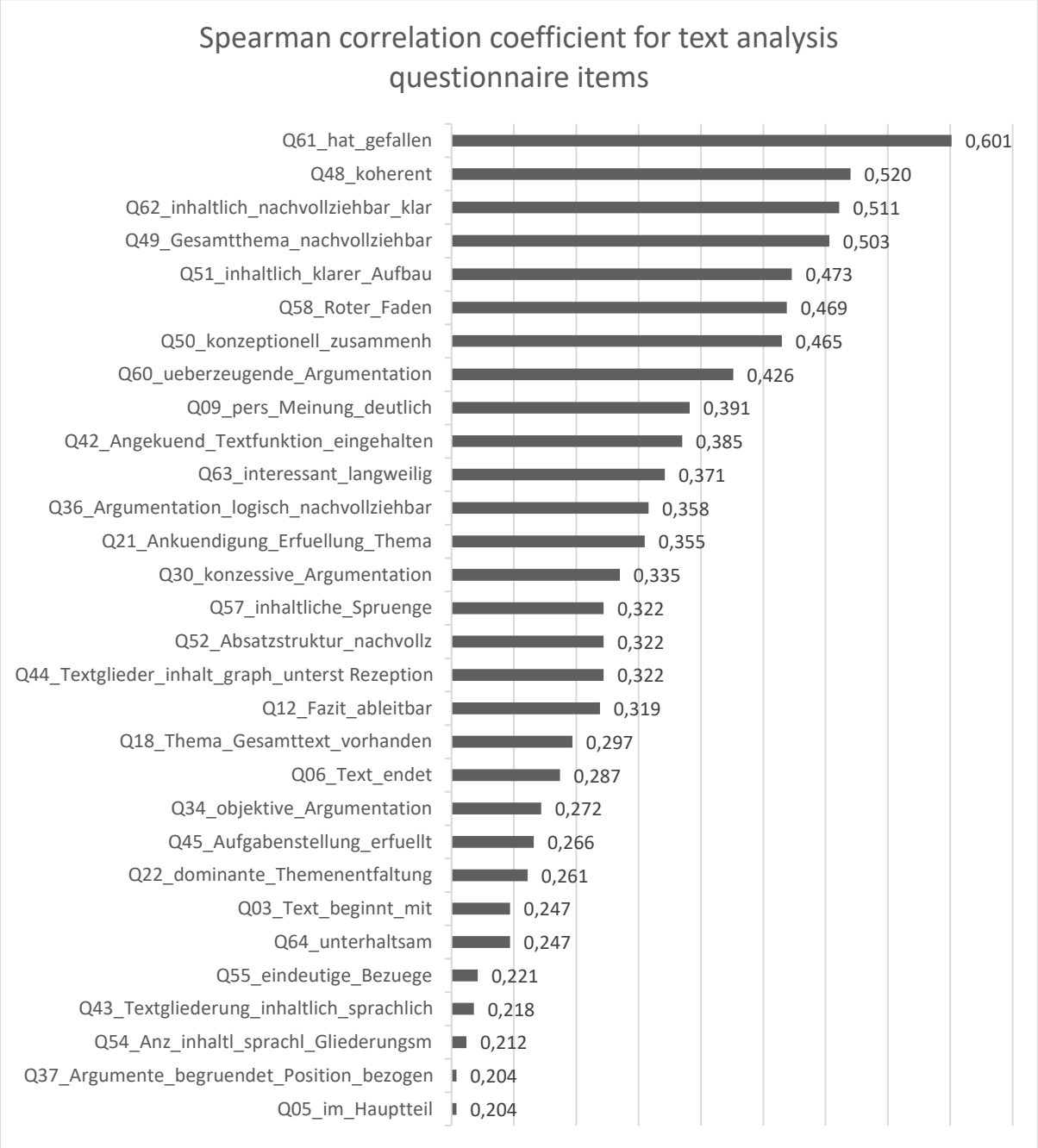


Figure 34: Rho coefficients for questionnaire items that are significantly (p-value < 0.01) correlated with the holistic grades for all items with rho > 0.2.

The heatmaps below show the correlations reported for each annotator in comparison. We see that Annotator C has much weaker relations between his/her text evaluations and the holistic grade he or she assigned for the text. Annotator B has for many variables the strongest correlation between the evaluation and the holistic grade. This is particularly obvious with the analytic evaluations he or she gave for the evaluations regarding the clarity of the text structure, the existence of a conclusion at the end of the text and for the questions regarding the clarity of the argumentation. Annotator A differs from the other annotators as he or she had stronger correlations for rather subjective evaluations (e.g. whether he or she liked the text, the interestingness of the text and whether he or she evaluated the text as entertaining). Some of the variables were not informative when modelling all texts together but have high correlations when investigated on just one annotator.

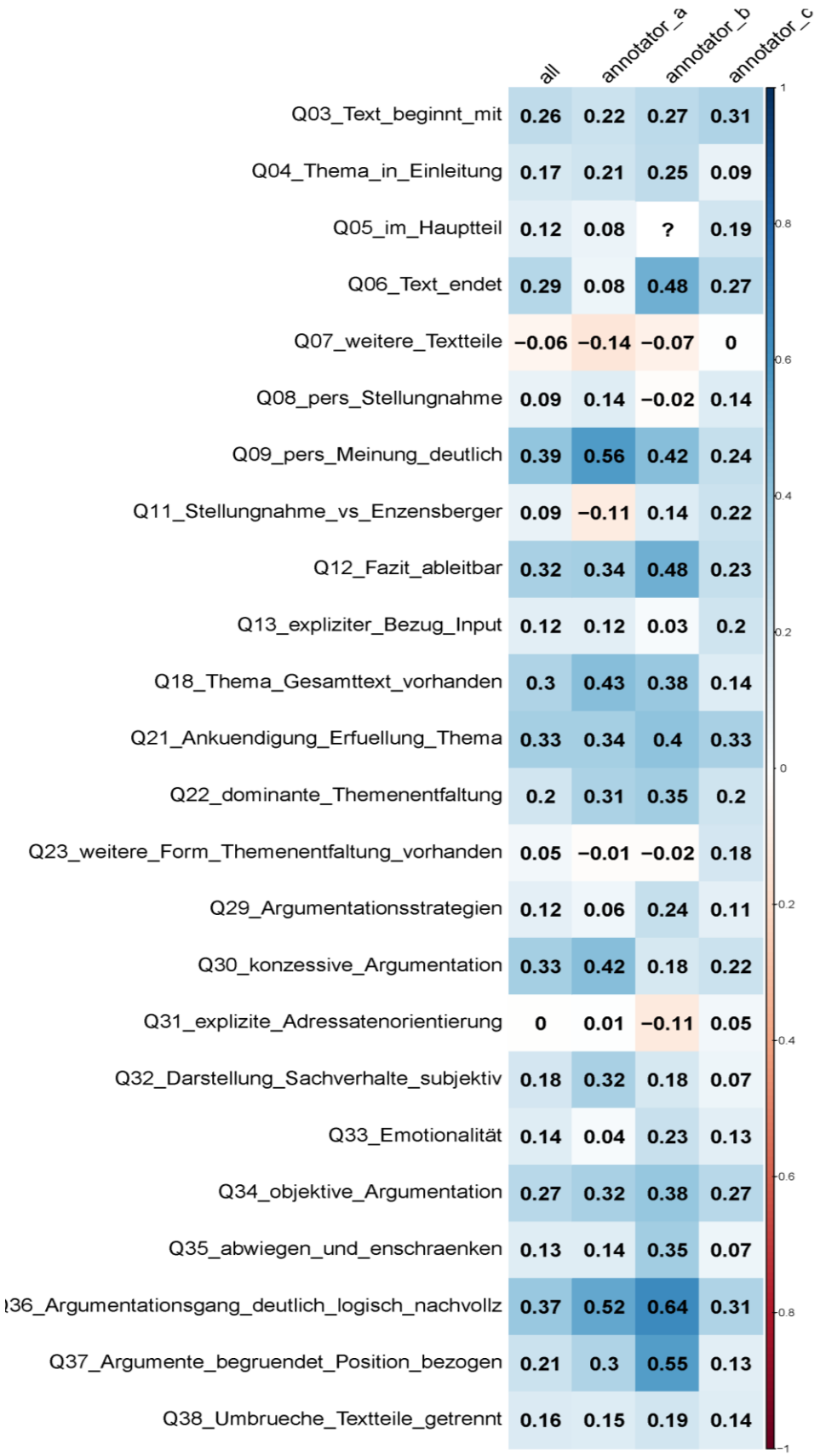


Figure 35: Heatmap showing correlation between individual annotators and questionnaire items (part 1).

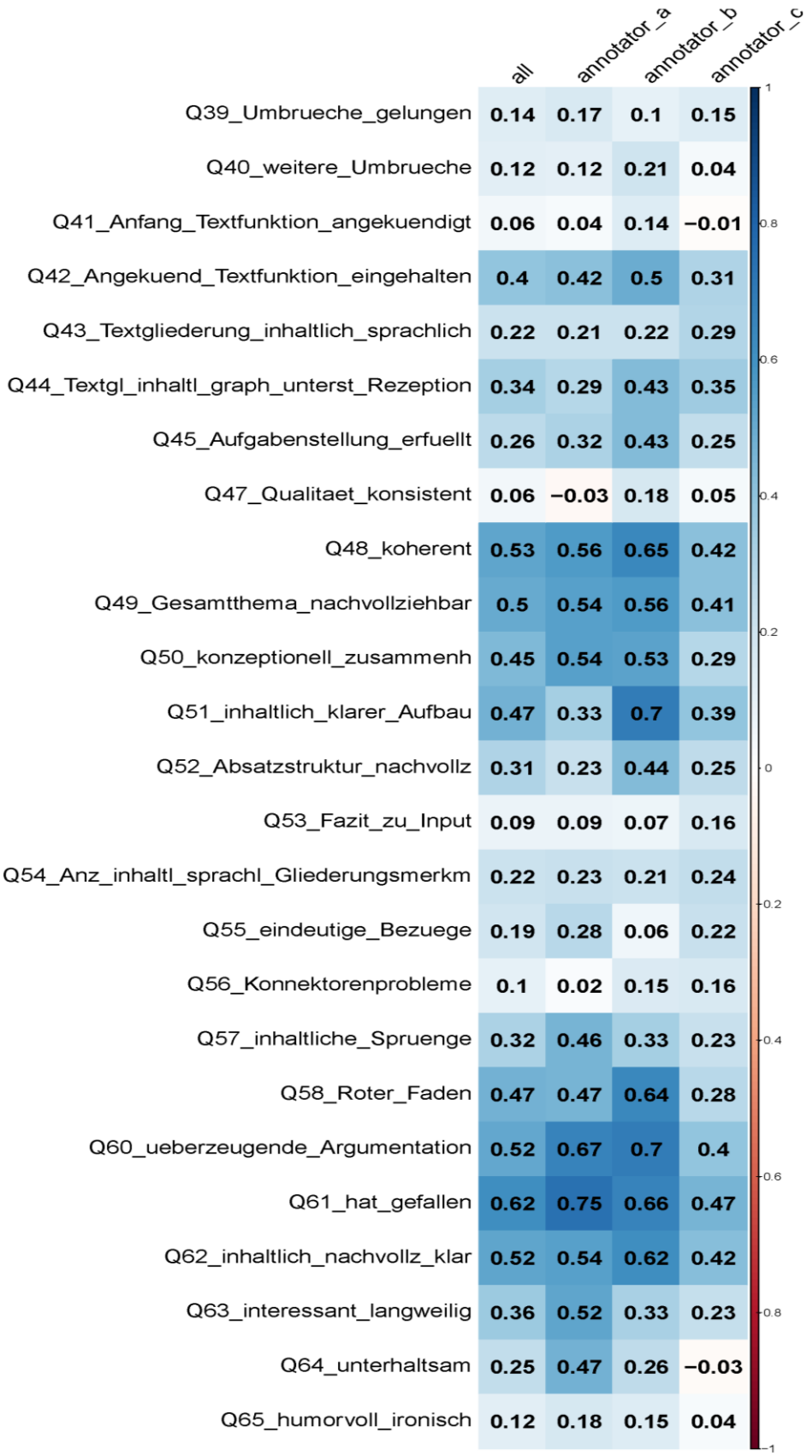


Figure 36: Heatmap showing correlation between individual annotators and questionnaire items (part 2).

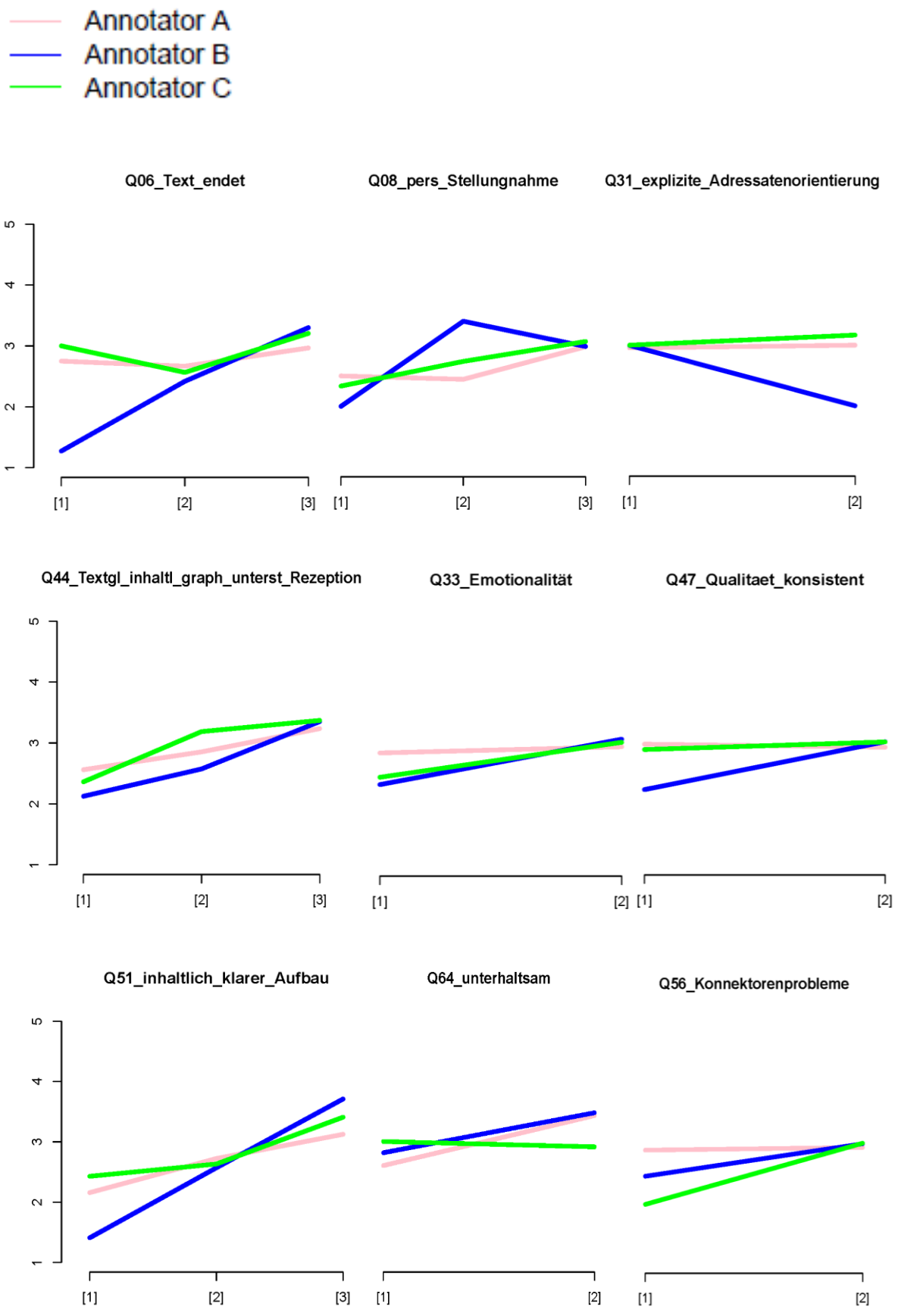


Figure 37: Interaction plots for Rater effects.

Figure 37 shows interaction plots for some of the variables where annotators visibly assign different grades. The plots show mean values for assigned grades for each annotator and each level of the predictor. We can see that the behaviour of Annotator B is different from the others when it comes to texts without proper ending, texts with explicit address to the reader, texts where the quality is inconsistent and texts without clear content structure. In all those cases, Annotator B gave, in average, lower grades. Annotator C on the other hand was usually rather moderate. However, he or she gave lower grades to texts showing problems in the use of connectives but did not give better grades for entertaining or funny texts (contrary to the other annotators). Annotator A seems to be the least influenced by the illustrated text analysis factors. His or her ratings for the emotional or unemotional text, texts with or without connectives, consistent or inconsistent quality and text with or without an explicit address of the reader did not differ.

10.2.4.2 Other measures for bivariate (monofactorial) analysis.

Alternatively, also other information theoretic association measures, such as information gain, (pointwise) mutual information or gain ratio are frequently used in the context of machine learning to estimate how much information is provided by one variable, when predicting another variable, i.e. how important the predictor variable is for explaining the dependent variable.

These measures are often used as part of the training procedure within learning algorithms or also for preliminary feature selection before the actual training step. However, they can also be used to investigate variable importance in general. I used another approach to investigate the importance of individual analytic evaluations for explaining the grades. In this experiment I rank all the variables using the information gain metric. The information gain metric describes the reduction of entropy (i.e. impurity) of the dependent variable when using the independent variable for partitioning. Contrary to the previously used correlation coefficient, we cannot see the direction of the relation. Whereas the correlation coefficient provided a clear positive or negative sign that indicates the type of (symmetric) relation, most feature selection metrics like information gain show how much this variable contributes to explaining the dependent variable but are less directly interpretable in terms of the quantitative effect of the variable. Table 32 shows the ranking and the respective information gain of the 15 top ranked questionnaire items coded as ordinal variables.

Rank	Text evaluation	Information gain
1	Q61_hat_gefallen	0.32428
2	Q48_koherent	0.29786
3	Q62_inhaltlich_nachvollziehbar_klar	0.24380
4	Q51_inhaltlich_klarer_Aufbau	0.22693
5	Q49_Gesamtthema_nachvollziehbar	0.22589
6	Q60_ueberzeugende_Argumentation	0.22000
7	Q50_konzeptionell_zusammenh	0.20323
8	Q58_Roter_Faden	0.20307
9	Q63_interessant_langweilig	0.15259
10	Q22_dominante_Themenentfaltung	0.14857
11	Q21_Ankuendigung_Erfuellung_Thema	0.14749
12	Q09_pers_Meinung_deutlich	0.14206
13	Q18_Thema_Gesamttext_vorhanden	0.13882
14	Q44_Textgliederung_inhaltlich_graphisch_unterst_Rezeption	0.12166
15	Q12_Fazit_ableitbar	0.11339

Table 32: Feature ranking using information gain (top 15 features).

Although we are missing the information on strength and direction of correlation, the information gain rank for the text analysis questionnaire items shows a strong overlap with the rank created using the correlation coefficient from the Spearman correlations.

10.2.4.3 Mixed-effects regression modelling

Next, I build and compare linear regression models using the variables of feature set 1 that showed a moderate correlation (rho > 0.2) in the previous correlation analysis.

When using all text analysis questionnaire forms (including those that referred to the same text but were made by a different annotator) in order to predict the holistic grades assigned by the annotator, I introduce hierarchy in the observations. Hierarchy in the observations violates the assumption of independent data points that is needed for methods like linear or logistic regression. In the following experiments I therefore use a mixed-effects model architecture (sometimes also called multi-level model, hierarchical linear model or mixed model), instead of regular linear regression, to account for the hierarchical data as well as for a possibly significant random effect introduced by the raters' low inter-rater-reliability.

After automatic stepwise model selection using the `lmerTest` package of R, the following predictors remained in the final model of the analysis[130].

Q61_hat_gefallen (Appeal)
Q22_dominante_Themenentfaltung (Genre)
Q63_interessant_langweilig (Interestingness)
Q48_koherent (Coherence)
Q52_Absatzstruktur_nachvollz (Paragraph structure)
Q09_per_Meinung_deutlich (Opinionatedness)
Q21_Ankuendigung_Erfuellung_Thema (Topic announcement)
- (1 | annotator)

In the manual stepwise model selection using forward selection we reach the same final model as with the automatic approach. The predictors explain 49,3% of variance in the grade level (R^2 for the main, i.e. fixed, effects in the mixed-effects model).

At last, I tried fitting a model without random effects structure for the annotators. An ANOVA comparing this fixed-effects model (not accounting for a possible annotator bias) with the previous mixed effects model showed that the more complex mixed-effects model (accounting for annotator differences) performed significantly better than the model without. Besides, it deals with the problem of repeated measures drawn from the annotators and therefore the violation of independence of data points and should be preferred for these reasons.

The built model can then be investigated, inspecting the effect of each variable on the predictions of the model. Through this we can observe the type, direction and effect size of the relations in the data. The variable effects plots for the variables in the built model show relatively small but positive effects for all fixed effects variables (see Figure 38).

[130] The model was significant with $p < 0.001$.

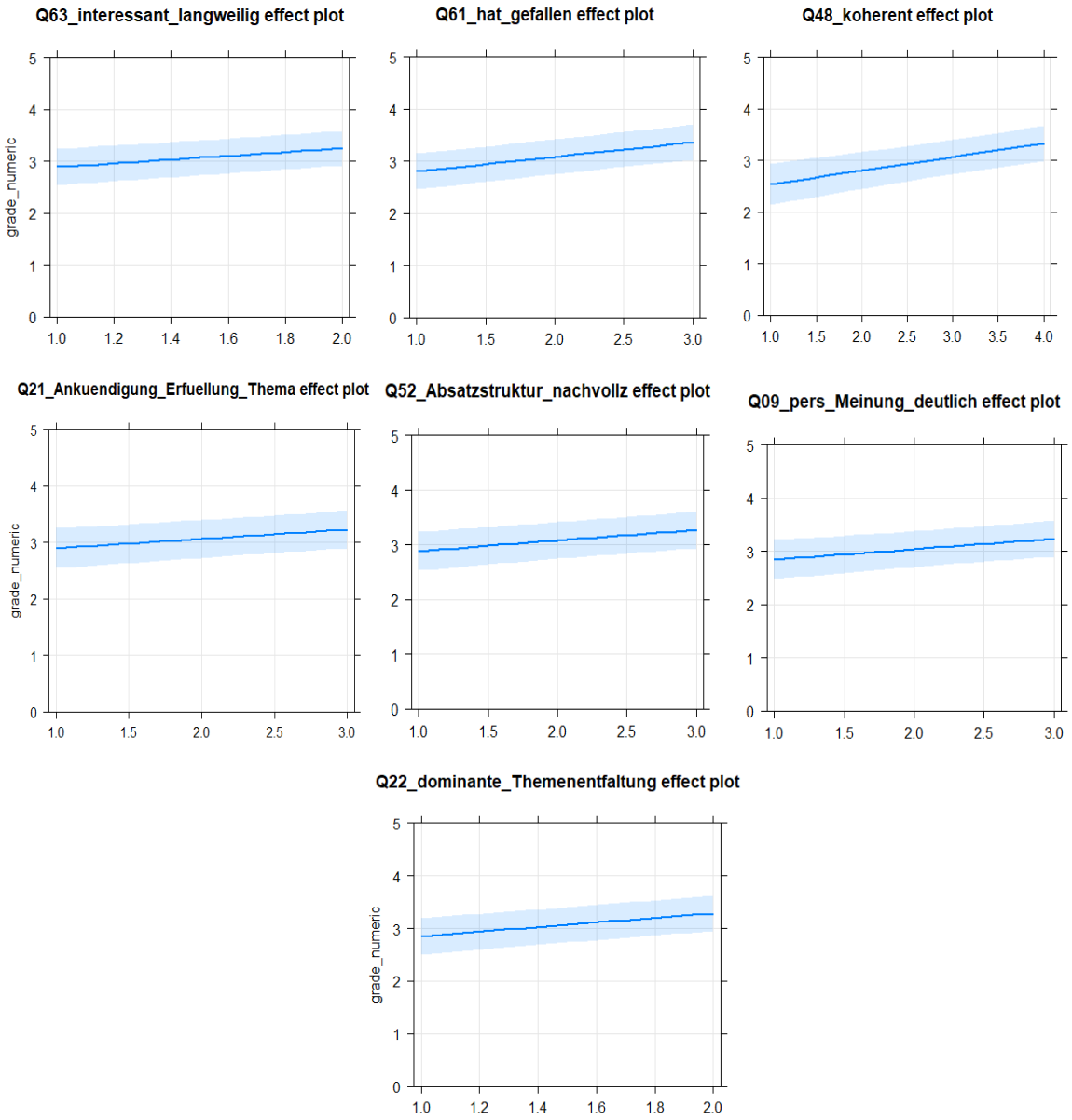


Figure 38: Effects plots for significant main effects in the mixed-effects regression model for predicting the holistic grade.

However, the plots calculate an average effect, when visualising the effect of the annotators in different intercepts, for example for the effect of the feature Q48_koherent, we can see that the intercepts for Annotator B is much higher than for the other two annotators (see Figure 39)[131].

[131] The same is true for the other variable effects, which are not printed here for simplicity.

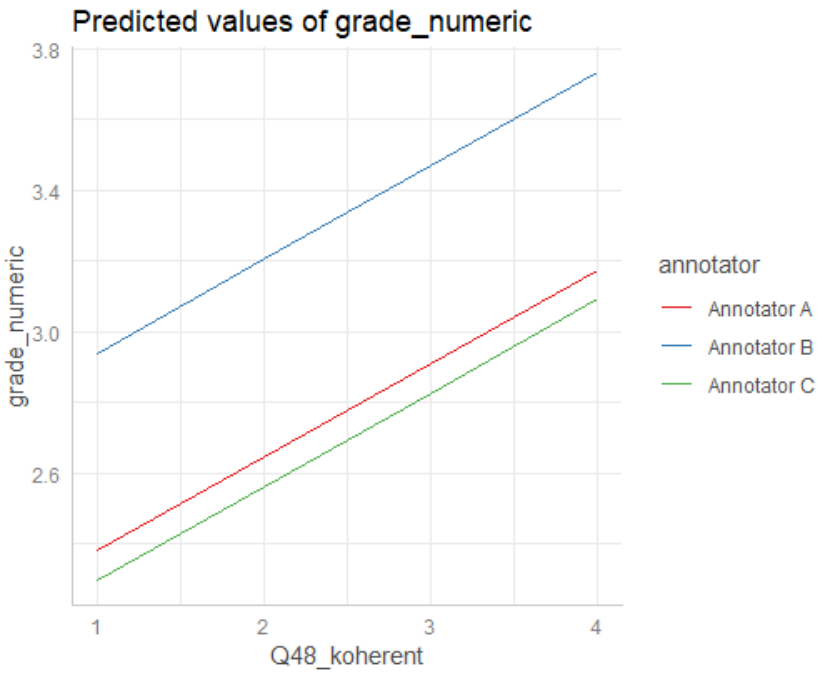


Figure 39: Effects plot visualizing the different intercepts for the effect of text coherence per annotator (random effect).

10.2.4.4 Interactions and Multicollinearity

The correlation matrix schematically presented in Figure 40 shows that some of the variables have weak to moderate correlations with each other. This is not very surprising, considering the conceptual relatedness of the variables, but it might also have an influence on the choice of analysis methods. Linear regression models for example, but also other learning algorithms do not deal well with collinearity in the data. In the dataset the variables with the highest correlations between each other were `Q21_Ankuendingung_Erfüllung_Thema` (the evaluation whether the topic is mentioned in the introduction) and `Q04_Thema_in_Einleitung` (the evaluation whether the announced topic is adhered to) with a correlation coefficient rho of 0.72; `Q22_dominante_Themenentfaltung` (whether the text genre was argumentative or not) and `Q45_Aufgabenstellung_erfuellt` (whether the task was fulfilled or not) with a correlation coefficient of rho = 0.77; and `Q43_Textgliederung_inh_sprachl` (whether there were textual structuring signals or not) and `Q54_Anzahl_inh_sprachl_Gliederungsm` (how many textual structuring signals were present in the text) with rho = 0.8.

Furthermore, the variance inflation factor for the regression model, provided by the `vif` function of the car package for R shows some multicollinearity especially for the factors `Q61_hat_gefallen` and `Q48_koherent`, meaning that the respective effect of the two different factors cannot be kept apart. However, the correlations are not very high, and the factors are in general also conceptually expected to be related, which is why the interpretation is not further hampered.

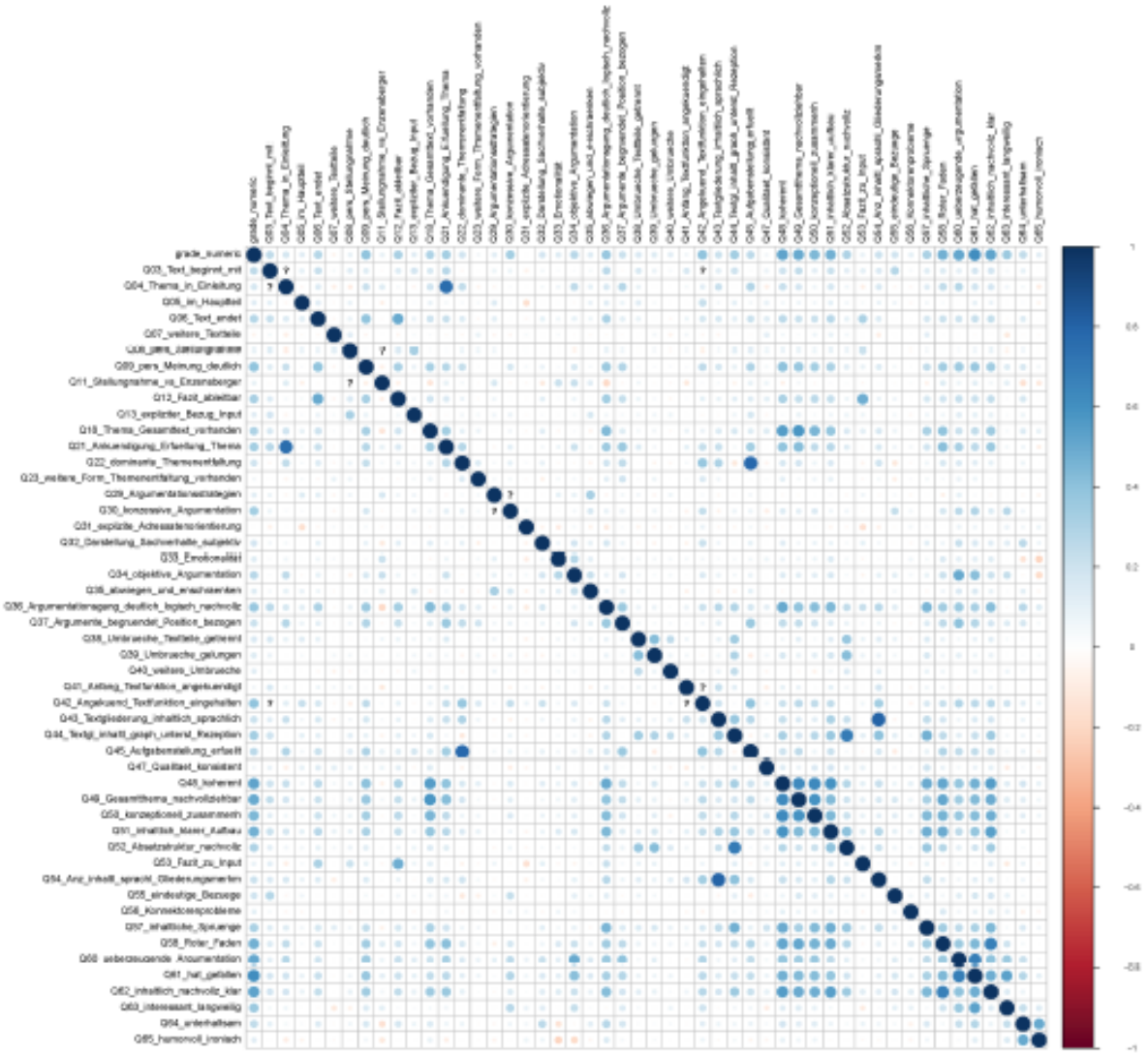


Figure 40: Plot for correlation matrix for questionnaire items showing mutually correlated items.

Summary

In this final section on the text analysis questionnaire, grade levels as well as the originally categorical items have been reformulated to numerical (ordinal) values and monofactorial and multifactorial methods have been used to investigate the relations in the data. The analysis showed moderate to high correlations between the holistic grades and many items of the text analysis questionnaire (30 out of 50 features had a Spearman correlation coefficient above 0.2). The information gain metric for the feature list showed a very similarly ranking of features as the correlation measure. However, in the multivariate mixed-effects regression model, only the features Q61_hat_gefallen, Q22_dominante_Themenentfaltung, Q63_interessant_langweilig, Q48_koherent, Q52_Absatzstruktur_nachvollz, Q09_per_Meinung_deutlich and Q21_Ankuendigung_Erfuellung_Thema remained as significant main effects in the model, while other highly correlated features did not contribute significantly (probably due to multicollinearity in the data). Furthermore, the conducted experiments also showed the existence of a rater bias in the data. Not only could we observe a significant bivariate association between annotator and text quality tested with chi-squared test of

independence (p-value < 0.001). The annotator is also one of the most important features for the prediction of text quality on this dataset and regression models accounting for the annotator as a random effect perform significantly better.

Conclusion

The experiments above showed monofactorial as well as multifactorial methods for analysing relationships between the items of a text analysis questionnaire and the holistic grade labels of the argumentative student essays in the KoKo corpus. In terms of multifactorial methods using predictive modelling and machine learning, both classification and regression models have been built using once the original dataset and once a dataset where inherently ordinally structured variables have been transformed to numerical values. Classification and regression experiments showed results that clearly differed from the monofactorial analyses, in terms of which variables contributed most to explaining the variance in the grade levels and led to the best prediction results. While monofactorial methods ranked conceptually similar multicollinear variables (e.g. Q58_Roter_Faden and Q62_inhaltlich_nachvollziehbar_klar) similarly high, multifactorial methods showed that the information in those variables is probably redundant and can be complemented better with other variables such as Q09_pers_Meinung_deutlich. Furthermore, the experiments showed a significant rater effect, indicating that the holistic grades are highly influenced by the annotators. Further explorations of this rater effect made it possible to identify strategies and tendencies of individual raters (e.g. that annotator A put more emphasis on rather subjective holistic evaluations of text appeal and interestingness).

11 Dealing with issues of observational data

Feature set 2: Error annotations

This section discusses some challenges in the analysis that are caused by non-normal distributions, outliers, and complex or noisy datasets. It uses frequency data for error annotations (feature set 2), a frequently used set of features for corpus linguistic studies. It shows constraints and limits of the previous standard analysis methods and hints at resulting problems of interpretation.

11.1 Methodology

After the first investigations on correlates of text quality using the text metadata provided via the text analysis questionnaire items of feature set 1, this section evaluates the use of frequency information, in particular the use of frequency tables for error annotations (feature set 2) for text quality prediction.

The use of frequency data is in its nature more closely related to corpus linguistic research, but also bares difficulties that arise from such types of data. Corpus frequencies are rarely normally distributed and usually show positive skewness (McEnery & Wilson, 2001), making the use of methods that assume normally distributed data inappropriate. Lexical frequencies, in particular, usually follow a Zipfian distribution, which is an instance of the power law distribution and is characterized by a small number of very high frequencies and a large number of very low frequencies (Baroni, 2009; Desagulier, 2017). Frequencies also usually depend on the length of the samples of text under comparison (i.e. the size of the compared subcorpora or the length of the text) and therefore generally need some sort of normalization that takes varying text lengths into account. If we compare texts from different authors (especially when looking at their stylistic features), we also must expect individual variation

(Baker, 2010; Baroni & Evert, 2009). This results in biased corpus comparisons, when writers are over- or underrepresented in a sample.

For these experiments, the used dataset has to be reduced, as not all texts have received all annotations. As shown in section 8, only 349 texts have been graded with a holistic grade and annotated for all four sets of error annotations. The dependent variable of this analysis is again the holistic grade provided as a 5-point scale text quality evaluation. The predictor variables used for the following experiments are described in section 9.2.2 and are based on frequency lists of error types that have been found and annotated in the texts (i.e. numerical predictors). In addition to the individual error types, the feature set contains variables that aggregate the error types on all levels of hierarchy according to the annotation scheme designed by the authors of the corpus.

I first investigate distributional issues in the data, looking for outliers and bias introduced by different text lengths. Then, I perform a correlation analysis using Spearman rank correlation and visualize results in a heatmap to get a first impression of the data and conduct a series of experiments to illustrate difficulties with complex feature sets when doing data-driven analysis on such data. While the classification experiments in section 10 are prevalently done in the WEKA data mining software, the rest of the experiments are conducted using R.

11.1.1 Distributional issues I: Text length and error frequency

As mentioned before, I expect error frequencies to depend on the text length. I therefore normalize the raw error frequencies (after controlling for a possible relationship between text length and number of errors) calculating the relative error frequencies per 1000 words for each text.

11.1.2 Distributional issues II: Outliers and investigation of individual texts

The distribution of error types presented in section 9.2.2 shows a negatively skewed distribution of error frequencies for most error types, i.e. the majority of texts usually has no errors of that kind, while a few texts show a high number of such errors. It seems that writers either have severe problems with one error type that results in various instances of this error type in one text or don't make the error at all. Therefore, I investigate outliers for feature set 2 and subsequently transform the feature set into binary features, indicating for each error type whether it was present in the text or not.

11.1.3 Dealing with complexity in the data

I then approach the question of complexity for data-driven analysis with data of this kind and address the following issues:

(1) Small relations and many features
Problems when dealing with small effect sizes and many features are investigated. I try to reduce noise in the data using feature selection and compare classification performance for the whole, noisy feature set and reduced feature sets (X, Y and Z features).
(2) Binary features vs. frequency features
Classification performance and model fit for both variable representations, the relative error frequencies giving detailed information on the frequency of errors, and the binary representation of error types with reduced data complexity are compared.
(3) Feature aggregation and division
Correlations and prediction results are explored for different granularities of error categorisations by aggregating and splitting error types.

11.2 Results

The frequency and type of occurring errors is a very common variable for the analysis of text quality or competence in student essays. However, such data can also introduce difficulties in the analysis due to non-normally distributed values and complex, hierarchical annotations schemes. As described in section 8, there are 349 texts in the KoKo corpus that provide annotations for orthography, punctuation, grammar and lexis errors, resulting in a feature set of error frequencies with 229 features distributed over various hierarchical levels. The distribution for this subset of the data is illustrated in Figure 41. The data description in section 9.2.2 showed that most error types in this feature set are skewed. Many of them only occur in a small fraction of the texts, but when they occur, they seem to occur more often within the text. In the following, we approach the frequency data with various techniques for data mining, including different strategies to prepare and select the variables and subsequently build, evaluate and interpret a predictive model based on this data.

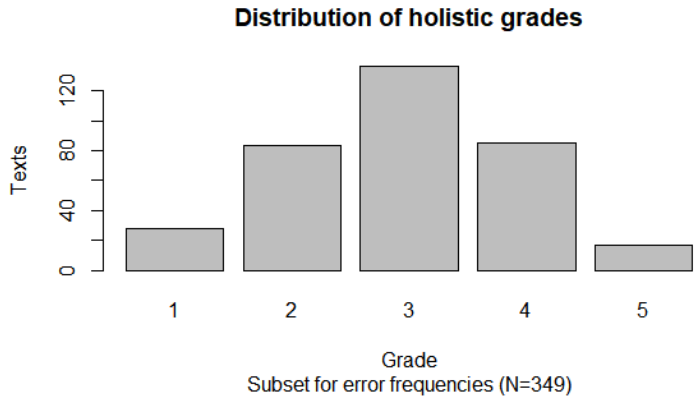


Figure 41: Distribution of holistic grades for subset of 349 fully error annotated texts.

11.2.1 Distributional issues I: Text length and error frequency

In a very first step, I test the relationship between raw error frequencies and the holistic grade. Using Spearman rank correlation in order to account for the ordinal character of the dependent variable we find no significant correlation between the number of errors and the holistic grade. If we inspect the correlation matrix for the variables holistic grade, text length and number of errors in Table 33, we find two moderate correlations: one between the holistic grade and the text length indicating that longer texts have better grades (rho = 0.4, p-value < 0.001) and one between the text length and the total number of errors in the text (rho = 0.47, p-value < 0.001), saying that longer texts also tend to have more errors. However, if one normalizes the error frequencies, e.g. by using relative frequencies per 1000 words (Table 34), thereby removing any influence of the text length variable, the correlation between errors and holistic grade turns around and a higher ratio of errors is now negatively related with the holistic grade (rho = -0.18, p-value < 0.001), which is what we would have expected from the beginning.

	Num errors abs.	*Text length*	*Holistic grade*
Num errors abs.	*1*	-	-
Text length	*0.467*	*1*	-
Holistic grade	*0.093*	*0.399*	*1*

Table 33: Correlation matrix for absolute (raw) error frequencies.

	Num errors rel.	*Text length*	*Holistic grade*
Num errors rel.	*1*	*-*	*-*
Text length	*-0.133*	*1*	*-*
Holistic grade	*-0.181*	*0.399*	*1*

Table 34: Correlation matrix for relative error frequencies (per 1000 words).

As one can expect, the differently distributed text length variable has a moderating effect on the relationship between raw error frequency and holistic grade. Confounding variables like the text length can have substantial influence on the conclusions drawn from the data. In this example the raw error frequencies are almost unrelated to positively related with the grade, the relative error frequencies on the other hand are clearly, if weakly, negatively related. Usually, one controls for any possible confounders before the actual analysis.[132] In corpus linguistics, this is often done a priori, by sampling the data right from the beginning to make it representative for a special purpose. This is usually easier when new data is collected for the specific study, as the collection can be directed in order to balance the data for the confounding variables or hold them constant in general. If the study aims to work with existing corpora, sampling becomes more difficult. First, because any possible confounding variables must be present in the data or metadata and second, because the corpus has to be large enough to allow such sampling, as this usually involves removing a big part of the available data, which is in turn neither helpful for statistical analysis nor for any learning algorithm.

Another possibility is to control the effect of the allegedly confounding variables during the analysis, by including the confounders in the model. While monofactorial analysis does not account for the effect of confounders or interactions, multifactorial analysis like generalized linear models (linear and logistic regression models) but also machine-learning-based models can single out the individual effect of the variable of interest when confounding variables are included as predictor variables in the model.

Regression models, for example, report regression coefficients that allow to see the individual ("partial") effect of the variable, while the other predictor variables (e.g. confounders) are held constant. However, although it is not a strict assumption of regression models that independent variables are completely unrelated, correlations between predictor variables (as is always the case with confounding variables) can influence the regression coefficients up to the point where they are no longer reliable. If we want to interpret the regression coefficients, high multicollinearity among predictors must be avoided and should be controlled for any regression model prior to interpreting the regression coefficients, e.g. by calculating the variance inflation factor (VIF) (cf. section 10.2.4.4).

Below we see the output of a simple linear regression model for the normalized error frequencies.

```
Call:
lm(formula = grade_numeric ~ errors_all)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09773 -0.88026  0.01084  0.87658  2.24617

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.252836   0.102358  31.779  < 2e-16 ***
errors_all  -0.006687   0.001893  -3.532 0.000469 ***
---
```

[132] Making use of domain-knowledge and intuition to identify the possible confounders.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9823 on 347 degrees of freedom
Multiple R-squared:  0.03469,  Adjusted R-squared:  0.03191
F-statistic: 12.47 on 1 and 347 DF, p-value: 0.0004689
```

Model output 5: Linear regression model for grade ~ total number of errors.

In comparison, the output of a multiple linear regression model for absolute error frequencies, including the text length feature, shows that the model was able to account for the confounding text length factor. The regression coefficients for the feature ‘*total number of errors’* are now almost the same (-0.0067 for the relative frequencies and -0.0077 for absolute frequencies). The variance inflation factor for the second regression model with both text length and raw error frequency is, at 1.24, relatively low and can be ignored for this model.[133]

```
Call:
lm(formula = grade_numeric ~ length + errors_all, data = questionary)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64265  -0.66457   0.03158   0.57621   2.94458

Coefficients:
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   1.9914388    0.1255690    15.859     <2e-16 ***
length        0.0021806    0.0002332     9.349     <2e-16 ***
errors_all   -0.0077463    0.0029934    -2.588     0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8915 on 346 degrees of freedom
Multiple R-squared:  0.2072,   Adjusted R-squared:  0.2026
F-statistic: 45.21 on 2 and 346 DF, p-value: < 2.2e-16
```

Model output 6: Linear regression model for grade ~ text length in tokens and total number of errors.

In the effect plots for the multiple regression model (Figure 42), showing the regression lines for individual marginal effects, we see the positive effect of text length and the negative effect of error frequency on the model prediction for the holistic grade. As can be seen in the left plot, the linear regression model also produced predictions above grade level ‘*5-excellent’* as the model does not know about the inherent ordinal and limited scale of the dependent variable.

[133] James et al. (2013) states a variance inflation factor of 5 or 10 to be a general rule of thumb for multicollinearity concerns. The threshold of 10 for the variance inflation factor is also used by Gries and Deshors (2014).

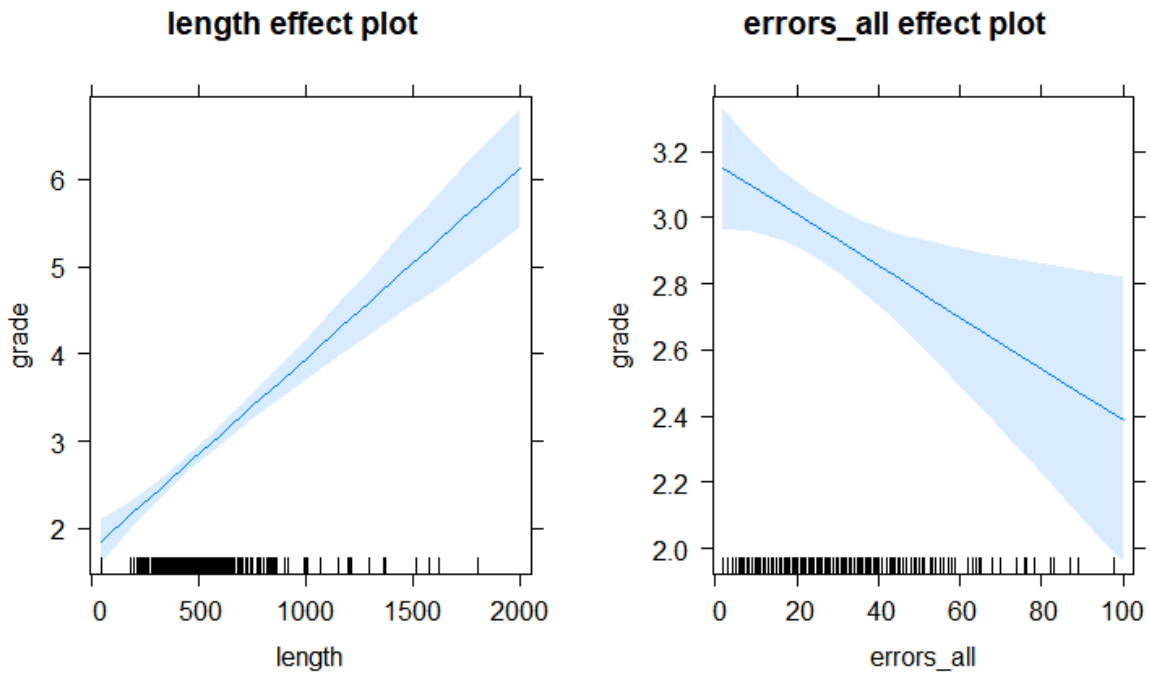


Figure 42: Effects plots for text length and total number of errors for absolute error frequencies.

Summary

Contrary to statistical analysis with regression models, the typical predictive modelling in computational linguistics or computer science does not necessarily account for confounders or multicollinearity, because the focus there is to build a model with the lowest possible error rate (i.e. highest possible accuracy). Individual feature importances and feature effects including confounders, moderators, or other types of biases are thus often ignored, as their contribution to the ultimate goal of better prediction performance is not directly obvious.[134]

However, if a predictive model is to be used for interpretation of feature importances and feature effects in order to get insights from the data (data mining), one needs to be aware of confounding effects. This entails normalizing the data and thus controlling for confounders preliminarily, or alternatively including known confounders and singling out the effect of individual variables. Moreover, strategies for detecting possible confounders in complex feature sets need to be taken, because the presence of confounding effects might not be immediately obvious, when using large, possibly automatically extracted, feature sets.

For the following experiments, I use the normalized error frequencies (relative number of errors per 1000 words), in order to investigate model performances for different error types without the bias introduced by different text length. Table 34 shows the updated correlation matrix for relative error counts.[135]

[134] But see e.g. Guidotti et al. (2019) for reasons that model interpretation and domain-knowledge are equally important for prediction focused tasks.

[135] This normalization strategy was used instead of including the text length variable in order to be able to evaluate if the learning algorithms can actually predict the grade only on the basis of error frequencies. Knowing that the text length is a relatively good predictor of text quality, the models would probably perform better than the baseline using only the information from the text length.

11.2.2 Distributional issues II: Outliers and investigation of individual texts

As one can see in the feature set description in section 9.2.2, many of the error frequency features have very skewed distributions. Skewed distributions are often problematic for data analysis, as there is little evidence for high values (or vice versa low), which makes predictions in those areas less reliable. Moreover, if there are clear outliers in the data, the model might be highly influenced by these values (e.g. through biased regression lines or inflated variance of features in neural networks).

Indeed, most variables are very skewed, which can be seen when looking at descriptive statistics for the relative error frequencies of the data (Table 35) and specifically in the high standard deviation compared to the mean value. The high maximum values also suggest the existence of outliers in the data.

Error	***min***	***max***	***mean***	***median***	***sd***
all	*6*	*230,6*	*46,4*	*41,0*	*27,8*
orth	*0*	*143,5*	*13,7*	*10,6*	*14,4*
orth_insertions	0	12,9	1,1	0,0	1,9
orth_lcp.cap	0	47,1	5,0	3,2	6,3
orth_omissions	0	65,9	2,4	0,8	5,0
orth_other	0	9,9	1,1	0,0	1,9
orth_sep.tog	0	20,0	2,9	2,2	3,1
orth_transpos	0	28,7	1,3	0,0	2,4
punc	*0*	*80,4*	*18,1*	*15,7*	*12,9*
punc_doppelp	0	5,4	0,2	0,0	0,8
punc_punkt	0	7,1	0,5	0,0	1,2
punc_nonkomma	0	11,9	1,0	0,0	1,8
punc_komma	0	78,6	17,1	14,4	12,7
punc_unkn	0	3,4	0,1	0,0	0,4
gram	*0*	*42,6*	*7,6*	*6,8*	*6,0*
gram_anak	0	5,2	0,3	0,0	0,8
gram_corr	0	24,5	4,4	3,6	4,1
gram_inco	0	21,3	1,0	0,0	1,9
gram_infl	0	8,0	0,6	0,0	1,1
gram_redu	0	21,3	0,1	0,0	1,2
gram_uncl	0	6,8	0,3	0,0	0,8
gram_unknown	0	10,1	0,6	0,0	1,1
gram_woor	0	3,7	0,2	0,0	0,7
lex	*0*	*55,4*	*19,0*	*18,0*	*10,4*
lex_form	0	21,3	4,5	3,6	4,0
lex_meta	0	21,3	1,2	0,0	2,2
lex_semdenot	0	29,1	6,8	6,0	5,3
lex_semkonnot	0	45,5	5,8	4,8	5,3
lex_stil	0	7,7	0,7	0,0	1,2
lex_EW	0	49,5	12,0	11,2	7,4
lex_FS	0	30,8	7,0	6,4	5,3

Table 35: Descriptive statistics for error types in subcorpus of 349 fully error annotated texts.

The box-and-whiskers plots in the following (see Figure 43 to Figure 47), show outliers for the relative error frequencies for the two highest levels of error categories in feature set 2 (coarse error categories *orthography*, *punctuation*, *grammar* and *lexis* as well as for the first level of subcategories of these errors). The plots have been produced with the Boxplot function from the R package `car` that allows to plot and label a custom number of outliers with the respective text ids. The outliers shown in the figures, represent values that are above an upper fence, defined as Q3+1.5*sd. The numbers in the figures show the text ids for the three highest outliers.

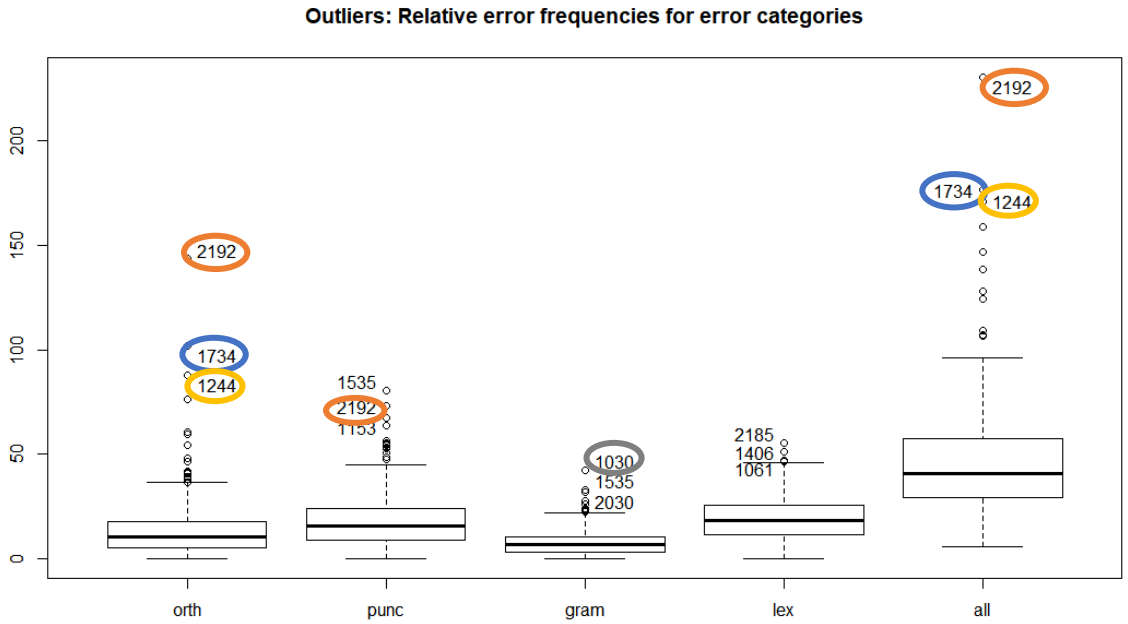


Figure 43: Outlier analysis using box-and-whiskers plots with highlighted outliers for all four error categories.

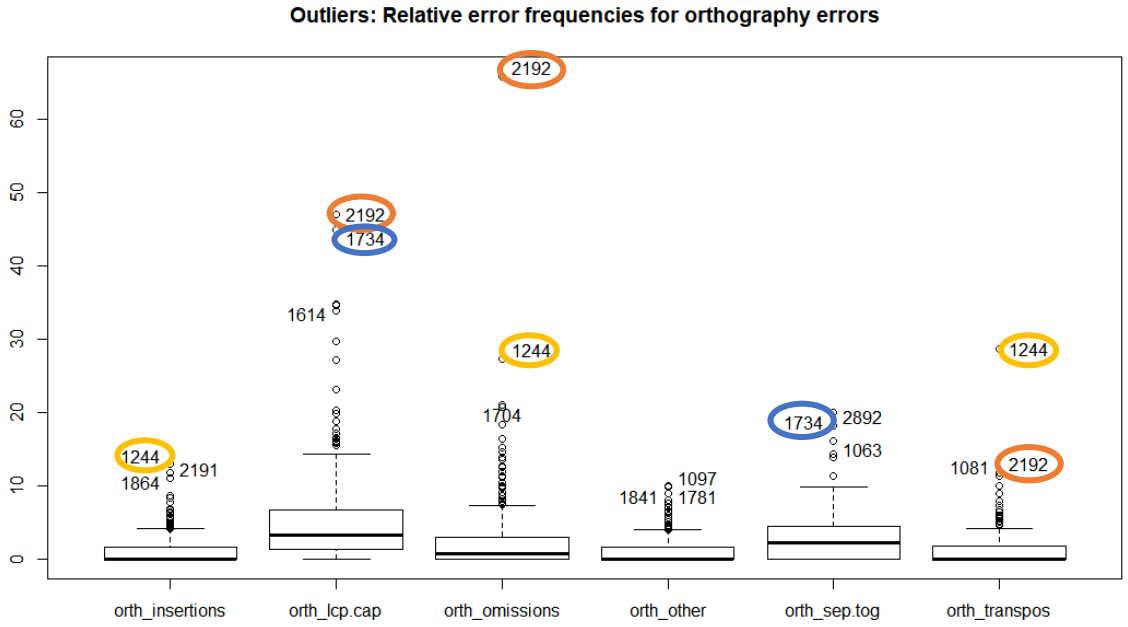


Figure 44: Boxplots with outliers for orthography error types.

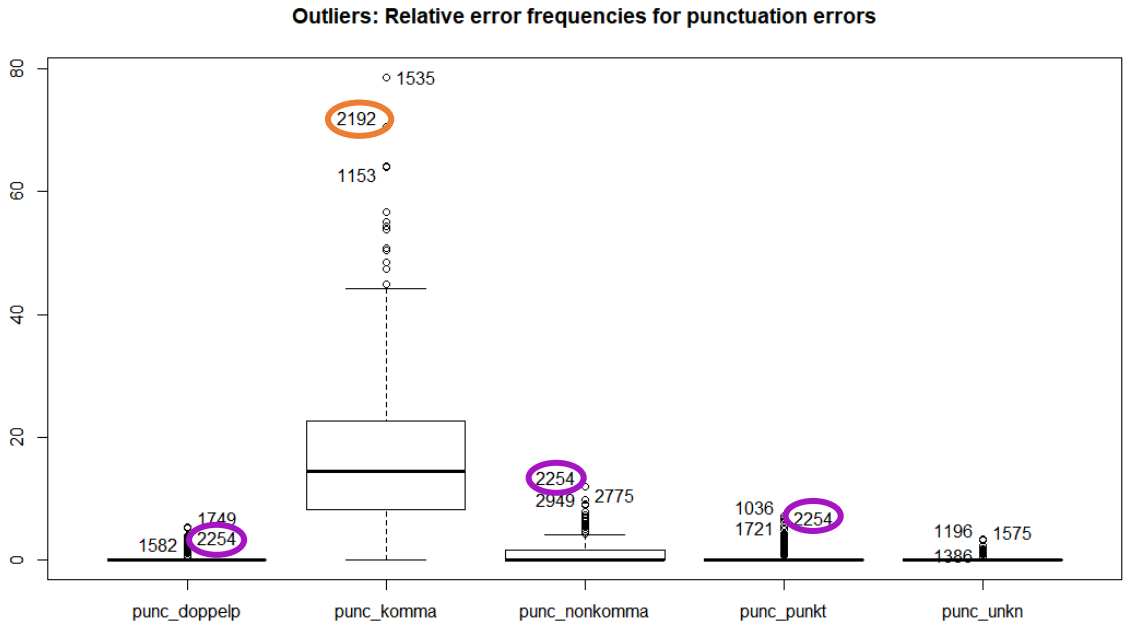


Figure 45: Boxplots with outliers for punctuation error types.

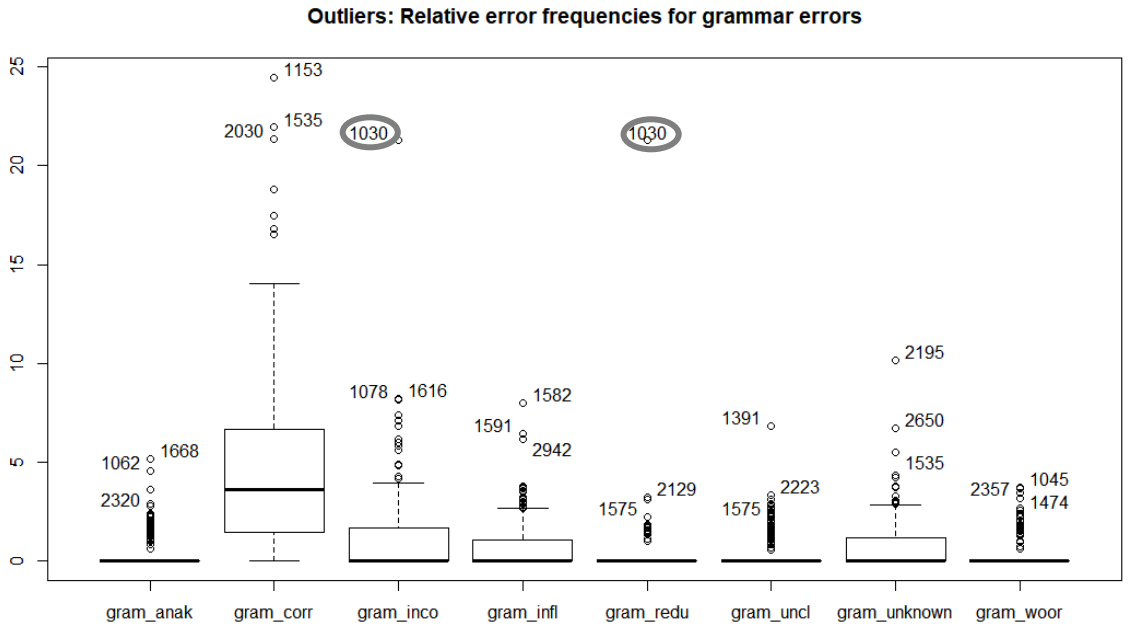


Figure 46: Boxplots with outliers for grammar error types.

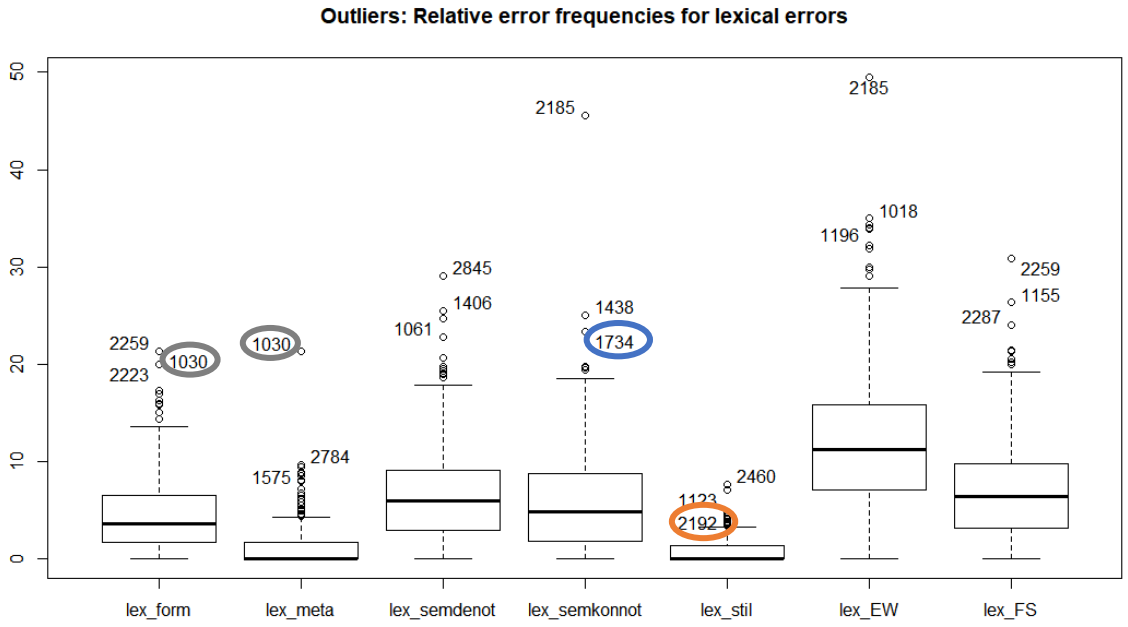


Figure 47: Boxplots with outliers for lexical error types.

These outliers can stem from errors in the data but also be signs of particularly interesting data points. In any case, it is worth investigating the most suspicious data points in detail. When counting the occurrences of text ids in the outliers, we can see that there are a number of text ids that are suspicious for various types of error frequencies[136]. ID 2192 occurs eight times in the figures and has particularly high values for orthography, punctuation and as well as errors regarding lexical style. The same is true for ID 1734 that shows high values for five error categories and has the second highest sum of errors. ID 1244 has very high values for three different types of orthography errors and a very high sum of errors as well. ID 2254 is only salient for punctuation errors, where it occurs as outlier in three of the five error categories. Finally, ID 1030 is suspicious for its high error frequencies in grammar and lexical errors. When looking into the evaluations of these texts, we can indeed see some concerning characteristics of the texts. While the texts for ID 2192, ID 1734 and ID 1244 have, as expected, lower grades (grade level '*2*' for ID 2192 and ID 1734 and grade '*3*' for ID 1244) and probably a weak writer with problems in linguistic accuracy (see excerpt of text ID 2192 below), the texts with ID 2254 and ID 1030 scored the grade '*4*' and '*5-excellent*'. For further investigation, the texts for ID 2254 and ID 1030 as well as an excerpt of the text for ID 2192 are shown and discussed below.

Text ID 1030

> Meiner Meinung nach hat der Schriftsteller völlig recht.
>
> Man muss wirklich froh sein, wenn man die die Jugend überlebt, denn jeder "normale" Jugendliche ist mindestens 100 in einer lebensgefährlichen Situation.
>
> Holzwerkstoffe aus Massivholz. ein genannter Dummkopf.
>
> ******
>
> **Grade**: 5-excellent

Looking into the text for ID 1030 we see that this text does indeed represent a dubious observation. While the outliers only indicated a high value for grammatically incomplete and reduced sentences,

[136] The texts discussed are highlighted with different colours in the graphs.

the writer of the text did in fact not respond to the input prompt but expresses his lacking commitment to fulfil the task and participate in the study. Moreover, the excellent rating for the text might be erroneous or at least not what we would expect from the rater. The text has very few form errors but does show some incomplete sentences that are of course greatly overvalued in the normalized error frequencies that extrapolate the amount of errors for 1000 words. This is a problem for accuracy measures[137] that base on normalized error frequency. If the normalization unit is higher than the average text length, writers appear in the analysis as if they would have made more errors than they actually did. However, looking at the outliers for the error frequency dataset, we could identify this atypical text and are left with the decision whether to exclude it in the further analysis as it does not seem to be sensical or will, because of the excellent rating, at least not contribute to the predictive performance.

Text ID 2254

> Das Interview mit Hans Magnus Enzensberger behandelt das Thema der Jugend. Der Schriftsteller und Essayist zeigt seine kritische und negative Meinung gegenüber jungen Menschen und setzt sich mit deren Schwächen auseinander.
>
> Sind jedoch die Jugendlichen derartig undiszipliniert und schlimm, wie es von Herrn Enzensberger behauptet wird?
>
> Die Jugend ist ein sehr oft behandeltes Thema. In Zusammenhang mit ihr fallen oft Begriffe wie Alkohol, Drogen, Chaos, Verwüstung oder auch Egoismus und Faulheit. Es wäre eine Lüge, würden die Jugendlichen dies abstreiten. Jedoch ist eine Verallgemeinerung bezüglich dieser Themen auf alle Jugendliche nicht korrekt und vor allem nicht fair. Jeder Erwachsene war einmal ein Jugendlicher und weiß, dass auch dieser Abschnitt nicht immer leicht ist. Es gibt viele Höhen und Tiefen und jeder geht auf seine Weise damit um.
>
> "(...) die Jugend ist sowieso keine beneidenswerte Phase des Lebens. (...) Ein junger Mensch ist labil, unsicher, schwankend, hat keine Souveränität, macht jede Dummheit mit." Die Phase der Jugend ist meiner Meinung nach die wichtigste Zeit des Menschen. Es ist die Zeit des Überganges vom Kind zum Erwachsenen. Dies ist also eine lange Ausbildungsphase, damit man später in der Lage ist eigenständig zu denken, auf eigenen Füßen zu stehen und natürlich bereits Erfahrungen hat bezüglich Beruf, Liebe aber auch Alkohol: Ich bin auch der Meinung, dass es jener Zeitabschnitt ist, in welchem man die meisten Fehler macht und die meisten Dummheiten begeht. Doch ohne diese Fehler und Dummheiten wäre der Mensch nicht in der Lage, sich zu verbessern und etwas zu lernen.
>
> "Denken Sie nur an diese Klamottensucht". Alle kennen das Sprichwort "Kleider machen Leute". Dies wird jedoch auch von Erwachsenen praktiziert und nicht nur von jungen Menschen. Selbstverständlich wollen die Jugendlichen gut aussehen, schließlich geht es um die ersten Beziehungen der Menschen zum anderen Geschlecht. Wenn man hierbei gut aussieht, ist bereits ein großer Schritt getan. Grund hierfür ist wahrscheinlich eine der großen Schwächen der Jugend, nämlich die Oberflächlichkeit.
>
> Das Leben in der Diskothek ist auch ein bereits oft behandeltes Thema, da es vor allem mit Alkohol in Verbindung steht. Diskotheken sind jene Orte, wo Jugendliche unter sich sind, ohne gestört zu werden von den Eltern oder anderen Erwachsenen. Sie können hier den Stress und die Anspannung des Alltages vergessen und sich den Freunden und Partnern widmen. Natürlich kommt es dabei nicht selten zum Alkoholkonsum, doch ich finde jeder sollte seine Grenzen selbst kennenlernen und so viel konsumieren, wie er es für richtig hält. Bei Grenzüberschreitungen wird man aus diesen Fehlern lernen und man erhält sowieso eine entsprechende Strafe dafür.
>
> Zusammenfassend bin ich also der Meinung, dass Jugendliche keineswegs fehlerfrei sind, jedoch hoffe ich, dass jeder Mensch eine gewisse Akzeptanz für diese Fehler aufbringt, damit junge Menschen daraus lernen können. Somit ist für mich die Aussage "Man muss froh sein, wenn man das überstanden hat" umzuändern in "Man kann froh sein, wenn man dies überstanden hat", d.h. ich finde jeder soll selbst entscheiden, ob er die Zeit der Jugend genießt oder nicht.
>
> ******
>
> **Grade**: 4

Text ID 2254 has also been rated relatively well. This text is especially salient in terms of not comma-related punctuation errors, according to the values in Figure 45. However, if we inspect the text, it seems to be of rather high quality, not atypical for student essays and probably just salient in the

[137] See Housen and Kuiken (2009) and Wolfe-Quintero et al. (1998) for the CAF-Model (complexity – accuracy – fluency) model of second language acquisition that defines linguistic accuracy, i.e. formal correctness of the language, as one of the main elements of language competences, as well as Polio and Shea (2014) for methods to measure linguistic accuracy.

outlier analysis because of its systematic deviance from the normative quoting style[138] that is high in comparison to other texts as many texts do not quote at all and therefore cannot make those mistakes[139].

Text ID 2192

> Der Schriftsteller und Essayist Magnus Enzenberger behauptet in einem Interwife dass, die Jugent eine Schlimme Zeit sei.
>
> Ich finde er hat zum teil Recht.
>
> Ja di Jugend macht jeden Plözin oder Dummheit mit, oder wenn einer sich etwas super tolles neues kauf dan möchte ein anderer das auch.
>
> Ich bin deshalb in disem Punkt seiner meinung da ich es oft mit erlebe mit Verwanten oder in meinem Freundes Kreis.
>
> [...]

Text 2192 on the other hand is responsible for eight of the highest outlier values reported in the boxplots. It has the highest number of errors in total (119 errors in absolute) as well as the highest number of orthography errors. The high ratio of omission errors is especially salient for this text but also capitalisation errors, transposition errors and comma errors are particularly frequent in this text. Furthermore, stylistic lexical misuses are occurring more often than in other texts. A brief look into the actual text is enough to expect a writer with serious weaknesses in German orthography[140]. There are systematic errors concerning vowel length, capitalization, tenues, etc. and the sentences are noticeably short. Nevertheless, the essay received the grade *'2'* from the annotator, which means a sufficient score. The text seems to be particular, but still valid as a realistic observation of a writer with difficulties in linguistic accuracy.

Summary

The examples showed that not all of the outlier values one can observe in the data are unrealistic and therefore invalid measurements. Some outliers are natural outliers, i.e. they can occur in such form in real life (e.g. when writers have dyslexia). Datapoints including such outliers do not necessarily need to be excluded but may be treated differently in order to limit the effect of the outlier on the full model (e.g. by capping the values to a reasonable maximum). Furthermore, some outliers exist, because the used measure (relative error frequencies) is probably not the best possible approximation of formal correctness of a text (extrapolated errors, existence of error in a structure vs. not used structure). The use of other operationalizations for the concept of formal correctness could be considered for future studies. Other outliers led to texts that should probably not be included in the corpus. If the text does not present a valid observation of the indented sample, or the label of the dependent variable seems to be erroneous (grade level *'5-excellent'* for non-serious text) it is reasonable to exclude the observation in the analysis. This also hints at the fact that the investigation of bivariate outliers (e.g. texts, which show very different relationships between predictor and descriptor variables) like the unrealistic holistic grade for text 1030, should probably be considered in further studies. However, for the remainder of this study I limit myself to excluding text ID 1030.

[138] According to the standard orthography and punctuation rules defined in (Duden, 2005, 2006), quotations have to be introduced with a colon and ended with a comma, if the sentence does not end at the end of the quotation.

[139] This is another known problem with error frequency-based accuracy measures, as errors are only possible when the according structure is used. This puts correctly performed structures at the same level as unused structures (Buttery et al., 2012).

[140] Words that deviate from the standard are highlighted in red.

11.2.3 Dealing with complexity in the data

11.2.3.1 Small relations and many features – noise and classifier performance

Finally, when trying to build a prediction model that classifies the holistic grade (5 levels) with the relative error frequencies using the same methodology as in experiment 10.2.1.1, there is no classifier that exceeds the classification accuracy of the majority baseline (see Table 36).

Dataset	Baseline	NB	LogReg	PART	J48	MLP	SMO	RandF
Accuracy	38.97	32.72	27.17	29.05	27.62	29.32	31.12	36.37

Table 36: Classifier Accuracy for classifiers trained on the full feature set of relative error frequencies.

I thus try to reduce the complex, hierarchical feature set by removing less relevant features using the ReliefF feature ranking algorithm. I choose the ReliefF method, as it showed the best results in the previous experiments (see section 10.2.2.4). Table 37 compares the accuracy results for the whole feature set with the best 200, 100, 50, 25 and 10 features selected according to the ReliefF algorithm.

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
best200	38.97	30.84	29.84	29.97	29.51	34.01	30.75	36.71
best100	38.97	29.43	31.92	30.72	29.43	35.91	35.27	37.04
best50	38.97	27.87	36.22	30.40	29.57	38.68	34.44	36.39
best25	38.97	24.32	34.96	30.73	31.11	**40.60**	34.25	34.41
best10	38.97	26.65	37.96	31.75	31.40	**40.63**	34.24	33.29
all	*38.97*	*32.72*	*27.17*	*29.05*	*27.62*	*31.12*	*29.32*	*36.37*

Table 37: Comparison of classifier accuracy for selected feature sets for relative error frequencies

There are values that are slightly higher than the baseline (SMO with the best 25 or best 10 features).

However, in this example, none of the reduced feature sets showed a significant increase over the baseline, i.e. it seems that none of the classifiers could learn anything about the holistic grade from the relative error frequencies.

11.2.3.2 Binary features vs. frequency features: A detailed classifier comparison

Next, I try to simplify the data and reduce skewedness in the variable distributions with two different techniques. First, I binarize the error features so that the predictors now show error presence instead of error frequency. Second, I cap very high outlier values to a maximum of the upper fence (3rd quartile + 1.5*IQR) to reduce the variance in the predictor variables. When we compare the results achieved with the relative error frequencies from the table above, with those achieved for the binarized features and the outlier capped relative features (Table 38), we see that binarization seems to have a slight positive influence on the classifier performance, but none of the accuracies actually reached the baseline accuracy of 38.97%. The capped features seem to perform even worse than the non-capped features with the outliers. This might be due the fact that capping the features resulted in a reduced feature set, as many features turned into zero-variance feature after capping. In general, it seems that there is so little evidence for relations between error features and the holistic grade that reducing the skewedness of the predictor variables had no real effect either.

Dataset	Baseline	NB	LogReg	PART	J48	SMO	MLP	RandF
Binarized all	*38.97*	*29.14*	*23.63*	*29.43*	*29.83*	*27.61*	*29.31*	*36.57*
best200	38.97	30.63	22.69	29.97	28.48	30.15	31.23	37.80
best100	38.97	33.43	32.65	31.68	32.61	34.08	35.96	37.68
best50	38.97	33.23	36.26	34.58	34.21	34.81	36.59	**39.28**
best25	38.97	32.66	35.87	35.50	35.81	37.05	34.29	38.28
best10	38.97	33.29	34.96	33.35	35.56	38.17	33.99	31.38
Capped all	*38.97*	*25.67*	*24.33*	*28.05*	*27.10*	*26.88*	*25.47*	*33.24*
best50	38.97	25.84	29.31	29.10	28.93	29.28	28.69	33.94
best25	38.97	25.27	34.09	30.68	31.43	37.77	32.66	36.81
best10	38.97	26.50	36.61	33.40	34.81	**39.28**	35.01	36.39
Relative all	*38.97*	*32.72*	*27.17*	*29.05*	*27.62*	*31.12*	*29.32*	*36.37*

Table 38: Comparison of classifier accuracies for transformed error frequency features. Binarized values show importance of error presence, capped values show relative frequencies were outlier values have been capped to reduce variance in the predictor variables

11.2.3.3 Zooming in & zooming out: What to gain from aggregating or splitting error types

At last, I make use of different levels of granularity encoded in the hierarchical error annotation scheme by approaching the feature set conceptually and choosing and comparing theoretically informed subsets of features instead of using automatic methods for feature selection or statistical methods for variance reduction. I try to reduce noise by manually removing redundant or overly fine-grained variables with almost no information and compare different levels of the hierarchical feature set, as well as different categories.

Given that the effort needed to annotate and categorize errors increases with the level of granularity, I therefore investigate the effects of errors on different levels of hierarchy. I first investigate simple bivariate correlations and visualize them according to the hierarchical structure of the variables, to allow better interpretation of the results. Then I compare the prediction performance of a simple model based on the coarse error categories (orthography, punctuation, grammar and lexis) with the performance of more fine-grained but probably noisy error categorizations, using the WEKA experimental setup from the previous section as well as hierarchical regression using R.

Bivariate correlations on different levels of granularity

I investigate bivariate correlations between the holistic grade and the error categories on two different levels of granularity. The four coarse categories of punctuation, grammar, orthography and lexical errors make it possible to see general trends. The values displayed are total frequencies for all errors of this category. Furthermore, the four categories are split into groups of errors as described in section 3.3.3. All further (lower) levels of error categories present in the data were ignored in this analysis in order to limit the complexity and only used for the classification experiments presented above.

Investigating Spearman rank correlation coefficients (see also the visualizations below), we can observe significant but very weak correlations for punctuation errors, orthography errors and lexical errors. In general, orthography, punctuation and lexis are – as expected – negatively correlated with the text quality. However, correlation coefficients are very low (between -0,15 and -0,17). There was no significant correlation between the normalized number of grammar errors and the holistic text quality judgments.

Total # of errors (-0.181)					
Orthography (-0,111)	*Punctuation (-0,170)*		*Grammar (-0,054)*	*Lexis (-0,150)*	
- Insertion (0.005)	*Comma (-0.183)*	*Noncomma (0.024)*	*- Anacoluthon (0.023)*	*EW (-0.127)*	*FS (-0.122)*
- Omission (-0.108)	*- Wrong (0.000)*	*- Full stop (0.074)*	*- Correspondence (-0.010)*	*- Form (0.022)*	*- Form (-0.071)*
- Transposition (0.068)	*- Missing (-0.193)*	*- Other (-0.08)*	*- Incompleteness (0.004)*	*- Meta (-0.081)*	*- Meta (0.102)*
- Capitalization (-0.053)			*- Inflection (-0.063)*	*- Sem. denotation (-0.063)*	*- Sem. denotation (-0.044)*
- Compounding (-0.167)			*- Reduction (-0.012)*	*- Sem. connotation (-0.081)*	*- Sem. connotation (-0.039)*
- Other (-0.031)			*- Unclear (0.022)*	*- Style (0.009)*	*- Style (-0.048)*
			- Unknown (0.007)		
			- Word order (0.044)		

Figure 48: Hierarchical visualization of Spearman rank correlation coefficients for error types. Significantly related error types are coloured (p-value < 0.05).

When we take a deeper look into error types, investigating the subcategories of errors, we see weak negative correlations for text quality judgments and orthographic errors in general. However, the correlations regard especially compounding errors and (although less) character omissions. The correlations between the orthographic categories of capitalization, insertion or transposition of graphemes are very low. For punctuation errors that are significantly negatively correlated with text quality on a general level, only the missing comma is significantly related with the holistic grades for text quality. For lexical errors the total frequency of single word or phraseme errors in the lexicon show weak negative correlations with text quality, while individual error types show only very low correlations. There is no significant correlation with any of the grammar error categories.

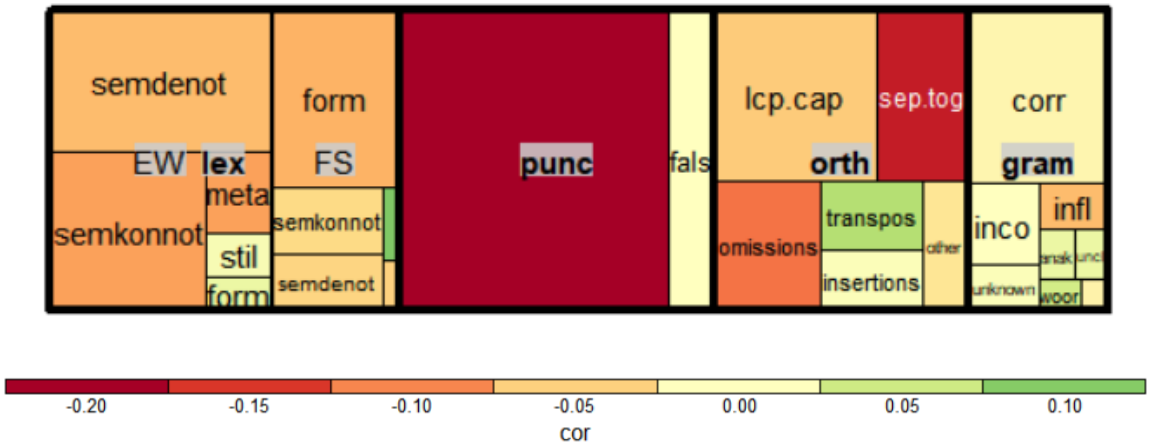


Figure 49: Treemap diagram displaying the proportional amount of errors per category and subcategory in the dataset. The diagram is coloured according to the correlation coefficients for bivariate Spearman rank correlations.

Classification approach

For the classification approach, I used SMO and random forest, as these two algorithms performed best in the previous analysis with this feature set and are relatively fast to train. For both data representations, relative and binarized errors, I compared the performance for 5-class prediction and the complexity-reduced 3-class prediction for:

- all features
- coarse categories (orth, punc, gram, lex)
- all second level categorization
- all third level categorization
- all orthography error types (and second level subcategories for orthography errors)
- all punctuation error types (and second level subcategories for punctuation errors)
- all lexis error types (and second level subcategories for lexis errors)
- all grammar error types (and second level subcategories for grammar errors)

No model had a significant increase over the baseline accuracy. However, there were a few classifiers that showed some (although insignificant) increase over the baseline for the 3-class prediction results (see Table 39).

Feature (sub)set	***SMO***	***RandF***
Two levels for orthography (bin)	42.10	29.97
All lexical error types (rel)	37.62	41.87
All grammar error types (rel)	37.68	41.55
Two levels for orthography (rel)	39.48	40.75
All orthography error types (rel)	38.60	40.15
4 coarse categories (bin)	36.59	40.12
Two levels for punctuation (bin)	39.02	40.03
All errors (rel)	39.18	39.12
All punctuation error types (bin)	34.38	39.09
Two levels for lexis (bin)	39.88	38.91

Table 39: Accuracies higher than the baseline (38.97) for 10-fold CV for SMO and random forest classifiers trained with different subsets of the relative and binarized error frequencies.

Hierarchical regression

In the present analysis the data contains various levels of granularity in the independent variables (columns), as frequencies for errors are aggregated error types or split categories of theoretically similar error types. Complex, hierarchical datasets are often problematic in data analysis. But while hierarchy in the observations is usually dealt with more complex analysis methods like mixed-effects models (see section 10.2.4.3), hierarchy in the independent variables can be addressed by reasoned adding of variables into the model and comparing its performance. For regression modelling this strategy is often called hierarchical regression.

I therefore use hierarchical regression in order to compare the predictive power for error categories of different types of granularity and control the final model for the multicollinearity introduced by possibly related predictor variables. I first build a model using just the coarse error categories orthography, punctuation, grammar and lexis and reduce the model using backwards model selection till the model only contains significant predictors. This first model shows two significant predictor variables: the (relative) sum of all punctuation errors (`punct`) and the (relative) sum of all lexical errors (`lex`) and reaches a model performance of 0.04505 R^2 and 0.03953 adjusted R^2. In a similar manner I train and select a final model for each of the four categories using its subcategorizations. For orthography, only errors regarding word separation and compounding (`orth_sep.tog`) and errors of the residual class *others* (`orth_others`) were significant predictors (0.04472 R^2 and 0.0392 adjusted R^2), for punctuation, the relative frequency of missing commas (`punc_komma_fehl`) was the only significant factor, for lexis only the (relative) amount of all lexical errors (`lex`). The model trained on the subcategories of grammar errors was not significant at all and none of the correlation coefficients for the grammar subcategories was significantly different from 0. Finally, I use all the significant factors In the resulting final model, only the variables for the (relative) sum of lexical errors and the orthography errors regarding word separation and compounding remained with variable significant effects, leading to an overall model performance of 0.04754 R^2 and 0.04203 adjusted R^2 (see also Table 40).

	significant predictors (F-test)	**R^2**	**adjusted R^2**
coarse error categories	punct**, lex*	0.04505	0.03953
orthography subcategories	orth_sep.tog**, orth_other*	0.04472	0.0392
punctuation subcategories	punc_komma_fehl***	0.03166	0.02887
grammar subcategories	-	*(model not sign.)*	*(model not sign.)*
lexis subcategories	lex**	0.02189	0.01908
combined predictors	orth_sep.tog**, lex*	0.04754	0.04203

Table 40: Hierarchical regression modelling results for explaining the variance in the holistic grade with relative error frequencies.

The remaining features of the two highest levels of error categories that significantly contributed to explaining the variance in the dependent variable in this model are the sum of lexical errors and orthography errors regarding compounding and word separation. Both variables effect the predictions for the grade level negatively (see effect plots in Figure 50). However, in total they only explain 4,2% of the variance in the dependent variable. The model is significant at a confidence level of 99.9%. The variance inflation factor (see also section 10.2.4.4) for the features in this model is relatively low (1.016 for both features), so that we can ignore possible effects of multicollinearity in the model.

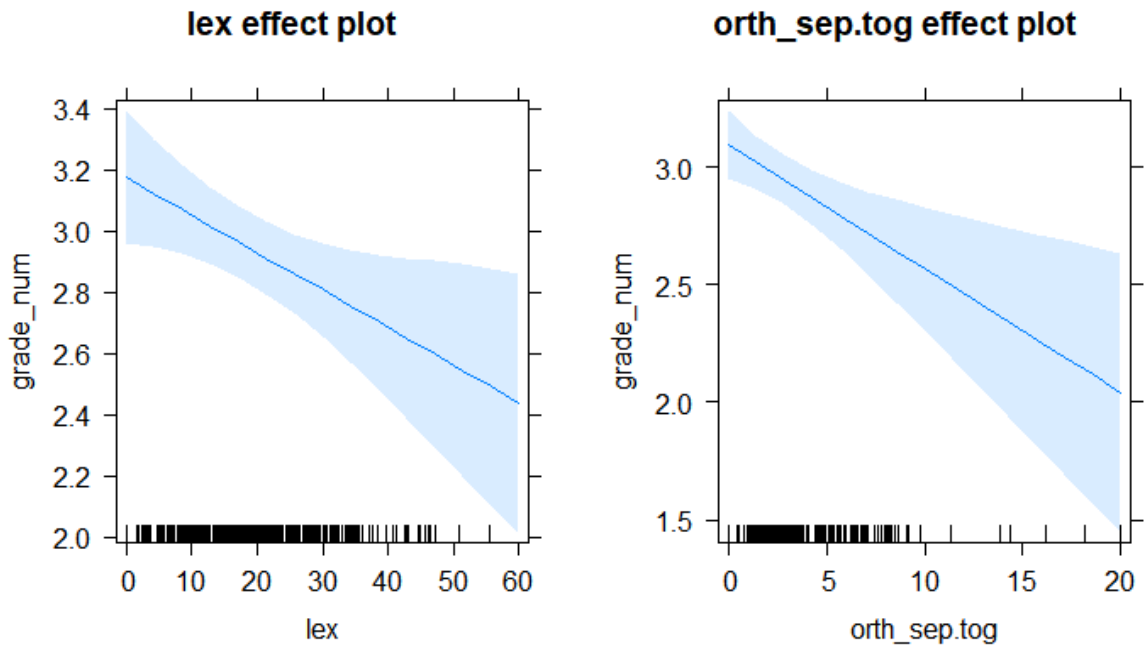


Figure 50: Effects plots for lexical errors (total number) and orthography errors concerning word separation and compounding.

Summary

In total, manually investigating the different hierarchical levels of error annotations for the student essays only yielded few insights into the data. The bivariate correlation analysis showed very weak correlations between individual error types (e.g. missing commas, or the total amount of lexical errors) and the holistic grade. Some of the significant correlations from this monofactorial analysis could not be found in the presence of other predictor variables in the hierarchical regression model. They seem to be relevant but already represented by other features or biased by text length. This might be the case for the correlation between missing commas and the holistic grade, as only longer and more complex and elaborate texts might even have enough complex sentences to forget commas. Furthermore, the low correlations and effect sizes as well as the machine learning models that fail to learn from the data, show that there is not enough evidence for an effect of form errors on the holistic grades.

Conclusion of the analysis on the error annotation dataset

The experiments in this section showed some of the problems one can encounter when working with frequency data from language corpora and in particular when working with error frequencies of native writers from secondary school. Frequency counts for linguistic annotations like error frequencies need to be normalized for text length if the text length is not held constant in the data and outliers or high variance features can naturally occur in this type of data. Furthermore detailed, hierarchical annotation schemes provide additional information but also raise the complexity of the dataset tremendously, serving for models that are barely interpretable. Most of all, however, the low monofactorial correlations, the low effect sizes, and the non-predictive models (machine learning models that fail to learn from the data), show that there is barely any evidence for a relationship between form errors and the holistic grades.

12 Interpretation of complex feature sets

Feature set 3: Measures of linguistic complexity and text cohesion

While the previous analyses focused on

a) manually annotated high-level text characteristics with relatively high correlations between the text characteristics and the holistic grade and
b) inherently structured, manually chosen frequency features for form errors,

the third set of experiments now makes use of a large set of automatically extracted linguistic features regarding linguistic complexity and text cohesion (see feature set description in section 9.2.3). Using this feature set several experiments are performed to illustrate issues in the application of data mining methods in linguists. This section thus focuses particularly on state-of-the-art analysis and interpretation of complex feature sets.

12.1 Methodology

In this section, I use the whole set of 646 text evaluations in order to have as much data as possible for training complex models. I thus duplicate the values for the linguistic complexity features for text ids with more than one evaluation. Additionally, I remove the observation for text 1030 (erroneous outlier, cf. section 11.2.2) and the three consensus evaluations (as in section 10). The dataset for the following analyses thus contains 642 observations.

12.1.1 Monofactorial correlation analysis

As a first step, a preliminary correlation analysis is performed to get an intuition of the feature set and possible relations using Spearman rank correlation for the ordinally ranked holistic grades. I report all significant (alpha = 0.01) correlations above a minimal correlation coefficient of 0.2 and give a short summary on the found relations.

12.1.2 Limits of intrinsically interpretable models

Next, the limits of so-called interpretable models are discussed and illustrated on the basis of this large multicollinearity-prone feature set.

(1) Performance – Complexity – Interpretability

To illustrate the trade-off between interpretability and complexity, interpretability and performance, and performance and complexity, I compare performance measures and quantitative interpretability measures (cf. Molnar et al., 2019) for tasks of different complexity (categorical vs. numerical variables, varying number of classes and number of features).

(2) Difficulties in model selection

Attempts with linear and logistic regression models are made to illustrate and discuss the interpretability of regression models and problems with the current practice of model selection when using complex datasets (large number of predictor variables or nested observations).

12.1.3 Black box interpretation

Finally, I show model-agnostic strategies for interpretation and explanation, starting with simple planned feature set comparisons, feature ranks provided by feature selection methods that can be easily conducted via WEKA and ending with methods from state-of-the art interpretation tools that are applied to complex, well-configured black box models.

(1) Comparing classifier performance

In order to get a first impression on the relevance of different types of complexity measures, I compare classifier results for subgroups of feature set 3. Measures for lexical complexity, syntactical complexity, text complexity and cohesion are divided and used in different training instances, to show their respective effect on the classification accuracy. I also compare the results to the ones generated with the other feature sets in the previous analysis.

(2) Feature selection and feature ranks

The set of features is then subsampled using feature selection and the changing classifier results as well as the feature ranks are discussed.

(3) Interpreting model internals: Measures for variable importance, variable effect and interaction effects in complex predictive models

The final part shows which classifier results can be achieved with further optimized, complex black box models and how these models can be interpreted with recent interpretation tools. In order to do so, three different model setups have been used to train text classifiers for 5-class prediction of holistic grades using the scikit-learn machine learning library for Python[141].

- *Linear neural network*
  The first model was a simplistic linear neural network model, based on the MLPRegressor (scikit-learn implementation for a multilayer perceptron for regression) with one hidden layer of one neuron. The limited-memory BFGS (‘lbfgs’) solver for weight optimization was used, as it is recommended when working with smaller datasets.

- *Non-linear neural network*
  The second model was a non-linear neural network model, based on the MLPRegressor with one hidden layer of three neurons. Instead of the linear identity activation function this model uses a rectified linear unit function (f(x) = max(0,x)) as the activation function. The limited-memory BFGS (‘lbfgs’) solver for weight optimization was used, as it is recommended when working with smaller datasets.

- *Random forest*
  The third and last model was a tree-based random forest model, based on the RandomForestRegressor of the scikit-learn library. The model was trained with a maximum tree depth of 6 and a minimum sample split criterion of 5 as parameters that stop the tree from growing too deep (cf. section 3.5.3).

[141] The Python environment allows to make more detailed custom adaptations and setups as well as advanced options for hyperparameter optimization. However, it also requires basic programming skills.

All three models made use of algorithms for regression models (*MLPRegressor* for neural network models and *RandomForestRegressor* for random forest models) and used the clamped and rounded continuous outcomes of the regression models in order to estimate prediction accuracy. In that way, it was possible to account for the ordinal nature of the holistic grades, while remaining comparable to the previous prediction results achieved with the WEKA data mining software[142].

Model selection and model evaluation
For each of the neural network setups, I trained several thousand instances of networks, with randomly initialized weights and a randomly chosen training set of 80% of the data. This approach was needed, as the dataset counts only 644 instances. With this dataset size most models do not have enough time to learn before they converge, making the initial weights crucial for the efficiency of the learning process. After the networks have been trained, I chose the best model according their performance on the remaining 20% of the data. In order to make sure that the model performance was not due to the distribution within the test set, I retrieved mean accuracies for different random weights for the exact same test set and used this as a second baseline, next to the majority baseline. For random forest models I used a similar approach, changing the initial weights for the random sampling used in the random forest algorithm to generate the base learners. The models have been evaluated using the OOB estimate (see section 4.3.3.2).

After I report on the best performance for the three different model setups for the feature set 3, the results are compared with the ones achieved with the standard configurations of the WEKA data mining software and with the results achieved when using feature set 1 or a combination of feature set 1 and feature set 3[143]. Finally the best performing model for feature set 3 is investigated using the SHAP interpretation library for Python (Lundberg & Lee, 2017).

12.2 Results

State-of-the-art research designs dealing with response variables that have multiple classes and/or operationalize abstract concepts, as well as a higher number of known or assumed predictors introduce many difficulties in the analysis and interpretation.

In the following I use an automatically extracted set of linguistic features designed to measure linguistic complexity and cohesion and analyse and interpret relationships (feature importances, feature effects and responses) as well as interactions in this data, before I add it to the previous feature sets in order to get one large, combined final feature set for the exploratory investigation.

12.2.1 Monofactorial correlation analysis

In both section 10 and 11, we saw that a simple bivariate analysis of correlations or associations is not always sufficient to investigate and understand complex problems. Experiments 10.2.3.3, 10.2.4.3 as well as 11.2.1 illustrated how such analyses can be influenced by confounders and are not able to show combinatorial effects. However, although they give an overly simplified picture of the actual structures in the data, monofactorial analyses can still inform one's intuition of the feature set and hint to possible relations.

[142] A separate implementation was done using the Python library of the TensorFlow machine learning platform (Abadi et al., 2016). However, the implementation proved to be more complex while achieving the same results as the scikit-learn library. It was also more difficult to integrate with post-hoc interpretation methods.
[143] Feature set 2 is excluded as the results as well as the available dataset proved to be insufficient in the previous experiments.

Therefore, I start also this section by investigating Spearman rank correlations between the linguistic complexity measures and the holistic grades. Table 41 shows the measures with significant correlations (alpha 0.01.) of at least 0.2.

Measure	rho
1. rootTTR*	0,387
2. nTokens	0,381
3. squaredNonAuxVerbTypesPerNonAuxVerb*	0,345
4. correctedNonAuxVerbTypesPerNonAuxVerb*	0,345
5. nSentences	0,344
6. probObjNotsPerTransition	-0,333
7. probNotNotsPerTransition	0,332
8. probNotObjsPerTransition	-0,33
9. probSubNotsPerTransition	-0,31
10. probNotSubsPerTransition	-0,303
11. lemmaFreqsPerTypeFoundInDlex	-0,301
12. annotatedTypeFreqsPerTypeFoundInDlex	-0,3
13. typeFreqsPerTypeFoundInDlex	-0,295
14. logAnnotatedTypeFreqBand6PerTypeFoundInDlex	-0,29
15. typeFreqsPerTypeFoundInSubtlex	-0,276
16. typeFreqsPerTypeFoundInGoogle00	-0,276
17. typeFreqsPerTypeFoundInKCT	-0,276
18. coveragePeriphrasticTenses	0,272
19. lemmaFreqsPerTypeFoundInKCT	-0,261
20. logTypeFreqsPerTypeFoundInSubtlex	-0,254
21. logTypeFreqsPerTypeFoundInGoogle00	-0,254
22. coverageDeagentivationPatterns	0,252
23. coverageTenses	0,251
24. coverageModifierTypes	0,243
25. sdVerbClusterSize	-0,243
26. TTR*	-0,234
27. AoA_sumTypesMinAoAPerTypeFoundInKCT	0,225
28. AoA_sumLemmaMinAoAPerLemmaFoundInKCT	0,225
29. lexTypesNotFoundInKCTPerLexicalType	0,219
30. AoA_sumTypesAoAPerTypeFoundInKCT	0,218
31. logTypeFreqsPerTypeFoundInKCT	-0,218
32. AoA_sumLemmaAoAPerLemmaFoundInKCT	0,217
33. lexLemmasNotFoundInKCTPerLexicalLemma	0,216
34. lemmasFoundInKCTPerLexicalLemma	-0,216
35. typesFoundInKCTPerLexicalType	-0,216
36. logLemmaFreqsPerTypeFoundInKCT	-0,215
37. typesNotFoundInDlexPerLexicalType	0,209

Table 41: Measures of linguistic complexity correlated (rho > 0.2) with text quality judgments ordered by their effect size.

The measure with the highest Spearman rank correlation coefficients is a root-transformed version of type token ratio[144], indicating **lexical diversity** in the text. However, this measure is known to be dependent on the text length, which is furthermore a known factor for text quality measured in holistic grades. We therefore cannot clearly state its actual importance for text quality without referring to other modelling techniques that can account for this. The same is true for the two measures of verb diversity (features 3 and 4) that directly depend on the length of the text. Furthermore, the third and sixth most important measures are indeed descriptive text length measures (number of tokens and number of sentences), for which we already know from experiment 11.1.1 that they are related with the holistic grades.

Features 6-10 are **cohesion features** indicating object transition (measured in the probability that an object or subject of one sentence was taken up in the next sentence or not). This seems to be a rather important measure for text quality. We see a clear negative bivariate correlation between all transition features that indicate that objects or subjects were not taken up in the following sentences. Moreover, looking at the feature names, it becomes obvious that all these features belong together and measure different categories of the same concept of transition of grammatical roles (cf. Galasso, 2014), some of them representing the inverse or at least partial inverse concept of the others.[145] This is problematic for many predictive models, as it naturally introduces multicollinearity.

The next important group of measures concerns the **elaborateness of the used vocabulary**. Variables on rank 11-17 and 19-21 as well as 31 and 36 are all different measures that try to operationalize how common or basic the vocabulary in the text is compared to reference corpora (see also description in section 9.2.3.2). They look at all the words in the text that occur in the respective reference corpus, see how frequently they occur in the reference corpus[146] (and thus in a representative sample of the language in general) and then average the frequencies. The correlations calculated for these measures suggest that the use of frequent vocabulary is negatively related with the holistic grade. Here, predictive models might also run into problems with multicollinearity, because the measures differ only in the reference corpus used but are expected to behave similar as they operationalize more or less the same concept.

Similar to the above word frequency or elaborateness measures, measures 27-30 and 32-35 try to operationalize the typical **age of acquisition** of a word by observing if (and at what age) the words used in the text occur in a reference corpus of child language (Karlsruhe Children Text). The observed correlations say that the words used in texts with better holistic grades occur in later ages in the reference corpus, additionally better rated texts have less words that occur in this child language reference corpus. This is another clear sign for a relationship between the elaborateness of the vocabulary and the holistic grades.

The measures on rank 18 and 22-25 are all measures of the **syntactic variation**. They show how varied the text construction is, how many types of periphrastic or non-periphrastic verb tenses are used (18 and 23)[147], how many different de-agentivation patterns are used to make the text sound more objective (22) and how many different types of noun modifiers are used to enrich the language (24). All of these measures are positively correlated, indicating that texts with higher holistic grades use more

[144] See also the non-corrected simple type token ratio at rank 27.

[145] In this data, however, there is very little evidence for objects or subjects that were taken up in the next sentence, so that no significant positive correlations could be found for the reversed features (e.g. probObjSubjPerTransition, probSubjObjPerTransition).

[146] The used reference corpora were DLEX, Subtlex, Google 00, and KCT as described in section 9.2.3.2.

[147] The use of various periphrastic and non-periphrastic verb tenses can thereby indicate the use of active and passive constructions and nominal style, which are considered as signs of academic language (cf. Weiß, 2017).

different structures to make the language more varied. However, the verb cluster size (a number of verbs that fall together to construct a predicate, e.g. “das Argument, das gesagt werden hat müssen”) varies more in text with lower text quality rating (indicated by the negative correlation between holistic grades and the standard deviation for the size of verb clusters).

Summary
In total, the highest correlations based on monofactorial Spearman rank correlation for the linguistic complexity measures suggest lexical diversity and elaborateness of the used lexicon, text length, syntactic variation as well as missing cohesion as important factors for the assigned holistic grades. However, as in the previous analysis, it is not clear whether the combination of these features is indeed the best way to explain the holistic grades, nor which one of the inter-correlated groups of features to choose.

12.2.2 Limits of intrinsically interpretable models

If we want to do multifactorial analysis with the data, we need to refer to more advanced analysis methods, as for example predictive modelling strategies, and evaluate the existence as well as the type of relationships in the data by interpreting models that are good enough to generalize from the data.

The interpretation of predictive models, however, usually depends on the typology of the model used. While some model typologies are considered to be intrinsically interpretable and have been used for data analysis for many decades, others are often called black-box models as their interpretation is not straightforward.

In this section I discuss some methodological limits of intrinsically interpretable models like regression models, decision tree models, rule induction models or Naïve Bayes Classifiers. In particular I discuss the conflict between performance, complexity and interpretability from a machine learning perspective as well as methodological restrictions for regression models from a statistical perspective.

12.2.2.1 Performance – Complexity – Interpretability

In the classification experiments in section 10 we saw that the predictive performance of a model also depends on the typology or algorithm of the model trained. Often, the simpler, intrinsically interpretable model typologies, such as regression models, decision trees, rule learners or Naïve Bayes classifiers (cf. section 9) had a lower predictive performance when compared to the more complex, so-called black box models. This was especially the case for complex model setups, were many, probably irrelevant features were included.

Indeed, for a classification of text quality ratings with the complex features set 3 containing automatically extracted features of linguistic complexity, the predictive performance of simpler, intrinsically interpretable models is not sufficiently high to assume non-random predictions (see Table 42). Of all the tested models, only the two of the black box models, random forest and the neural network Multilayer Perceptron, achieved performances that were significantly above the baseline (p-value < 0.001, paired *t*-test).

		Intrinsically interpretable				Black box models		
Dataset	Baseline	NB	Logreg	PART	J48	SMO	RandF	MLP
complexity	38.75	28.95	41.14	42.70	41.33	42.64	**47.06***	**45.09***

Table 42: Comparison of classifier accuracies for intrinsically interpretable models and black box models.

When we interpreted the models in section 10.2.3, a reduced feature set was used to demonstrate interpretation possibilities while keeping the interpretation simple. However, if we would have used a larger feature set, the interpretation might have become difficult, because the actual interpretability of allegedly interpretable model typologies is often proportional to the complexity of the task. Using a higher amount of predictor variables usually leads to indigestible models as trees and rule sets grow bigger, probabilities are given for various competing classes, collinear regression coefficients compete against each other (cf. also 10.2.3.3).

Figure 51 and Figure 52 show the increase in tree complexity (measured in tree size, i.e. total number of nodes, and number of leaves) when the number of features used for training increases. Figure 51 displays the reported tree complexity measures for the categorical feature set 1 (text analysis questionnaire) for J48 decision tree classifiers trained with the WEKA standard configuration with different amounts of features. Analogously, Figure 52 displays the values reported for the numerical feature set 3 (linguistic complexity for different amounts of features[148].

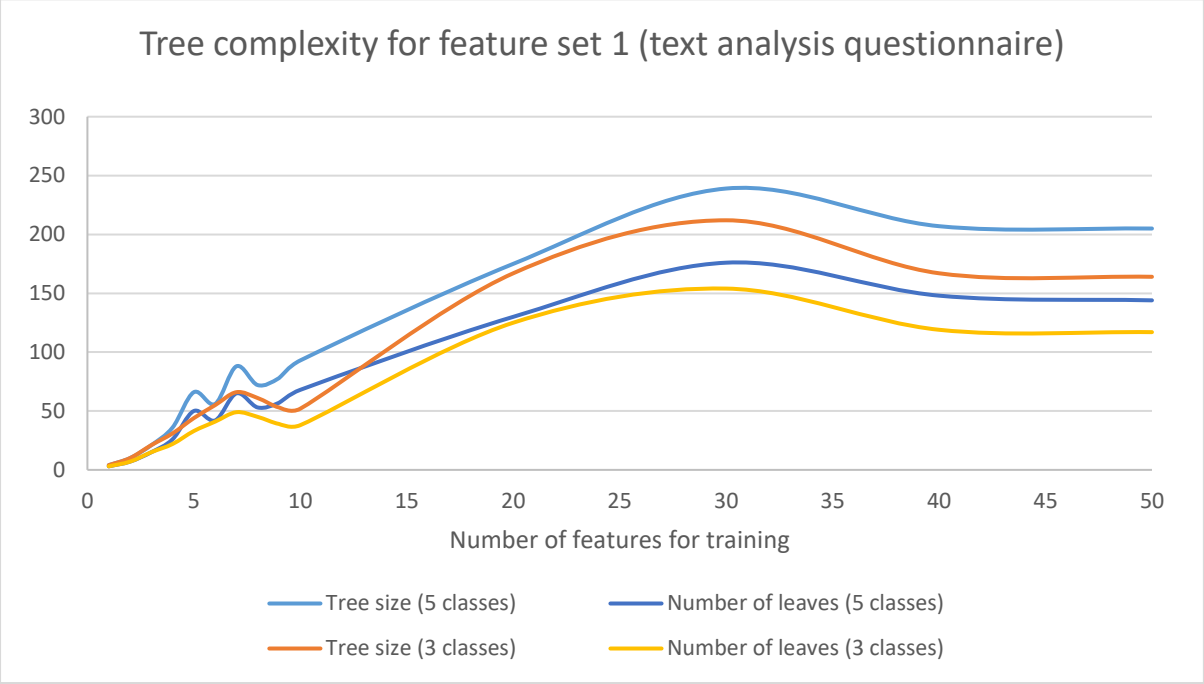


Figure 51: Tree complexity for predicting 3 or 5 grade levels on the categorical text analysis questionnaire data.

[148] In both cases, ReliefF feature ranking was used to reduce the feature set according to the importance of the features.

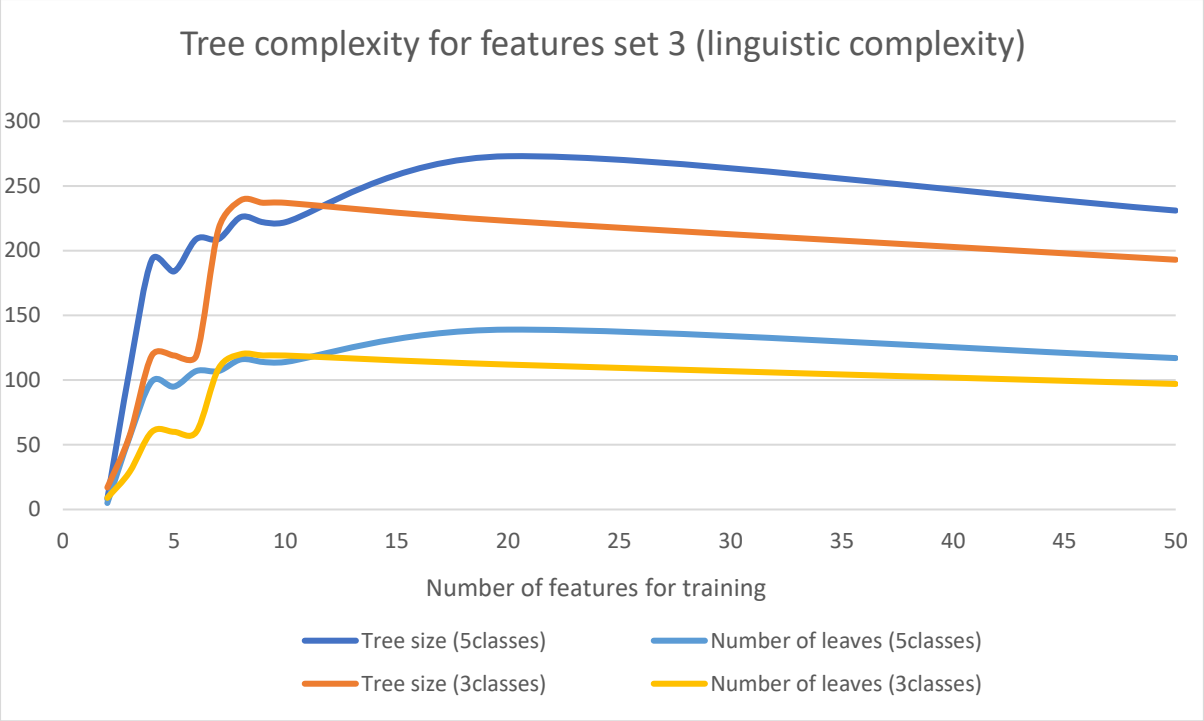


Figure 52: Tree complexity for predicting 3 or 5 grade levels on the numerical linguistic complexity feature set.

The graphs show that the complexity of the tree increases very fast with just a few added features. While the categorical feature set 1 has a slightly lower-paced complexity increase, the tree complexity for the numerical values of feature set 3 is immediately very high. This is due to the various possibilities to split the subset into further subsets that is limited to the maximum number of categories with the categorical features but can be infinite for continuous numerical features[149].

However, while the complexity of the trees trained with feature set 1 (left graph of Figure 53) is reaching sufficiently high values for the prediction of five grade levels as well as for the prediction of three grade levels even for trees trained with less than ten features, the accuracy curve for feature set 3 displayed on the right shows that the classifiers for predicting five grade levels only reach the baseline accuracy when trained with more than 50 features (and might still not be significantly above the baseline).

[149] Both graphs show that the complexity does not increase after a certain number of features. The reason for this is that the used stopping criterion prohibits the tree from growing any further.

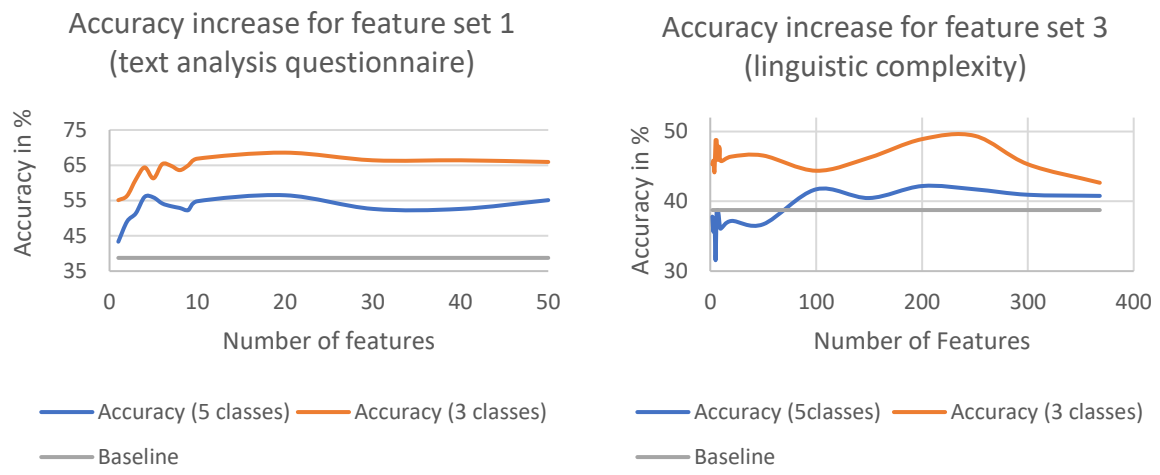


Figure 53: Accuracy increase with number of features for predicting 3 or 5 grade levels using feature set 1 and feature set 3.

Furthermore, the decision tree classifiers do not necessarily gain from the additional features, as accuracy results do not increase after a certain number of features as can be seen also in Figure 54 below. In contrast, the random forest classifier for the prediction of five grade levels, does not show this weakness and already performs above the baseline with only very few features.

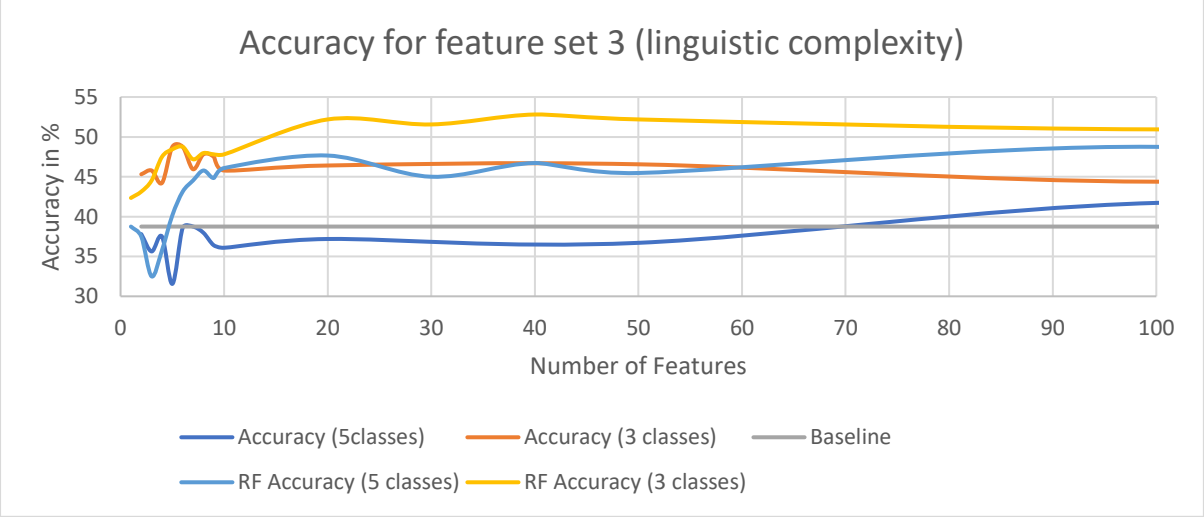


Figure 54: Accuracy increase with the number of features for predicting 3 or 5 grade levels with feature set 3 comparing decision tree and random forest classifiers.

12.2.2.2 Regression modelling and difficulties in model selection

In the correlation analysis we could observe possible confounding and multicollinearity introduced by natural dependencies between the features. One example for such dependencies are lexical diversity features. Lexical diversity is supposed to measure how varied the choice of words and lexical items (the vocabulary) of the text is. Naturally, the longer the text, the higher is also the chance to use a more diverse vocabulary. Therefore, operationalizations for lexical diversity are often biased by the text length (Francis & Kucera, 1967) and different measures have been developed over the last decades to account for this natural bias (see e.g. Bonvin & Lambelet, 2017; McCarthy, 2005).

The automatically extracted linguistic complexity features in feature set 3 contain seven different lexical diversity measures:

- TTR
  The type token ratio, one of the simplest but also most text length-dependent measures for lexical diversity.
- rootTTR
  The root transformed type token ratio that is supposed to be less dependent on text length.
- bilogarithmicTTR
  The bilogarithmically transformed type token ratio, also designed to be less dependent on text length than the simple TTR.
- UberIndex
  The Uber index is yet another transformation of the TTR (cf. Jarvis, 2002)
- YulesK
  Yules K, is a lexical diversity measure that bases on rank frequencies of the words in a text (cf. Bonvin & Lambelet, 2017)
- MTLD
  The Measure for Textual Lexical Diversity, proposed by McCarthy (2005) iteratively recalculates and refines the TTR for every word using an algorithmic strategy to create frames for the calculation.
- HDD
  Proposed by McCarthy and Jarvis (2007) the HD-D measure is calculated based on probabilities of occurrence of each type in a text in subsamples of the same text.

However, various studies point out that most measures, especially the derivatives of the TTR are still not independent from the text length (Bonvin & Lambelet, 2017; Durán et al., 2004; McCarthy, 2005; McCarthy & Jarvis, 2010).

Table 43 below shows a Pearson-Product-Moment correlation matrix for these features including holistic grades and the text length features nTokens (number of tokens in the text) and nSentences (number of sentences in the text). Only the Measure for Textual Lexical Diversity MTLD is not significantly related with the text length at alpha 0.01[150] (r = 0.08, p-value = 0.037). This also corresponds to results from other studies that investigated the text length dependency for lexical diversity measures (McCarthy & Jarvis, 2010).

	grade	nTokens	nSent.	MTLD	yulesK	HDD	Uber Index	TTR	Root TTR	Bilog TTR
grade	1,000									
nTokens	0,381	1,000								
nSentences	0,329	0,895	1,000							
MTLD	0,113	0,083	0,083	1,000						
yulesK	-0,097	-0,173	-0,195	-0,750	1,000					
HDD	0,132	0,207	0,226	0,830	-0,972	1,000				
uberIndex	-0,039	-0,231	-0,184	0,766	-0,540	0,646	1,000			
TTR	-0,271	-0,756	-0,663	0,413	-0,218	0,263	0,777	1,000		
rootTTR	0,396	0,790	0,730	0,545	-0,486	0,587	0,297	-0,319	1,000	
bilogTTR	-0,175	-0,587	-0,504	0,579	-0,362	0,437	0,894	0,966	-0,065	1,000

Table 43: Correlation matrix for lexical diversity measures and text length measures.

[150] However, the p-value is at 0.037 quite high as well.

However, the MTLD measure has only a very low correlation with the holistic grade (r = 0.11, p-value = 0.004), compared to for example the root-transformed type token ratio (r = 0.40, p-value < 0.001). This raises doubt if lexical diversity is indeed a good predictor variable for text quality grades.

A number of regression models is investigated to evaluate if we can detect an effect of lexical diversity on the holistic grades despite the one already observed for text length. In particular, I compare the following regression models:

1. Grade ~ text length (nTokens)
2. Grade ~ text length (nTokens) + rootTTR (lexical diversity measure with the highest correlation with the holistic grade)
3. Grade ~ text length (nTokens) + MTLD (lexical diversity measure that is not correlated with the text length
4. Stepwise backward model selection choosing from the full scope of lexical diversity measures and the text length

For each model I then evaluate the overall model significance, the R^2 and adjusted R^2 values as a measure of model performance, the significance of positive or negative variable effects in the model, as well as the variance inflation factor for the model that gives an estimate on multicollinearity. The systematic comparison of those models shall demonstrate the problem with violating the assumption of independence in analysis and with multicollinearity in stepwise model selection.

Holistic grade ~ Text length

The first model shows the results for a linear regression for the text length (nTokens). The model is significant and shows a significant positive effect of the text length on the holistic grade, explaining 14.5% of the variation in the dependent variable (adjusted $R^2 = 0.1437$, multiple $R^2 = 0.1451$) (cf. Model output 7).

```
Call:
lm(formula = grade_numeric ~ F256_nTokens, data = complexity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.76834 -0.70716  0.04258  0.58262  2.41380

Coefficients:
                    Estimate Std. Error   value Pr(>|t|)
(Intercept)         2.0827519  0.0929093   22.42   <2e-16 ***
F256_nTokens        0.0016346  0.0001571   10.40   <2e-16 ***

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9134 on 638 degrees of freedom
Multiple R-squared:  0.1451,  Adjusted R-squared:  0.1437
F-statistic: 108.2 on 1 and 638 DF, p-value: < 2.2e-16
```

Model output 7: Linear regression model for grade ~ text length in tokens.

Holistic grade ~ Text length + rootTTR

The second model includes the lexical diversity measure with the best correlation with the holistic grade, but the highest correlation with the text length to the model (see Model output 8). This model is also significant but has a higher multiple R^2 (0. 1694) and adjusted R^2 (0. 1668), indicating better performance. Both features in this model have regression coefficients that are significantly different from 0, indicating that they both individually contribute to the model. The `rootTTR` feature has a

higher t-value and is therefore more important for the model than the text length feature. The variance inflation factor for this model is at 2.66 already rather high, indicating the there is some multicollinearity that might cause regression coefficients to be unstable.

```
Call:
lm(formula = grade_numeric ~ F256_nTokens + F117_rootTTR,
    data = complexity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.43055  -0.69202  0.02904  0.62760  2.52696

Coefficients:
               Estimate   Std. Error   value  Pr(>|t|)
(Intercept)    0.4717708    0.3839445   1.229   0.21962
F256_nTokens   0.0007726    0.0002526   3.058   0.00232 **
F117_rootTTR   0.1758303    0.0406942   4.321   1.8e-05 ***

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.901 on 637 degrees of freedom
Multiple R-squared:  0.1694,   Adjusted R-squared:  0.1668
F-statistic: 64.96 on 2 and 637 DF, p-value: < 2.2e-16
```

Model output 8: Linear regression model for grade ~ text length in tokens and root type token ratio.

Holistic grade ~ Text length + MTLD

Third, I compare the previous model with a model with text length and MTLD as predictors (Model output 9), where MTLD was not significantly correlated with text length in the bivariate correlation analysis. The model reports significance for both features, but the overall model performance is lower than in the previous model. Besides, the model puts much more importance on the text length feature than on the lexical diversity feature contrary to the previous model (t-values 10,2 vs. 2.2).

```
Call:
lm(formula = grade_numeric ~ F256_nTokens + F123_MTLD,
    data = complexity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.81458 -0.68208  0.04188  0.59458  2.51831

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.7124706  0.1890976   9.056   <2e-16 ***
F256_nTokens   0.0016054  0.0001571  10.216   <2e-16 ***
F123_MTLD      0.0028546  0.0012710   2.246    0.025 *

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9105 on 637 degrees of freedom
Multiple R-squared: 0.1518,    Adjusted R-squared: 0.1491
F-statistic: 56.99 on 2 and 637 DF, p-value: < 2.2e-16
```

Model output 9: Linear regression model for grade ~ text length in tokens and Measure for Textual Lexical Diversity (MTLD).

Holistic grade ~ Text length + TTR + rootTTR + bilogarithmicTTR + MTLD + yulesK + uberIndex + HDD (stepwise backward model selection)

Lastly, I use a stepwise, backward approach for model selection using all seven lexical diversity features as candidate features. I start with a full model with all seven features plus the text length feature. The model itself is significant, which means that it did learn from the data. However, for none of the predictor variables we see significant regression coefficients (see Model output 10).

```
Call:
lm(formula = grade_numeric ~ F256_nTokens + F123_MTLD + F121_yulesK +
    F122_HDD + F120_uberIndex + F116_TTR + F117_rootTTR +
    F119_bilogarithmicTTR, data = complexity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25490 -0.66628  0.00616  0.66958  2.49073

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)          -5.122e+01  4.793e+01   -1.069    0.286
F256_nTokens         -1.241e-04  5.881e-04   -0.211    0.833
F123_MTLD             1.767e-03  2.886e-03    0.612    0.540
F121_yulesK           7.269e-03  1.738e-02    0.418    0.676
F122_HDD              3.896e-02  4.054e-01    0.096    0.923
F120_uberIndex       -2.116e-02  2.224e-02   -0.951    0.342
F116_TTR             -1.710e+01  1.742e+01   -0.982    0.327
F117_rootTTR          1.268e-01  2.454e-01    0.517    0.605
F119_bilogarithmicTTR  6.789e+01  6.898e+01    0.984    0.325


Residual standard error: 0.8948 on 631 degrees of freedom
Multiple R-squared:  0.1884,   Adjusted R-squared:  0.1782
F-statistic: 18.32 on 8 and 631 DF, p-value: < 2.2e-16
```

Model output 10: Linear regression model for grade ~ text length in tokens and all lexical diversity measures.

The variance inflation factor reported for the variables signal the high multicollinearity (see Table 44). We can therefore assume that the multicollinearity in the model hinders the calculation of correct and reliable estimates for the coefficients.

Variable	VIF
F256_nTokens	14.598
F123_MTLD	5.374
F121_yulesK	39.667
F122_HDD	61.659
F120_uberIndex	21.587
F116_TTR	977.030
F117_rootTTR	97.929
F119_bilogarithmicTTR	830.082

Table 44: Variance inflation factors for lexical diversity measures and text length in tokens.

After iteratively removing the non-significant feature with the highest p-value for manual backward model selection, we remain with a final model that now only contains the non-adjusted, heavily text length-dependent type token ratio feature (Model output 11). The text length feature (nTokens) has been removed in one of the steps, so has the non-related MTLD feature and also the rootTTR feature with the highest bivariate correlation coefficient. The model is significant but only explains 7% of the variance in the dependent variable (cf. adjusted and multiple R^2), so about half of what we achieved with only the text length feature. It is clear that the stepwise model selection approach failed completely in this experiment.

```
Call:
lm(formula = grade_numeric ~ F116_TTR, data = complexity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47814 -0.79717 -0.01168  0.77860  2.33605

Coefficients:
            stimate  Std. Error  t value   Pr(>|t|)
(Intercept)  5.1950      0.3143   16.528    < 2e-16 ***
F116_TTR    -4.2154      0.5921   -7.119   2.94e-12 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9508 on 638 degrees of freedom
Multiple R-squared:  0.07359, Adjusted R-squared:  0.07214
F-statistic: 50.68 on 1 and 638 DF, p-value: 2.936e-12
```

Model output 11: Linear regression model for grade ~ non-corrected type token ratio.

In sum, we can say that the models built did not give a clear answer on the individual contribution of text length and lexical diversity. When we operationalize lexical diversity with rootTTR the model performs better, but multicollinearity is rather high. When we operationalize lexical diversity with the probably least intercorrelated MTLD feature, the model performance goes down. However, both predictors are significant in both cases, so that we can assume that lexical diversity and text length are both informative for text quality grades and add individually to its explanation. Which one is responsible for how much of the shared contribution, is, however, impossible to determine.

The observed problems regarding multicollinearity, model selection and model interpretation are not exclusive to lexical diversity measures in the linguistic complexity feature set but might hinder the analysis of other aspects with various operationalizations as well. It is thus impossible to fully trust in stepwise model selection procedures in case of multicollinearity, especially when many features are involved. Moreover, manual stepwise regression modelling becomes increasingly difficult when many features are involved. As regression models only account for individual effects of features on the dependent variable, combinatorial effects have to be added manually as interaction terms. However, adding all possible interaction terms (combinations of features that could interact with each other) soon results in overly complex model structures that need a lot of computation time and are difficult to interpret for humans. Besides, using categorical features instead of continuous features, or using classification instead of regression and the resulting explosion in output values for one-hot-encoded feature coefficients or class probabilities further increases the complexity of the model, when aiming at manual model selection.

Finally, for the sake of illustration the previous experiments were performed with an ordinary multiple linear regression. However, as the data has nested observations (i.e. various data points belong to the same annotator), we should use appropriate models for hierarchical data like mixed-effects models (cf. section 10.2.4.3). The model selection procedures for these models are, however, even more difficult and the observed effects even more complex to interpret (cf. Speelman et al., 2018).

Summary

The interpretation of predictive models trained on the complex feature set of linguistic complexity measures is limited when referring to the inspection of intrinsically interpretable models. First the prediction performance for more complex black box models is significantly higher than the one for intrinsically interpretable models, indicating that the simpler models miss out on important information that is actually present in the data. Second, the interpretability of intrinsically interpretable

models is impeded as model complexity rises with the number of features and model selection techniques can fail under the presence of multicollinearity. Hence, intrinsically interpretable models might lack the necessary performance and/or the actual interpretability in order to gain interesting new insights from the data.

12.2.3 Black box interpretation

In recent years, a number of approaches and methods for black box interpretation have been proposed as tools to make sense of complex predictive models (see also section 3.6). These can be used when the model internals are not inherently interpretable (e.g. black box models like random forests or neural networks) or when the built model is in general too complex to interpret. There are many model-specific as well as model-agnostic techniques to interpret models as a black box. However, this section refers, as an example, to model-agnostic interpretation methods that can be used to interpret intrinsically interpretable models as well as black box models.

12.2.3.1 Classifier comparison and feature engineering

One frequently observable strategy is to compare the predictive performance of different models that were trained on systematically chosen feature sets. In order to interpret the effect of the linguistic complexity measures for predicting the holistic grades, I therefore split the full set of linguistic complexity measures into measures regarding lexical, syntactical and text complexity as well as cohesion measures and test how well they perform individually as predictors of the holistic grade. Table 45 shows that the best prediction results were achieved with the syntactic complexity features and a random forest classifier. Although the monofactorial analysis of correlations between the features and the holistic grade reported more lexical features among the highest correlated features, the multifactorial prediction approach did learn more from the syntactic features. However, the difference between the prediction results for syntactic features, lexical features and text features was not significant. Indeed, lexical complexity features and text complexity features also achieved classification results that were significantly above the majority baseline. The cohesion measures appear to be the least informative feature when it comes to holistic grades assigned to the student essays.

The best performance, however, was achieved with the random forest classifier trained on text complexity as well as lexical and syntactic complexity features. The performance for this classifier was even higher than the one with all complexity features (including the cohesion measures).

Dataset	Baseline	NB	Logreg	PART	J48	SVM	RandF	MLP
cohesion	37.92	15.63	37.34	37.59	37.11	40.47	43.56	36.11
text_complexity	37.92	20.27	38.31	38.42	38.41	39.80	**45.20***	35.56
lex_complexity	37.92	26.81	39.09	40.39	40.03	42.00	**45.39***	42.77
syn_complexity	37.92	28.28	34.13	39.69	38.50	39.86	**45.97***	41.98
Ablation test								
complexity	37.92	28.95	41.14	42.70	41.33	42.64	**47.06***	**45.09***
no_cohesion	37.92	33.67	40.23	42.02	41.59	42.81	**47.11***	44.36
no_cohesion_no_text	37.92	33.38	39.34	41.77	42.05	43.06	**46.77***	**44.45***

Table 45: Classifier comparison and ablation test for different categories of linguistic complexity measures.

12.2.3.2 Feature selection and feature ranking

The rank reported by the ReliefF algorithm states the following 15 most important features (Table 46).

Feature	ReliefF score
annotator	0.0217388
coverageModifierTypes	0.0199672
rootTTR	0.0199416
nSentences	0.0166510
correctedNonAuxVerbTypesPerNonAuxVerb	0.0163167
nTokens	0.0161786
squaredNonAuxVerbTypesPerNonAuxVerb	0.0155279
lexTypesNotFoundInKCTPerLexicalType	0.0121755
lexLemmasNotFoundInKCTPerLexicalLemma	0.0115628
lemmasFoundInKCTPerLexicalLemma	0.0115628
typesFoundInSubtlexPerLexicalType	0.0115353
probSubNotsPerTransition	0.0115195
coveragePeriphrasticTenses	0.0114943
typesFoundInKCTPerLexicalType	0.0113792
lexicalTypesPerLexicalToken	0.0112065

Table 46: ReliefF scores for 15 most important features in linguistic complexity feature set.

The ReliefF algorithm reported the control variable annotator as the most important feature in total, meaning that according to this feature selection approach, knowing the annotator contributes most to explaining the holistic grade.

The most important complexity measures reported by the algorithm are (similarly to the correlation analysis before) features related to text complexity, i.e. the text length (number of sentences, number of tokens), and lexical complexity, i.e. lexical and in particular also verb diversity (rootTTR, lexicalTypesPerLexicalToken, correctedNonAuxVerbTypesPerNonAuxVerb, squaredNonAuxVerbTypesPerNonAuxVerb) as well as features related to lexical sophistication (Types and Lemmas found or not found in the reference corpus for child language, types found in a reference corpus of TV subtitles). The coverage of modifier types and the coverage of periphrastic tenses, both syntactic complexity-related variables indicating variability in the text, are also important according to the ReliefF algorithm (these measures were among the ones with the highest Spearman rank correlation coefficient as well). Furthermore, there was one text cohesion measure, among the 15 best-rated features (the probability that the subject of the sentence is not referenced in the next sentence).

12.2.3.3 Interpreting model internals: Feature importance, effects and interactions for complex models

Finally, the creation and interpretation of complex, non-intrinsically interpretable models is attempted by using interpretation methods from interpretable machine learning and explainable artificial intelligence. I train a number of models on the feature set 3 using random forests, as well as a linear and a non-linear neural network architecture. Table 47 shows the results for the best models.

Model setup	Dataset baseline	Test set baseline[151]	Accuracy of best model
Linear neural network	*37.92%*	40.3%	53.5%

[151] The baseline for the test set is higher, as the model was not evaluated on the full dataset (as in cross-validation) but on a test-trainset split of 0.2.

Non-linear neural network	*37.92%*	42.6%	57.4%
Random forest	*37.92%*	41.1%	48.1%

Table 47: Baselines and accuracy of best model for three complex model architectures.

The two random forest models built with different implementations (once using the WEKA data mining software and its standard configuration for random forest classification and once using the `RandomForestRegressor` implementation of the `scikit-learn` machine learning library for Python, evaluated with capped and rounded predictions) differ only little in terms of predictive performance. However, the situation is not the same for the neural network models. The accuracy for the best Python neural network is significantly higher than the one achieved with the WEKA standard configurations (45.09%) and also significantly higher than the one for random forest (47.1% in WEKA, 48.1% in `scikit-learn`). It is, however, important to note that these results refer to individual networks that have been trained with randomly initialized feature weights. In order to account for the non-deterministic nature of neural networks (induced by these randomly chosen initial weights), a series of differently initialized networks has been trained and evaluated. The average accuracies of the randomly initialized neural networks are not higher than the baseline for the train set. Better results could only be found in some of the networks. This is probably due to the small data size. The models highly depend on well initialized weights, while the actual "learning" of the neural network, i.e. updating the weights with the loss and activation function, can hardly take place. However, there were various instances of networks that did learn within the few iterations and achieved good accuracy results on a test set of 20% of the data (where the remaining 80% have been used for training).

I also trained network instances for the same model setups on feature set 1 (text analysis questionnaire) and feature set 1 and 3 together (see Table 48). The best model in absolute was the non-linear neural network with one hidden layer of three neurons, trained on 80% of the data using both text analysis features and linguistic complexity features. The model achieved an overall accuracy of 76% on a test set that would yield a 34% accuracy if the classifier would only assign the majority class.

	Feature set 1: Text analysis		***Feature set 1 + 3***[152]	
Model setup	**Test set baseline**	**Accuracy of best model**	**Test set baseline**	**Accuracy of best model**
Linear neural network	*37.2%*	*68.2%*	*37.2%*	*71.3%*
Non-linear neural network	*37.2%*	*72.1%*	*34.1%*	*76.0%*
Random forest	*37.2%*	*68.2%*	*34.1%*	*64.9%*

Table 48: Comparison when using feature set 1 or combining feature set 1 and 3.

Apart from the comparison of accuracies for different feature sets and the investigation of monofactorial feature selection criteria as shown in the previous subsection, these models can be investigated using recently developed interpretation techniques and tools for the calculation and visualisation of variable importances, variable effects and interactions. For this case study, the Python library SHAP is used exemplarily to attempt an interpretation of the best models.

Global model interpretation

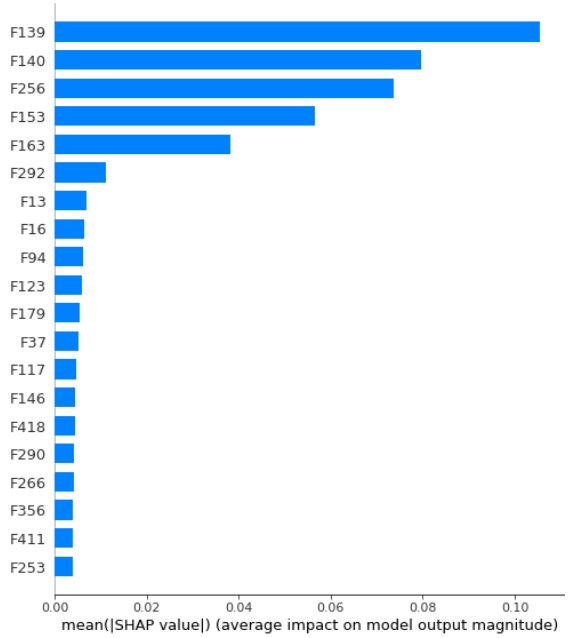
The `SHAP` library for Python provides an infrastructure to calculate Shapley values (indicating local feature effects) for the model and visualize them for local and global model interpretation. The first

[152] The continuous variables of feature set 3 have been normalized for the combined feature set in order to account for the big differences in variance.

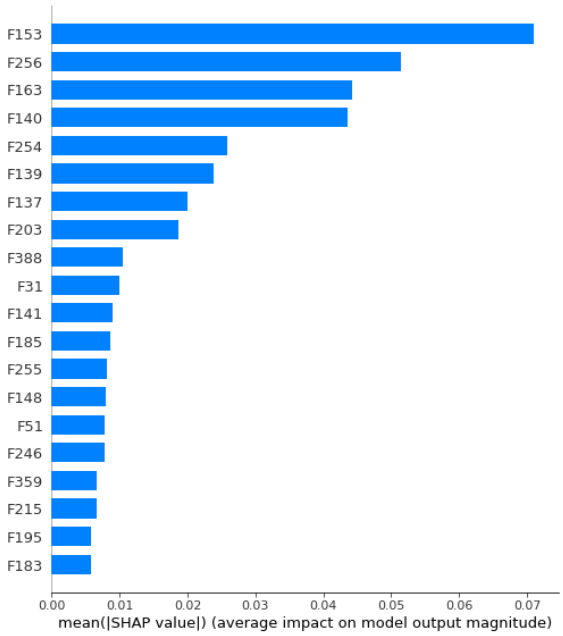
three graphs show the global variable importance (measured in the mean Shapley values) for the best model of each of the model types (Figure 55). Each of the models has a few variables that are visibly more important than the rest. However, which of the complexity measures was among the most important variables depends on the model type. While the difference between the important variables and the less important variables is very clear cut in the linear network and the random forest model, it is less marked for the non-linear network. The random forest model bases its predictions mainly on the number of tokens in the text and on the (also length-dependent) lexical diversity measures *root type token ratio* and *corrected type token ratio*. The network models on the other hand chose measures for lexical sophistication (frequency of the words from the essay according to reference corpora Dlex, Subtlex and Google 2000) next to the text length measure. Figure 56 shows the respective numerical values.

Feature importance

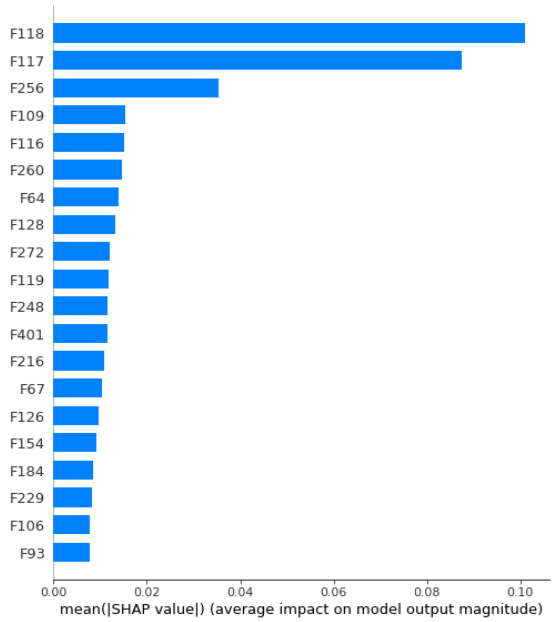
Linear Neural Network



Non-linear Neural Network



Random Forest



Legend

F64: complexTUnitsPerTUnit
F109: NpModifiersPerNP
F116: TTR
F117: rootTTR
F118: correctedTTR
F137: adverbsPerLexicalToken
F139: annotatedTypeFreqsPerTypeFoundInDlex
F140: typeFreqsPerTypeFoundInDlex
F153: typeFreqsPerTypeFoundInSubtlex
F163: typeFreqsPerTypeFoundInGoogle00
F203: -istTPerToken
F254: nSentences
F256: nTokens
F260: probSubNotsPerTransition
F292: 1PPersPronounsPerTokenInSentencePerSentence

Figure 55: Shapley values for 20 most important features for the linear neural network, the non-linear neural network and the random forest model.

Linear Network

column_name	shap_importance
F139	0.105595
F140	0.0795317
F256	0.0736628
F153	0.0565047
F163	0.0380553
F292	0.0111975
F13	0.00683454
F16	0.00627923
F94	0.0060814
F123	0.00596072
F179	0.00528389
F37	0.00522408
F117	0.00465401
F146	0.00437475
F418	0.00428643
F290	0.0041163
F266	0.00406551
F356	0.0039663
F411	0.00394273
F253	0.00380663

Non-linear Network

column_name	shap_importance
F153	0.0710676
F256	0.0513975
F163	0.0441878
F140	0.0435277
F254	0.0258539
F139	0.0238157
F137	0.019988
F203	0.0185984
F388	0.0105205
F31	0.00992562
F141	0.00892366
F185	0.00868443
F255	0.00821193
F148	0.00802651
F51	0.00781384
F246	0.00776522
F359	0.00659219
F215	0.00658452
F195	0.00585803
F183	0.0058125

Random Forest

column_name	shap_importance
F118	0.101109
F117	0.087393
F256	0.0353473
F109	0.0154974
F116	0.0150743
F260	0.0147633
F64	0.0140139
F128	0.0132211
F272	0.0121286
F119	0.0117817
F248	0.0116136
F401	0.0116075
F216	0.0108369
F67	0.0103797
F126	0.00973328
F154	0.00915427
F184	0.00855354
F229	0.00823399
F106	0.00779789
F93	0.00778121

Figure 56: Ordered list of Shapley values for the 20 most important features for all three models.

Feature effects

We can inspect the feature effects with global summary plots, and partial dependence plots for individual main effects.

Global summary plots

Figure 57 and Figure 58 summarize the variable effects for all the predictions for the 20 best variables of each model. For each line all predictions are illustrated as coloured dots, depending on the observed value for this variable (high values are pink, low values are blue). The dots are then distributed on the x-axis, depending on how much the individual prediction was influenced by the variable. The linearity, uniformity and magnitude of the effects can be interpreted by the spread of the dots and the distribution of colours over the span of the x-axis. Pink dots that are located mainly on the right-hand side and blue dots that are located mainly on the left-hand side are indicative of a positive relation.

The two neural network models show that many predictions are centred at the middle of the x-axis, indicating that there is no effect of the variable on the prediction. Especially for the less important variables, the effects often concern only individual predictions. For example, in the graph for the linear neural network below we see a negative rather uniform relationship of the `F139` variable (the lexical sophistication measure `annotatedTypeFreqsPerTypeFoundInDlex`). While the variable influences the predictions in both directions, the Shapley values for the variable `F140` (a variant of the previous variable, the `typeFreqsPerTypeFoundInDlex`) show almost exclusively negative effects. Comparing the two graphs for the linear and non-linear neural network model, we can also observe that the reported effects for various variables is different between the two models. Variable `F153` (the lexical sophistication measure `typeFreqsPerTypeFoundInSubtlex`), for example, is negatively related with the text quality judgments in the linear model and positively in the non-linear model. This aspect is shown below in the local model interpretation graphs (Figure 57).

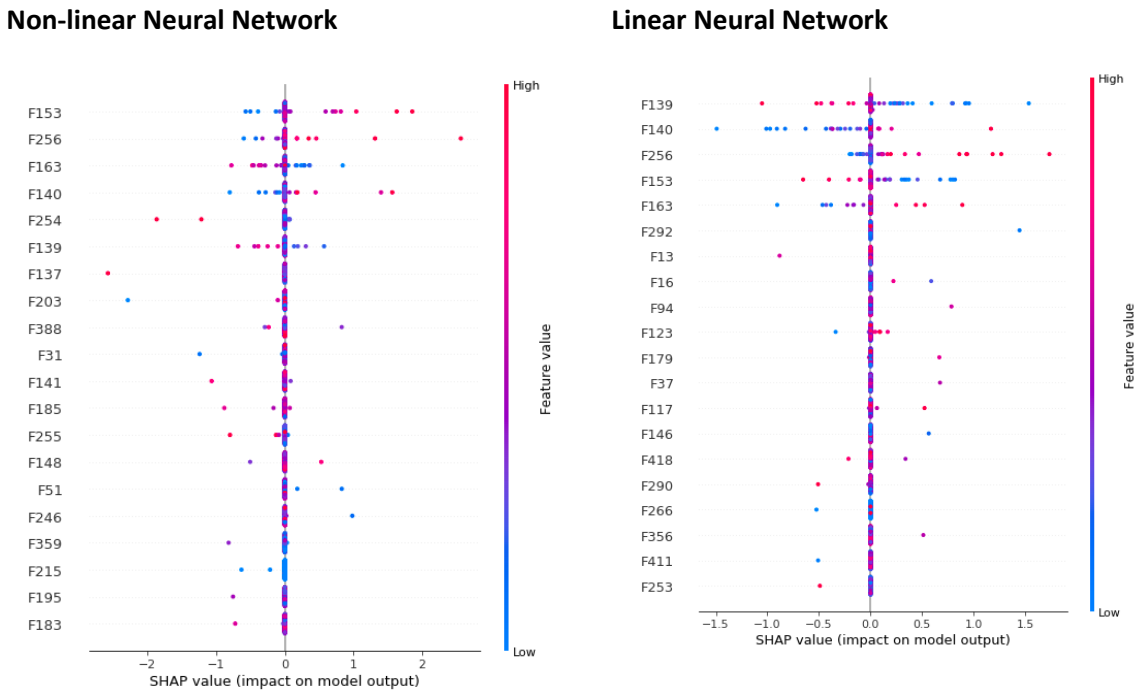


Figure 57: SHAP summary plots showing the Shapley importance for each data point for each of the 20 most important variables for the linear and non-linear neural network models trained with the linguistic complexity feature set.

The random forest model on the other hand shows effects for all or almost all predictions, variables without effect (Shapley value of 0) barely exist for the 20 best variables. Effects for less relevant variables cluster at a certain low Shapley value, either with few exceptions to one or both sides of the cluster (e.g. F109_NpModifiersPerNP, F116_TTR) or with two clusters on both effect directions that can also be completely separated from each other in case there is a non-linear variable effect that changes abruptly with only little change in the observed variable value (e.g. F64_complexTUnitsPerTUnit, F117_correctedTTR).

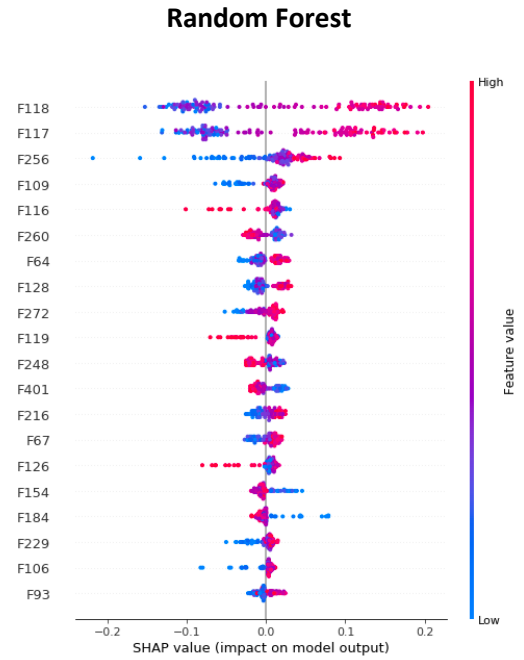
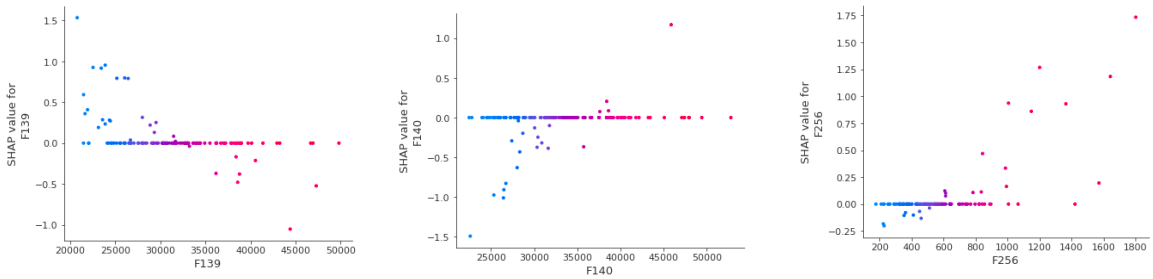


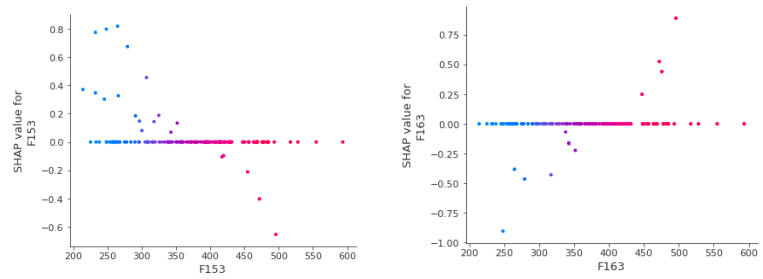
Figure 58: SHAP summary plots showing the Shapley importance for each data point for each of the 20 most important variables for random forest model trained with the linguistic complexity feature set.

Main effects for best features (Partial dependence plots)

For individual main effects, the SHAP library allows to print two-dimensional partial dependence plots for main effects (based on the Shapley values for the predictions). Figure 59 to Figure 61 show the main effects for the best variables for each model.

Linear Neural Network



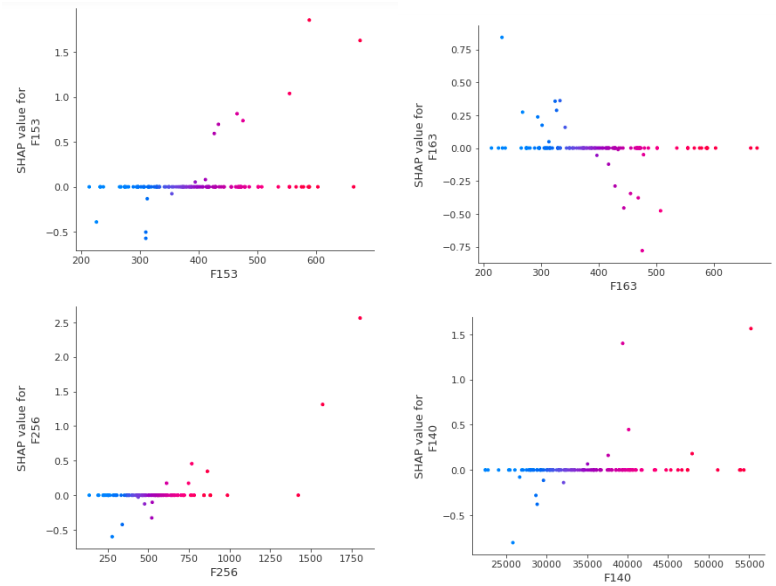


F139: annotatedTypeFreqsPerTypeFoundInDlex
F256: nTokens
F140: typeFreqsPerTypeFoundInDlex
F153: typeFreqsPerTypeFoundInSubtlex
F163: typeFreqsPerTypeFoundInGoogle00

Figure 59: Partial dependence plots for most important features of the linear neural network.

For the linear neural network, five variables were visibly more important than the rest of the variables. Four of these variables are lexical sophistication measures, measured in the frequency of the occurring words according to a reference corpus (Dlex, Subtlex or Google00). The 3rd best variable (F256) is the text length. The graphs above show the relationship between the variables and the predictions of the model. The effect of the text length variable is relatively clear (albeit only on a limited amount of predictions). The longer the text, the better is the predicted holistic grade. For the lexical sophistication features we would expect a negative relation. Better texts are expected to have less frequent vocabulary. However, we observe this relationship only for the Subtlex corpus and a (corrected) version of the measure on the Dlex corpus, while the unmodified Dlex measure and the measure for the Google00 corpus show a contradictory relation. These contradicting effects are due to neural network learning mechanism that learns through penalization of wrong predictions. In the case of high collinearity, this can lead to final feature weights that counterbalance each other.

Non-linear Neural Network



F153: typeFreqsPerTypeFoundInSubtlex
F256: nTokens
F163: typeFreqsPerTypeFoundInGoogle00
F140: typeFreqsPerTypeFoundInDlex

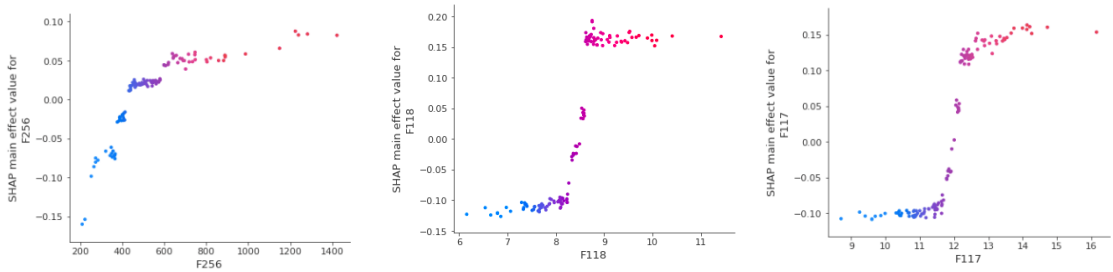
Figure 60: Partial dependence plots for most important features of the non-linear neural network.

For the non-linear neural network, the text length feature shows a positive effect on the outcome variable for some predictions as well. However, the affected data points are even less[153], because the model is more complex than the other. The other highly relevant features are again lexical sophistication measures. While the relationship of the measure for the Subtlex corpus was negative in the linear

[153] This is probably caused by the increased complexity of the model.

model, the non-linear model shows a positive relationship for this feature. On the contrary, the lexical sophistication measure based on the Google reference corpus, shows a negative effect in the non-linear model, while it was positive in the linear model. This hints to the conclusion that the way the collinear feature effects get cancelled out among each other is rather arbitrary and individual feature effects cannot be trusted. While collinear factors in the regression models lead to insignificant factors and difficulties with model selection, the variable effects were more interpretable, in the sense that variable effects would not be inverted because of the internal logic of the model. For the partial dependence plots for the neural network models, this cannot be guaranteed.

Random Forest



nTokens rootTTR corrected TTR

Figure 61: Partial dependence plots for most important features of the random forest model.

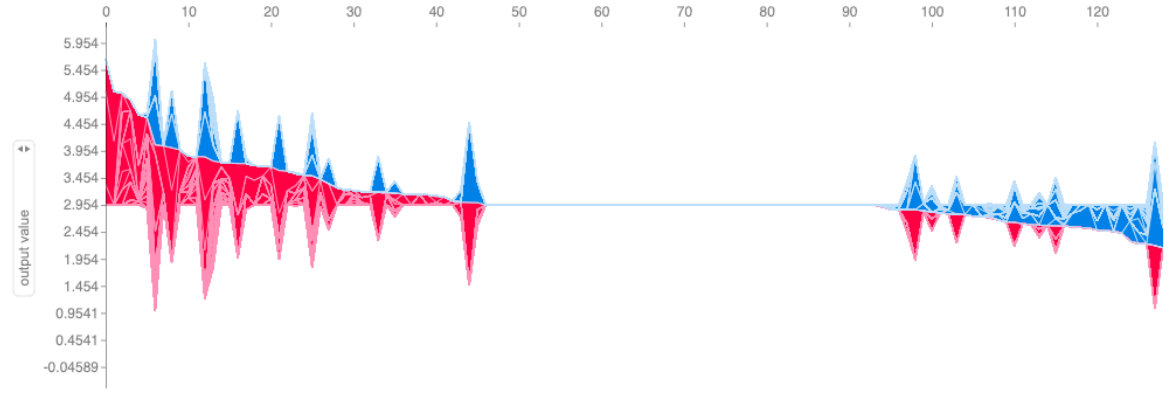
The interpretation of the main effects for the most relevant variables in the random forest model based on the SHAP dependency plots is less ambiguous. The three most important variables in the random forest model are the number of tokens in the text (`F256_nTokens`), the root corrected type token ratio (`F118_rootTTR`) and another corrected version of the type token ratio (`F117_correctedTTR`), where tokens are doubled before root transformed (see also first three features of the global summary plot for the random forest model). The text length as well as the knowingly text length-dependent type token ratio measures are, as expected, positively related with the predictions for the holistic grade. There are no contradictory effects as observed in the neural network models. However, the model tells us that only highly text length-dependent features are used to predict the holistic grade and all other features have only marginal importance. What is more, the lexical sophistication measures that clearly dominated the neural network models, do not even occur among the 20 most important features of the random forest model.

Local model interpretation

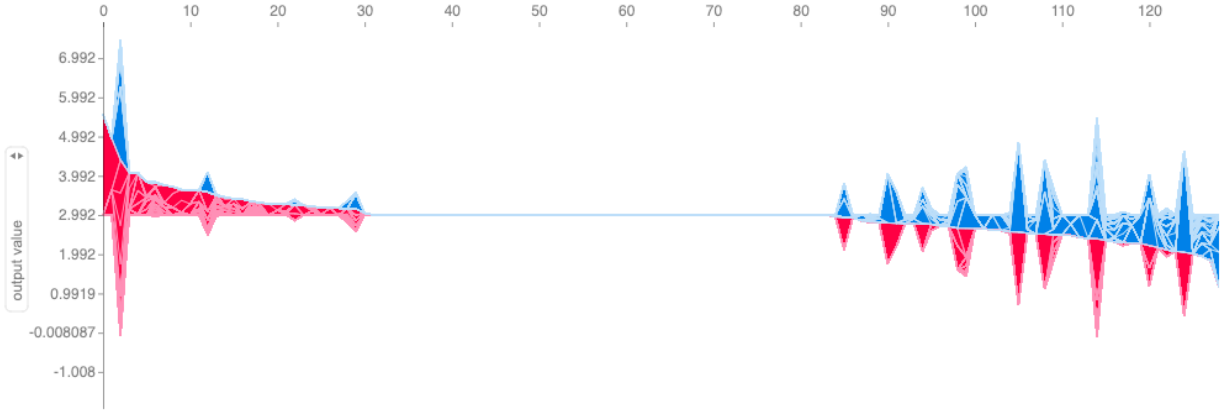
Locally we can interpret the models using individual or stacked local effects plots.

The stacked local effects plots can be used interactively and allow to switch from a summary of all variables (Figure 62) to a graph for an individual variable effect (Figure 65). The y-axis shows the predictions and can be ordered by the original train set order of observations, by data point similarity or by the predicted outcome value. This helps to compare local effects that occur in a broader local (or global) area.

Linear Neural Network



Non-linear Neural Network



Random Forest

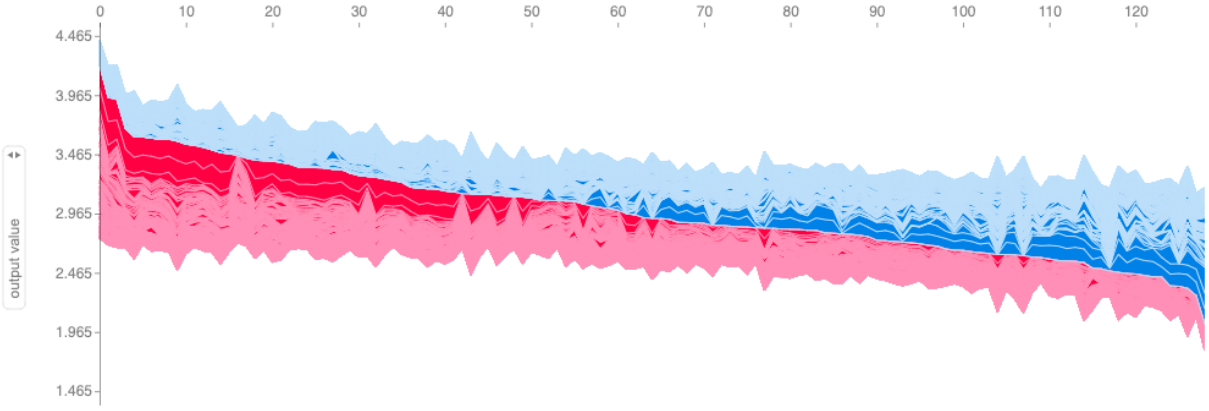


Figure 62: Stacked local effects plots for local model interpretation.

The graphs show summarized effects of the variables on the individual predictions. The observations are ordered on the x-axis by the predicted outcome value (non-rounded predictions for the regression task. The line in between the blue and pink band is the prediction for the individual observation. The blue bands on top show negative effects of variables, the pink band below shows positive effects of individual variables on the predictions. The brighter lines in between the blue and pink area separate the effect of different variables.

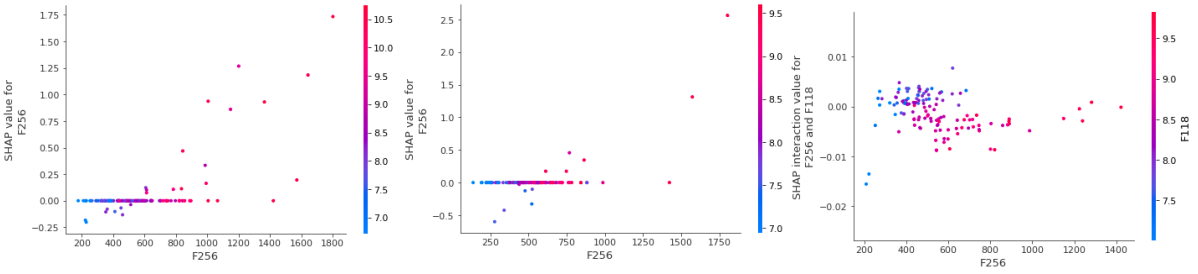
The graphs show the different approaches of neural network models and random forest models. Both neural network architectures do not show any effect for any variable in the middle part of the predictions. For all those parameters the model did not intervene at all and just predicted the mean value for the training set (2.95 for the linear network and 2.99 for the non-linear network). For all the predictions that did differ from this value, variable effects are marked in pink or blue. We can see in the two network models that both sides of the prediction spectrum are marked by separated high peaks that display mostly individual variables that had a strong effect on the prediction. The affecting variables, however, are barely overlapping, which can be seen by the fact that there are no broader bands over adjacent predictions (and therefore similar values) as compared to the graph for the random forest model. Instead many effects are displayed as triangles of one colour that is most often counterbalanced by another triangle of the other colour. Exploring the interactive visualization for this graph shows that the both sides of the resulting diamond represent different but most probably collinear variables (e.g. typeFreqsPerTypeFoundInGoogle00 and typeFreqsPerTypeFoundInDlex). The absence of broader bands of similar variable effects for adjacent predictions is particularly obvious in the second, non-linear network, where almost all predictions are made on the basis of varying individual variables.
Contrary to the neural network models, in the random forest model all features affect each single prediction, according to the SHAP values. However, while there are a few variables that are important for most of the predictions (i.e. the variables that occur on the highest positions of the aggregated tree models, illustrated by broader bands of pink or blue areas), most of the variables have only little effect on the predictions. While it is possible that individual predictions are only positively affected (negatively affecting variables are not considered at all to make up the prediction), or vice versa, the random forest model predictions are always a combination of various positive and negative variable effects.

The graphs in Figure 65 follow the same principle, but instead of visualizing all the effects for all the models, individual variables are inspected. For each model I chose the most important variable, indicated by the mean Shapley value of the variable, and one of the less important variables that was still among the ten best variables for illustration purposes. If the predictions are affected positively by the variable, we expect the graph to show pink values on the left-hand side and blue values on the right-hand side. The left graph for the random forest model for example shows the effect of the root type token ratio on the predictions. The variable is important for almost all predictions and shows (with few exceptions) a positive effect between the root type token ratio and the outcome variable. The right graph of the random forest model illustrated the effect of the number of deverbal nouns per noun phrase[154]. The variable indicates the nominalisations that are often related to academic language use. Although the variable is reported to be among the 15 best variables in the model, the effect of the variable cannot be read from the graph. The non-determinable effect might be due to an interaction effect that we cannot see in this two-dimensional visualization. However, it could also be a sign of unrelated data, showing up as an effect as the random forest model always considers a certain number of variables, even though they might be unrelated. Contrary to the random forest model, the neural network model is able to ignore variables completely, resulting in variable effects plots that can, in the worst case, illustrate one single affected prediction (and thus lack any generalizability, cf. right graph of the non-linear neural network model for the effect of first person personal pronoun ratio that is used to indicate objectivity in academic writing). Hence, the methodological difference between neural networks and the tree based random forest model is visible also in these graphs.

[154] The relationship is, however, not linear, as the effect changes abruptly in the mid area.

Interaction effects

At last we investigate two expected interactions in the models, the interaction between text length and lexical diversity and the interaction between the text length feature and the rater giving the holistic grade. The two graphs on the left of Figure 63 show partial dependence plots for the neural network models that print the value of a second variable as colour in order to illustrate possible interactions. The graph on the right is an example of a specific interaction plot that can be printed for random forest models. In the plots w see the interactions for the text length feature `nTokens` and the lexical diversity feature `rootTTR` for all three models. The type token ratio is known to be dependent on the text length. This dependency of the variables can be seen in the distribution of colours according to the text length variable on the x-axis. As expected, the lower values of `rootTTR` (blue dots) are located on the left side of the graph, while the higher values (pink) are located on the right side. However, the plots for the neural networks do not allow to see the actual interaction effect, but only indicate the presence of the interaction. The interaction plot for the random forest model visualizes the actual interaction effect (effect on the prediction that is caused by a combination of both variables). It shows a high interaction effect for very low text length and `rootTTR` as well as for average or length texts. While a low lexical diversity is influencing the effect of the text length feature strongly negatively, for average length texts, a lower lexical diversity effects the prediction positively (or at least not negatively).

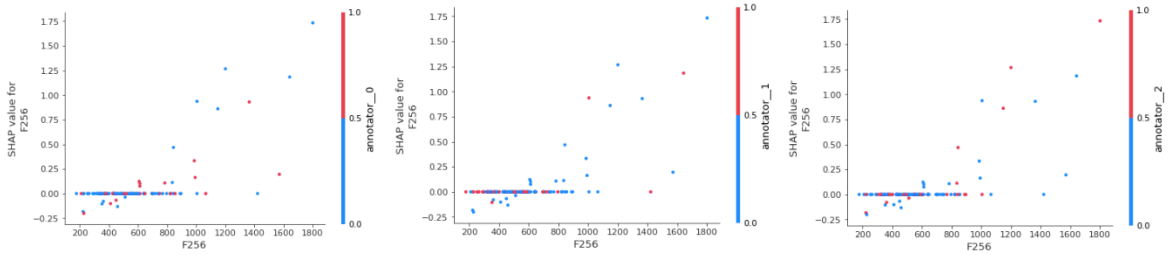


Linear Neural Network | Non-linear Neural Network | Random Forest

Figure 63: Observing interaction effects with partial dependence plots and interaction plots in SHAP.

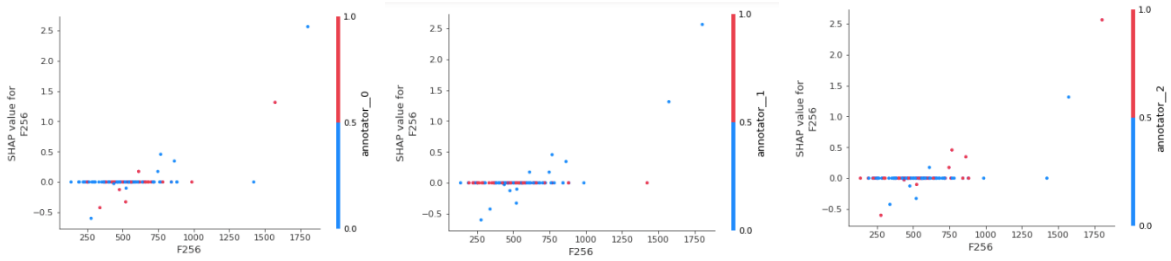
For possible interaction effects with the rater I use interaction effects with the text length feature (number of tokens) as an figurative example, as the variable was among the three best variables in all three models (see Figure 64). The coloured partial dependency plots for the neural network models do not show any visible relationship between the text length and the raters. Pink and blue spots are distributed randomly over the few predictions where text length had an effect. However, the interaction plots provided for random forest models made it possible to observe an interaction between Annotator B and text length. Annotator B was less harsh on very short texts but rated average length texts worse than others.

Linear Neural Network



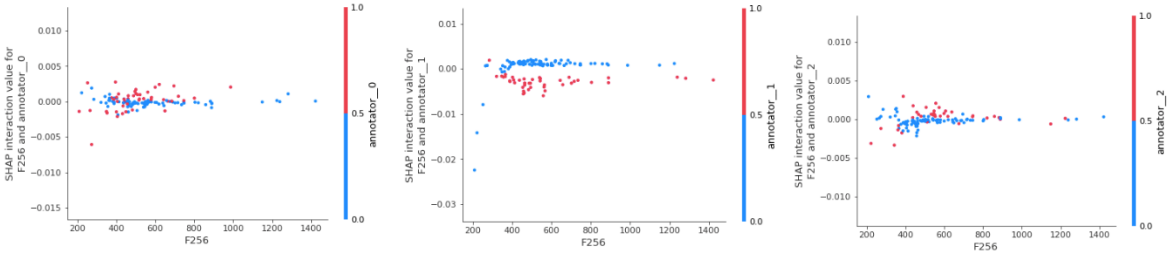
Annotator A Annotator B Annotator C

Non-linear Neural Network



Annotator A Annotator B Annotator C

Random Forest

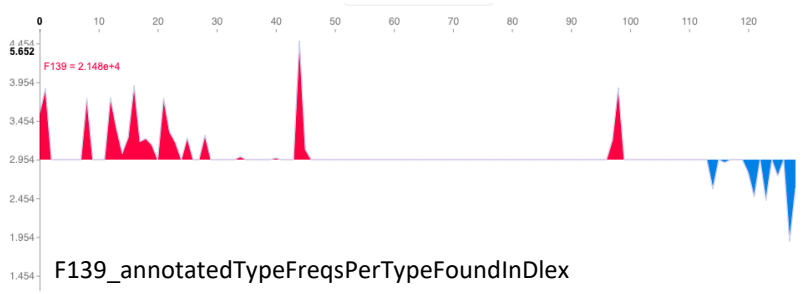


Annotator A Annotator B Annotator C

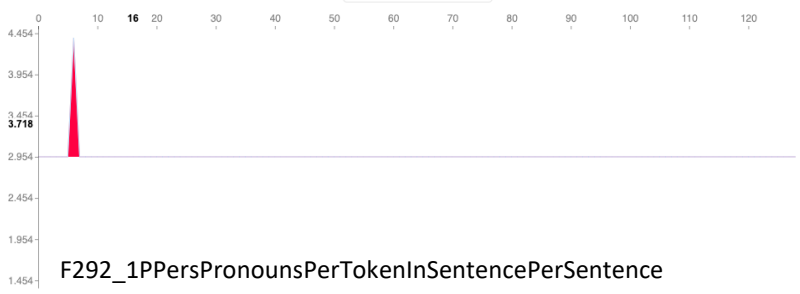
Figure 64: Interpreting annotator interaction effects in complex models.

Best feature vs. other highly ranked features

Linear Neural Network

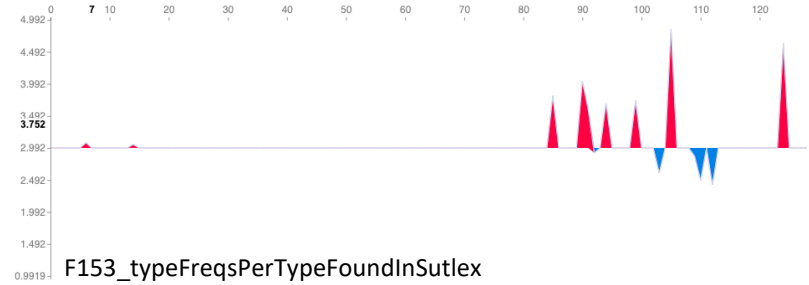


F139_annotatedTypeFreqsPerTypeFoundInDlex

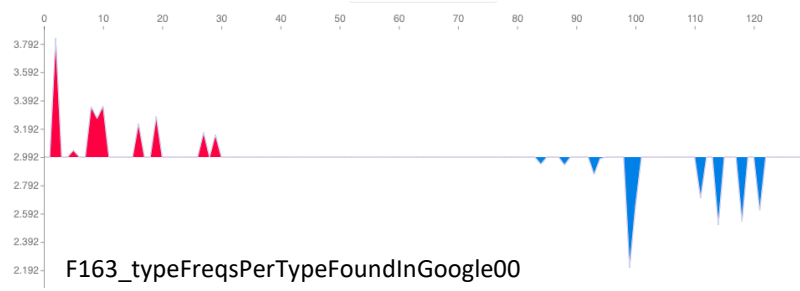


F292_1PPersPronounsPerTokenInSentencePerSentence

Non-linear Neural Network

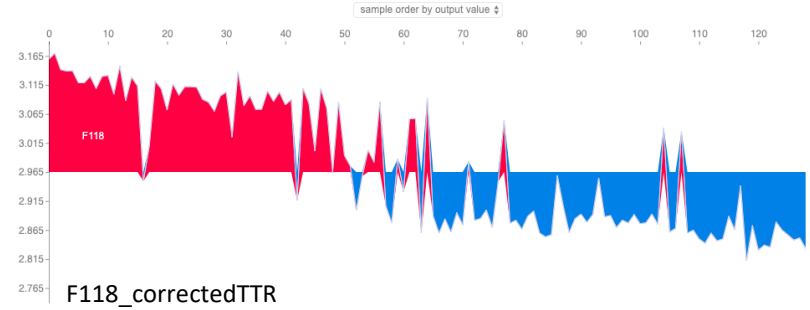


F153_typeFreqsPerTypeFoundInSutlex

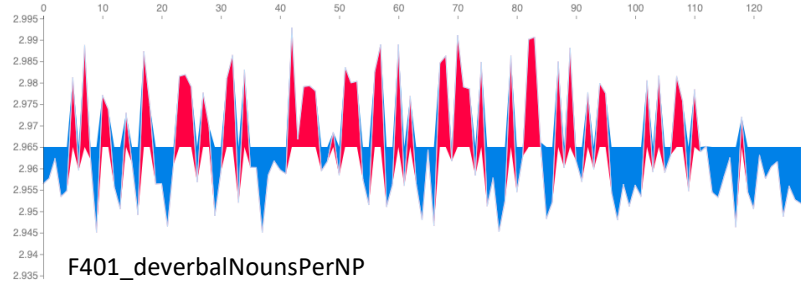


F163_typeFreqsPerTypeFoundInGoogle00

Random Forest



F118_correctedTTR



F401_deverbalNounsPerNP

401

Figure 65:Comparison of local effects for best vs. 4th best feature.

Summary

The interpretation methods presented above gave some insights on the model internals. We found that text length was an important feature in both model types. Apart from that, the neural network models highly depended on lexical sophistication measures measured in terms of the frequency in which the words of the essay would occur in other reference corpora. The random forest on the other hand based its decisions mainly on (text length-dependent) versions of lexical diversity. Text length, measured in the number of tokens in the essay was always positively related with the assigned holistic grades, however, there was visible interaction effect with the raters in the random forest model, where Annotator B did give lower grades to average length texts while he did not necessarily evaluate very short texts badly. Another interaction effect was visible for the root TTR lexical diversity measure which influenced average text length essays badly if it was higher. The main effects for lexical sophistication measures in the neural networks were contradictory and most probably only caused by multicollinear high variance variables that the neural network accounted for by counterbalancing feature effects. The main effects for lexical diversity measures in the random forest model showed a visible non-linear effect of the variable on the predictions that could, however, be caused by variables position in the base tree models of the tree ensemble and it's corresponding splitting. The local effects plots that allowed to explain individual predictions and local regions of the prediction spectrum yielded similar insights. However, the available data did not allow to see local effects that concern more than one or two data points in the highly complex neural network models. The local effects plots for the random forest models on the other side where noisy and variable effects for less important features are hardly interpretable.

Results of the analysis on the linguistic complexity dataset

The exploration of linguistic complexity measures related to holistic text quality judgments showed a number of relevant aspects for the prediction and explanation of holistic grades, such as the text length, the lexical variation in the text, as well as the elaborateness and diversity of the vocabulary and implicit cohesive devices such as the transition of grammatical roles and uptake of arguments from previous sentences. Although monofactorial analysis allowed to identify text quality-related measures of linguistic complexity, multifactorial methods with this complex feature set were difficult, as the interpretability of predictive models was severely impeded by lacking predictive performance, excessive complexity of the model, issues of multicollinearity, lacking robustness of interpretation methods, and by lacking evidence in the data when observing smaller effects in the variables in well-performing black box models also by.

13 Summary and discussion

This study presented a broad exploration of different linguistic aspects that were expected to be relevant for annotators while assigning holistic text quality grades. The study used metadata, in the form of manually annotated text characteristics corresponding to a text analysis questionnaire, linguistic text annotations in the form of error frequencies for orthography, punctuation, grammar and lexis, and automatically extracted computational linguistic measures for linguistic complexity. The analysis used monofactorial quantitative methods (correlations and other association measures for feature ranking) as well as multifactorial

methods based on predictive modelling and machine learning. However, as monofactorial methods have known shortcomings, the analysis focused largely on harnessing predictive modelling approaches.

In feature set 1 with the hand-coded, abstract and content-related features from the **text analysis questionnaire**, we could observe strong relationships between the features and the holistic grades and models built on that feature set had the highest predictive performance. Between the three main groups of features within this feature set, the features regarding the **text content** yielded the best prediction performance, followed by the features regarding the **completeness of the required elements of an argumentative essay**. The features regarding the text structure (e.g. paragraphs, textual structuring elements) did not allow to generalize to the unseen data points in the test sets. However, it was also possible to observe more detailed results regarding individual text quality items that were related with the holistic grades. The text's overall **coherence**, a **clear text topic**, the **conceptual uniformity, clear structure** and the **comprehensibility of the text development** were, next to more subjective items like the **text's general appeal** and the presence of a **golden thread**, consistently chosen as most predictive features by different feature ranking procedures. The features were also highly correlated in a monofactorial Spearman rank correlation analysis on the dataset ($rho > 0.4$, p-value < 0.001). However, these monofactorial results differed slightly from the ones gained from multifactorial methods, were redundancy in the features was found. In a stepwise forward model-selection process using mixed-effects regression modelling with a random effect for the rater, further features, **text appeal, text genre, interestingness** and **coherence** as well as a **comprehensible paragraph structure**, an **explicit opinion statement** and the **adherence to the initially announced topic** remained in the final model, with significant main effects that in sum explained 47% of the variance in the grade level. The model accounting for random effects by random intercepts for the **raters** performed significantly better, suggesting a significant interaction effect between the variables and the raters. We explored the interactions with a heatmap for monofactorial Spearman rank correlations on the three subdatasets for the raters, finding that Annotator B put a stronger emphasis on features regarding the argumentative text genre (e.g. logical argument structure or convincing argumentation). Annotator A on the other hand had higher correlations for the more subjective evaluations regarding text appeal, interestingness, entertainment than the others. The correlations for annotator C were in general more moderate. As already expected by the phrasing of the questions in the questionnaire, not all items were independent from each other. The correlation matrix visualized at the end of section 10.2.4.4 indicated clusters of variables that are not only theoretically but also empirically correlated (e.g. different aspects of text coherence).

These results on German student essays are comparable to other findings in automatic essay scoring and the analysis of text quality in student writing. In a similar study Crossley and McNamara (Crossley & McNamara, 2010), for example, compared the ratings for an atomic (analytic) scoring rubric with holistic text quality judgments in a corpus of English argumentative essays. Similar to this study, they found coherence to be one of the highest predictors of holistic text quality scores in a multiple regression model as well as features related to the clarity of the topic, the conclusion and the structure of argumentation and the used register. However, the study analysed a limited set of evaluations. Evaluations regarding text appeal and other more subjective characteristics of the text (interestingness, humour or entertainment, etc.) that showed high correlations and good prediction performance in this experiment were not present in their analysis. These evaluations allowed us to observe different rater strategies and systematic biases, while analysing a broad spectrum of possibly related

features. Instead, Crossley and McNamara limited multicollinearity and rater bias a priori designing and creating the data purposefully and ensuring thus methodological rigour. Contrary to that, this thesis worked with readily provided data and aimed at exploiting the full set of available features in a data mining approach. While this served for a less rigorous and targeted analysis design it allowed to explore various aspects in a relatively low-cost manner, giving grounds for further analysis on specific aspects.

Regarding the **error frequency features** of feature set 2, a feature type that is very relevant in corpus linguistic studies, less clear statements can be made. The observed features are not important enough, at least not in their unfiltered form, to allow the prediction of holistic text quality grades. Most features are very skewed, biased by other text features like the text length or synthetically exaggerated through normalization approaches. Although the **overall number of errors** was related with the holistic grade, the effect size was rather low (Spearman rank correlation of rho -0.18, p-value < 0.001). Furthermore, aggregating features into bigger groups and investigating monofactorial rank correlations resulted once in a conceptually new category that showed a higher correlation than the individual features (in the case of lexical errors) and once a synthetic group that did summarize features for which the relationship was of different effect size and different direction (in the case of punctuation features). Besides the total amount of errors, only **missing commas**, **wrong word separation or compounding** and the **total amount of lexical errors** were significantly related with the holistic grades. However, in a hierarchical regression model only the total amount of lexical errors and the word compounding errors remained as significant main effects. Grammar errors had no relevance at all. These results resemble the findings for similar corpora in other languages, which also showed very low correlations and predictive performances for spelling (orthography) and punctuation errors with native writers of a certain age, while equally showing no relationship at all between holistic text quality scores and grammar errors (e.g. Crossley, Kyle, Varner, et al., 2014). The detailed multi-level annotation scheme for error annotations could thus not give a lot of additional insights on the nature of holistic text quality judgments in the KoKo corpus. Interesting was, however, to see that some error types seemed to be related only by their total number, while for other error categories only some subtypes where relevant. Nevertheless, the features are interesting in terms of the depicted outliers that make it possible to detect erroneous data points or inspect interesting cases.

Finally, the **linguistic complexity measures** in feature set 3 revealed some more, albeit less strong and clear-cut relationships. The results of a preliminary correlation analysis showed weak to moderate significant rank correlations between the holistic grades and some of the linguistic complexity measures. The correlated measures address **lexical diversity, text length, lexical sophistication** (lexical frequency and age of acquisition of words), **syntactic variation** as well as some **measures for cohesion**. The found correlations correspond to other studies where similar measures were used for other languages (Crossley & McNamara, 2011b; McNamara et al., 2010; Östling et al., 2013). The automatically extractable feature set furthermore allowed to train predictive systems that yield prediction results that are significantly above the majority baseline. They are thus predictive for the holistic grade and can be used to explain the variance in the grade labels. However, the high-dimensional, noisy and not necessarily human-readable feature set only yields good prediction performance when using complex model typologies (random forest or neural networks) that do not allow immediate

intrinsic interpretation (e.g. by inspecting regression coefficients or tree visualizations)[155]. When interpreting these black box models, only few highly relevant variables could be detected and interpreted in greater detail. The variables concerned the **text length** (a well-known factor of holistic text grades, cf. Grabowski and Becker-Mrotzeck 2014, Crossley et al. 2014, Crossley et al. 2010) and **lexical sophistication** or **lexical diversity** features. For lexical diversity we could also observe a clear **interaction effect** regarding **text length** (average length texts are rated worse when lexical diversity is high). Another text length-related interaction concerned the **raters** (one of the raters is stricter on average length texts). However, variable effects for lexical sophistication and diversity had some questionable aspects (method-specific contradictory effects and text length dependency) and have to be taken with a grain of salt.

14 Conclusion

This first empirical corpus study used an existing corpus of argumentative first language student essays to explore features of holistic text quality judgments and compare the grading strategies of three different annotators. The study used a wide set of possibly relevant, already available or automatically extractable features, including text metadata from a text analysis questionnaire, relative error frequencies from a hierarchical error annotation scheme and automatically extractable computational features of linguistic complexity. It identified and interpreted relationships, possible confounding factors and interactions, as well as individual or maybe outstanding observations relevant for further, more targeted analyses by exploiting the available resources.

The study contributed to the linguistic analysis of text quality by discussing and bringing together different approaches to operationalize and analyse text quality from different scientific fields. It is one of the few studies on text quality prediction with NLP methods on German language (but see Horbach et al., 2015; Weiß et al., 2019). The results of this study were, however, comparable to similar studies in other languages and help to define non-language-dependent features able to characterize text quality. Despite the vast number of investigated features, the gained insights in terms of new, interesting patterns of language use are, however, few, as model performance or data size was in many cases too low to make statements for features with smaller effects or less frequent holistic grades (e.g. insufficient or excellent grades).

Future research should thus aim to improve the interpretability of the built models. Using categorical data for black box interpretation or refining used feature sets for linguistic complexity could be an easy first step to do so. Improved and newly developed methods for model-specific and model-agnostic post-hoc interpretation of complex models could possibly still add further value to the analysis of text quality. However, in order to interpret high-dimensional predictive models with such methods, larger amounts of data are needed.

[155] What is more, the simple, so called *intrinsically interpretable models* are also lacking concrete interpretability when complex feature sets are used, as they equally grow in complexity.

Empirical corpus study 2

Age-related language in South Tyrolean Social Media

15 Introduction

The second corpus study presented here addresses a series of research questions in the field of age-related language in social media. Methodologically the study aims to evaluate techniques for data-driven corpus analysis, using predictive modelling and data science methods in a confirmatory research design. It makes use of the methodological framework of age-prediction studies from computational sociolinguistics, building upon the knowledge of learning algorithms, feature types and operationalizations that were successful in similar studies. The study analyses the social variable age in a corpus of South Tyrolean social media texts of three different text types, identifying and interpreting linguistic phenomena that are related to the age group a writer belongs to.

At the basis of this study are theoretic concepts and previous observations of age-specific language in sociolinguistics. I first summarize existing knowledge in the field, stating previously analysed linguistic phenomena and observed confounding factors that will inform the following experiments and set the expectations made about the corpus. After this, I briefly present the DiDi corpus of South Tyrolean CMC (Frey et al., 2015, 2016; Glaznieks & Frey, Forthc.) that has been created with a similar research aim, investigating age-related social media language in South Tyrol, before I describe the research design of the current study, including research questions methodology, as well as the specific corpus subsets used for the following experiments.

The experiments test the difference in used language between generations and age groups; and describe language use of individual groups in an exploratory fashion to further elaborate the hypotheses. This way, the results triangulate the relationship between both entities, investigating sociolinguistic topics like digital generations, language change, age-and generation-specific language, youth language or language of the elderly and digital age.

Finally, I summarize the results and discuss the value of this study for our understanding of age-specific social media writing on Facebook and for South Tyrolean language use in non-institutional writings, setting an agenda for possible future investigations.

16 Theoretical background

16.1 Language and age

It is well known that a person's language use is related to his or her age. Indeed, age, next to gender, social class and education, is one of the most important variables in sociolinguistics. Linguistic features that are related to the speaker's or writer's age have been studied since the early days of sociolinguistics in the late 50s and 60s (Fischer, 2015; Labov, 1966; Trudgill & Trudgill, 1974). They were found to be either age-specific, i.e. "linguistic behaviour that is appropriate to and typical of different stages in a speaker's life span" (Cheshire, 2005, p. 761), or generation-specific, i.e. language features that "reflect language change, in that older speakers may not have undergone the linguistic changes that have affected younger generations" (Cheshire, 2005, p. 760). While the former is usually passed on from one generation to the other, always occurring in the same age spans and allowing *age-grading*, the latter is usually grounded in one generation and remains within this group of people. Cheshire (2005) also points out that there are few linguistic features that are age-exclusive, in that they only occur within a certain age group (e.g. elderly speakers' trembling voice), while most features are age-preferential, meaning that they occur more frequently in one age-group than in the others.[156]

A recurring and almost universally acknowledged example of age-preferential language is the preference of middle-aged people to use more prestigious forms and varieties of the language. While the pressure of individualization and rebellion among adolescents[157] and the reduced need to adapt one's (linguistic) behaviour to societal norms in elderly people[158] favours the use of less conventional, less prestigious forms, middle-aged people who need to adapt to society for work and social life, generally use prestigious standard varieties more frequently (see also Mattheier, 1987; Thimm, 2002). This leads to an approximately U-shaped distribution of vernacular language use over the lifespan of a person as can be seen in the figure below (Figure 66).

[156] See Neuland (1987) and Coupland (1997) for more detailed overviews on youth language and the language of elderly people respectively.

[157] This effect is also described in the adolescent peak principle (e.g. Chambers, 2003; Labov, 2001) that assumes that adolescents reach a peak of non-conformant language use between 15 and 17 years (see also Peersman et al., 2016).

[158] Compare also Justine Coupland et al. (1991) and Nikolas Coupland (1997) for a critical view on the often deficiency-oriented explanations for non-standard language use in older speakers.

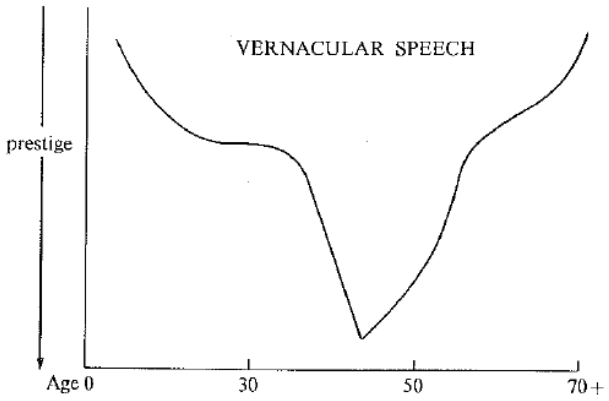


Figure 66: Age-preferential use of vernacular language over the lifespan of a speaker. Figure obtained from Downes (1998).

However, the social variable age can be established not only in terms of years since birth but also according to other factors such as physical maturity (e.g. biological age) or socially via one's affiliation to social groups and past experiences. Consequently, age-related language is not necessarily exclusively tied to the chronological age of a person. Instead, other factors can play an important role in how "young" or how "old" a person speaks or writes (e.g. a 30-year old student might show a different language use than a 30-year old office worker). After first studies in age-related language investigated primarily quantitatively measurable operationalizations of age in years since birth (Fischer, 2015; Labov, 1966; Trudgill & Trudgill, 1974), further research also explored other possible age concepts. Eckert (Eckert, 1997), for example, distinguished between *chronological age* (the number of years since birth), *biological age* (one's physical maturity), and *social age* (performed by group affiliations and experience). Although these different definitions of age can overlap, they do not necessarily need to. It is therefore suggested to consider not only the chronological age, but also other operationalizations of this social variable, when analysing age-related language. However, the chronological age still remains the default operationalization in most studies, as it can be easily measured.

Another known aspect of age-specific language is its frequent interaction with other social and contextual factors. More than one study reported the influence of gender on age-specific language, giving reason to always control one, when investigating the other (for studies referring to this interaction of age and gender in online contexts see for example Hilte et al., 2017; Nguyen, 2017; Nguyen et al., 2013). Besides, language use is usually adapted to the audience or interlocutor, which renders e.g. inter-generational communication difficult to analyse. Which age-specific features are used, and how much, can be instrumentalized intentionally to construct age identities and to make age relevant in the situation. In this way, age is not only defining how a person speaks or writes, but the linguistic behaviour is also defining how old or young he or she is perceived to be (Coupland et al., 1991; Georgalou, 2015a). One can see this clearly in the virtual communicative spaces of digital media, where linguistic features are strategically set to display identity, such as age identities (Georgalou, 2015a, 2015b) or translocal identities (Kytölä, 2016; Leppänen et al., 2009).

16.2 Age and computer-mediated communication

The social variable age is particularly relevant in the study of computer-mediated communication (CMC). Ever since Marc Prensky's theory of the Digital Native (Prensky, 2001, 2009;

Thornham & McFarlane, 2011), i.e. a person that was born and raised in times where the world wide web was accessible via personal digital devices, young people are commonly believed to be the drivers of CMC and the inventors of netspeak and similar linguistic phenomena that emerged with and are specific to digital media (cf. Androutsopoulos, 2006; Crystal, 2001; Siever et al., 2005). CMC is often connected with youth language and early sociolinguistic research on CMC often focused specifically on young people (Androutsopoulos & Georgakopoulou, 2003; Herring & Kapidzic, 2015; Leppänen, 2007; Siebenhaar, 2006; Tagliamonte & Denis, 2008). Consequently, the research concerning age and language use in CMC has a bias towards younger generations that is further reinforced by older generations being less active in the popular social media platforms that were used for linguistic investigation[159].

In terms of genres, however, there are studies for almost all possible text types or *communicational forms* (Dürscheid, 2005) that have emerged in the world wide web. Blogs, microblogs (e.g. Twitter), social networking sites (e.g. Facebook), instant messaging (e.g. Whatsapp), the more traditional SMS texting, wikis, and discussion forums have all been investigated for age-specific language.

While there are some qualitative, mostly ethnographic studies (Androutsopoulos, 2003; Androutsopoulos et al., 2013; Georgalou, 2015a; Stæhr, 2015) and qualitative as well as quantitative corpus linguistic approaches (e.g. Hilte et al., 2016; Peersman et al., 2016; Santillán, 2009)[160], a substantial amount of research was conducted in the field of computer science and computational linguistics in the form of age prediction studies (see section 16.3 below).

Age-related linguistic features have been investigated mostly on the lexical level, both in terms of content as well as style. Alongside non-standard language (Peersman et al., 2016; Siebenhaar, 2006), other language features that are specific to CMC, such as emoticons, character flooding, acronyms or fully capitalized words, have been reported as important linguistic features for describing the age of a person. Moreover, sentence and word length as well as lexical density have been frequently reported in studies. Young writers have been found to use more slang words, more CMC-specific phenomena, more out-of-dictionary words, more self-references as well as shorter words and shorter sentences (Goswami et al., 2009; Nguyen, 2017; Rao et al., 2010; Rosenthal & McKeown, 2011; Simaki et al., 2016b). With regards to CMC-specific phenomena, older writers use not only less emoticons (Hovy et al., 2015) they also do not replace punctuation marks by the use of emoticons equally often as younger writers tend to do (Spina, 2019b).

In general, research has not divided age-specific from generation-specific features of CMC as these categories are, in practice, hard to separate with synchronic corpora that only provide *synthetic age cohorts*. An exception is the work of Danescu-Niculescu-Mizil et al. (2013), who analysed two active online communities both on a community level and on the level of the individuals, and could show how certain lexical choices arise and dissolve with the constantly changing generations of users.

[159] Up until very recently, many platforms were rarely frequented by elderly people, which led to a severe lack of linguistic evidence for retired people or age groups over 50 years of age. Nguyen et al. (2014) and Peersman et al. (2016), for example, both report that they could not analyse older people's language behaviour because of lack of data in their corpora.

[160] See also Baron et al. (2012) for further references.

Apart from a the interest in age- or generation-specific features, research is also inspired by the question of how digital media changes the language practices of young people (Abel & Glaznieks, Forthc.; Dürscheid et al., 2010; Lobin, 2014; Prensky, 2001, 2009; Storrer, 2013). Most recently, studies have also focused on the intersection of online and offline writing (Hilte et al., 2016, 2017; Stæhr, 2015; Verheijen, 2017).

16.3 Age prediction and computational sociolinguistics

A substantial amount of research that deals with age-related language is conducted within computational linguistics. The computational linguistics community has been engaged with automatic authorship attribution and the more general tasks of author profiling for many years (Estival et al., 2007; Stamatatos, 2009). In author profiling, computational linguists try to automatically predict characteristics like age, gender, first language, or even the personality of the author of a text on the basis of their language (Rangel et al., 2016). This line of research draws substantially from previous research in register studies and stylistics (e.g. Daelemans, 2013; Goswami et al., 2009). Investigated phenomena thus often originate from stylometry and are mostly limited to text surface features like frequency lists of function words, vocabulary size, or other features that are independent from the content (Simaki et al., 2016a; Stamatatos, 2016)[161].

Few studies in age prediction have taken a deeper look into the linguistic correlates of age, as their focus is clearly on technical approaches to prediction[162]. One important, new strand of research that does deal with the actual interpretation of age prediction models is *computational sociolinguistics* as laid down by Nguyen et al. (2016) and taken up by Simaki et al. (2016a), Hovy and Johannsen (2016) and Dunn (2019), among others. It can be seen as an interdisciplinary cross-over between sociolinguistics and computational linguistics, where both communities mutually enhance each other's research agenda. This emerging field "integrates aspects of sociolinguistics and computer science in studying the relationship between language and society from a computational perspective" (Nguyen et al., 2016). It exploits the advances made in author profiling in order to analyse age-specific language linguistically (e.g. Peersman et al., 2016; Simaki et al., 2016b, 2016a), observe language change (e.g. Danescu-Niculescu-Mizil et al., 2013) or – more related to social sciences and digital humanities – investigate and discuss the possibility to operationalize social categories on the basis of language[163] (e.g. Kosinski et al., 2016; Nguyen, 2017; Nguyen et al., 2014)[164].

[161] If the features were not independent from the actual content, it would not be possible to distinguish between differences that are due to the author and differences that are due to the topic or genre that was written.

[162] However, see Hovy (2015) or Stoop and van den Bosch (2014) for examples and Ribeiro (2016b) for theoretical reasons why prediction-oriented computational linguistics and NLP can benefit from investigating the models built.

[163] For instance, Nguyen et al. (2013) compared age prediction results for chronological age and life stages (social age) and showed that both have comparable results, but that the social age might be preferred for theoretical reasons.

[164] It is worth noting that both age prediction studies and computational sociolinguistic studies are not limited to CMC. However, a major part of the research utilizes CMC data as it is more easily available and more easily processable than other types of data (e.g. spoken discourse). The larger the amount of data, the better the computational methods for author profiling will work.

16.4 Language in South Tyrol

South Tyrol is an autonomous province in northern Italy in which Italian, German and Ladin are officially recognised languages of the population. Italian and German are the official languages for public administrative and institutional communication across the whole territory and are obligatory study subjects for any South Tyrolean school student. The German language, although being a minority language when considered at a national level, is the language of most speakers in South Tyrol. Almost two thirds of the inhabitants officially declared German as their first language (L1), as reported by the last census in 2011 (ASTAT, 2016). The remaining part of the population is split into around 30% that declared to belong to the Italian language group and around 4% that declared to belong to the Ladin community. The Italian and Ladin populations are mainly concentrated in certain areas (Ladin being the most prominent language in the valleys Badia and Gardena, and Italian being more prominent in the capital and its suburban areas to the south) while the rest of the territory (especially the remote valleys) are primarily German-speaking (ASTAT, 2016). However, the centuries-long contact between the two languages that goes beyond the Italian annexation of the territory (Eichinger, 2002), as well as the public encouragement towards learning both German and Italian as well as other languages, has allowed for intercultural contact and an individual multilingualism next to the official institutional and social multilingualism (Abel et al., 2012). This can also be seen in social networking sites, where South Tyrolean residents express their multilingual literacies and intentionally construct identity as well as desired audience through language choice (Frey, 2018; Frey et al., 2016; Glaznieks & Frey, 2018).

With respect to the German-speaking population, South Tyrol shows another linguistically interesting phenomenon. As in many stable minority groups, dialectal usages of the language are widespread over many areas of social life and contribute to the local identity management of the German population (Riehl, 2007). While dialect is used in almost all areas of everyday spoken communication among German-speaking South Tyroleans, the standard variety is restricted to mostly written language use as well as administrative and institutional communication (e.g. in schools, official public speeches or radio and TV broadcasting) (Abel et al., 2012; Schober, 2007). The linguistic situation for these two German-language varieties can thus be considered as diglossic in the sense of Ferguson (Ferguson, 1959)[165]. This is also visible in the non-institutional and private social media usage of South Tyrolean residents, which is characterized by a frequent and consistent use of the South Tyrolean dialect (Glaznieks & Frey, 2018; Glaznieks & Glück, 2019; Glaznieks & Stemle, 2014)[166].

Regarding age-related language use, research has focused on traits of language contact and multilingual language competences in South Tyrolean youth language. It has been stated that contact-induced inferences enter the South Tyrolean standard variety of German through administrative language (Lanthaler & Saxalber, 1995) (e.g. *'Autobüchlein'*, '*Supplenz'*). Further contact phenomena can, however, enter South Tyrolean every-day language through use among young people in cities and "strongly bilingual groups" (Anstein, 2013). South Tyrolean youth language (of the German language group) is characterized by elements of Italian youth language as well as the language of youth from other German-speaking environments (Vikoler, 2016). Italianisms are used for cursing, greeting and addressing peers as well as for

[165] But see Lanthaler (2001) for critical voices against this classification.

[166] For example, Glaznieks and Glück (2019) showed that spelling conventions found in the Facebook communication of South Tyroleans resemble the dialect realizations found and reported over different areas within spoken language.

interjections (‘*dai’*, ‘*bo’*) and routinized formulaic sequences (e.g. *’ma va’*) (Glaznieks & Frey, 2018; Humenberger, 2012; Vikoler, 2016)[167]. Mixed code and code-switching are also often reported as a style instrument of young people (e.g. Eichinger, 2001). However, code-switches are mostly restricted to simple, non-integrated forms (e.g. nonce borrowings, greeting formulae) (Glaznieks & Frey, 2018), probably due to the lack of appropriate second language competences and emotional grounding (Abel, 2007; Glaznieks & Frey, 2018; Vikoler, 2016). The focus group study of Abel et al. (2012) revealed that the second language is “more of a school subject than a tool for communicating in everyday life” for young people, while the German dialect is regarded as an important tool for identity management, group affiliation and audience design[168]. This finding is supported by Vikoler (2016), whose questionnaire data revealed that young people perceive a mixed language style as a marker of their South Tyrolean identity rather than an expression of their second language competence. The stronger presence of dialect varieties of German in social media and mobile communication (SMS) of younger people was additionally attested by Glaznieks and Frey (2018) and Huber and Schwarz (2017). However, these results are based on qualitative explorations and comparisons of mean values that were not tested for statistical significance.

17 The DiDi corpus and its characteristics

The DiDi corpus of South Tyrolean CMC (Frey et al., 2015) contains authentic written language from a total of 133 South Tyrolean users of the social networking site (SNS) Facebook. The total sum of around 40.000 individual productions can be divided into texts that were published semi-publicly on the users’ Facebook walls (ca. 11 000 status updates and ca. 6 000 responsive comments) or privately via Facebook’s built-in instant messaging service (chat, ca. 23 000 messages). It is sociolinguistically annotated with relevant metadata on the age, first language, gender, education, employment and social media usage habits of the writers.

17.1 Corpus collection, building and availability

The DiDi corpus was created as part of the DiDi project that was conducted from 2013 to 2015 at the Institute for Applied Linguistics at Eurac Research Bolzano. The project, entitled *Digital Natives – Digital Immigrants. Writing on social networking sites* was aimed at documenting and analysing non-institutional, private language use in South Tyrol by building an approximately age-balanced corpus of Facebook texts that can be used for sociolinguistic analysis of age-related language.

The data was gathered via a specifically produced Facebook app that allowed the retrieval of user consent, language data and questionnaire forms with socio-demographic metadata from voluntary data donors from the northern Italian province South Tyrol/Alto Adige (Frey et al., 2014). Users were recruited via Facebook advertisements, word-of-mouth recommendation and sharing project information on public Facebook groups. After users agreed to donate their

[167] Humenberger (2012) analysed Italian swear words and curses in South Tyrolean youth language on Facebook and found that young writers tend to use stronger curse words and more non-obfuscated spellings compared to older writers.

[168] This pragmatic strategy is also of crucial importance in semi-public social media communication (Androutsopoulos, 2014; Tagg & Seargeant, 2014).

data, it was downloaded directly via the Facebook API. The texts were anonymized and non-standard spellings were annotated with standardized spellings, transcribing the original non-standard word to the corresponding word form that adheres to German orthography (cf. Frey et al., 2015). This step was needed to facilitate (semi-) automatic processing (part-of-speech tagging and lemmatization) and allowed the identification of dialectal texts and the analysis of non-standard spellings. The texts were then linked to user-related metadata retrieved from a questionnaire form where the users filled in their social-demographic data. Further available annotations regard the CMC-specific language phenomena, the language of the texts, occurrences of code-switching and whether German texts are written in dialect or not.

The anonymized corpus including metadata and annotations is available for research purposes and can be accessed and queried at https://commul.eurac.edu/annis/didi or downloaded in JSON or XML format via https://clarin.eurac.edu/repository/xmlui/handle/20.500. 12124/7.

17.2 Description of texts

The corpus shows the whole set of texts published during the year 2013 on the users' personal Facebook walls (status updates and comments) and/or their private Facebook chats. The entire corpus comprises 596 319 tokens in 39 825 texts. The mean text length is around 15 tokens per text. However, the text length has a high variance, with the maximum text length of 2 880 tokens. In comparison, the median and interquartile range of texts are very short (75% of the data is actually between 4 and 16 tokens long). Less than 4% of texts are more than 50 tokens long and texts longer than 100 tokens only make up 1.2% of the data. This is also due to the fact that the corpus contains different text types. It comprises semi-public initializing status updates as well as the more dialogic forms of semi-public comments and private chat messages. Table 49 shows the summary statistics for text length split by text types. Figure 67 demonstrates the slight differences in dispersion between the three genres.

	Mean	Median	IQR	99 Percentile	Max
Status updates	15.6	7	12	169	2880
Comments	14.5	9	12	93	528
Private chat messages	14.8	8	13	107	1105
Total	*15*	*8*	*12*	*113*	*2880*

Table 49: Summary statistics on text length in tokens.

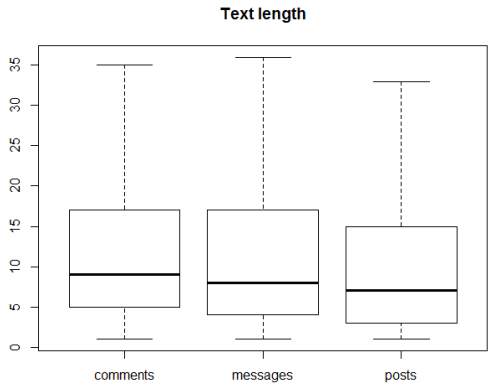


Figure 67: Text length in tokens (without outliers).

17.3 Description of users

All texts are linked to one of the 133 user profiles, for which additional socio-demographic information is available (e.g. age, first language, gender, education or employment status). The 133 collected profiles spread over writers of different first language backgrounds, genders, educational backgrounds and ages. While gender and age are rather balanced for the whole corpus, first language background has a strong bias towards German-speaking South Tyroleans. Employment and educational degree also differ in group size (see Table 50).

	Profiles	Texts	Avg. #Texts (Median)	IQR
Gender				
female	70	20 273	80	322
male	63	19 552	123	292
Education				
school (no high school graduation)	20	2 885	70	151
high school	46	18 248	159	349
university	47	11 972	87	317
Employment				
school	12	4 999	81	212
student	26	5 372	86	206
employed	56	15 180	136	299
freelance	20	9 806	188	630
unemployed	2	1 254	627	626
retired	11	1 692	33	59
First language				
German	105	29 883	87	293
Italian	9	4 260	295	471
German + Italian	11	4 165	126	560
German + other	5	1 110	84	204
other	3	407	153	123
Age group				
14-19	17	5 805	66	201
20-29	32	5 289	75	174
30-39	21	7 514	234	259
40-49	17	8 377	332	584
50-59	29	10 016	182	340
60+	17	2 824	61	131
Total	136	39 825	96	306

Table 50: Overview of user characteristics in the DiDi corpus.

Moreover, the number of texts per user is not uniformly distributed. Table 51 and Figure 68 show the distribution of texts over user profiles for each of the text types.[169]

	Mean	Median	IQR	Max
Total	299	96	306	2844
Status updates (N=117)	84	28	92	780
Comments (N=116)	49	20	52	523
Private chat messages (N=57)	167	0	111	2303

Table 51: Summary statistics on texts per user.

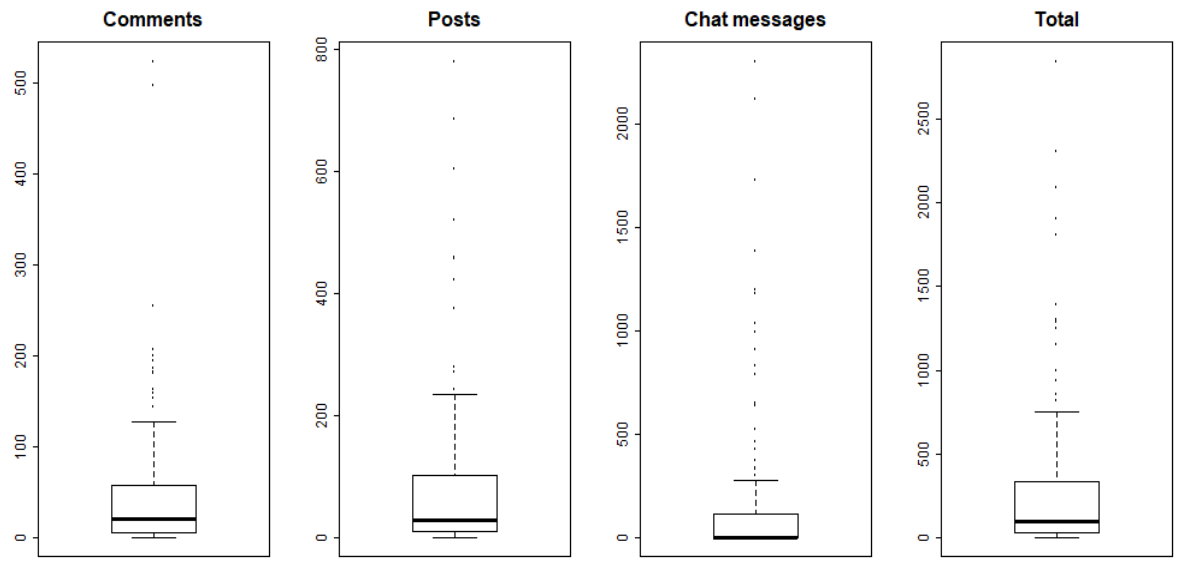


Figure 68: Boxplots showing the distribution of texts per user.

So far, the corpus has been used to study spelling variants related to the South Tyrolean dialect, multilingual repertoires, and cohesive devices as a sign of text quality (Abel & Glaznieks, Forthc.; Glaznieks & Glück, 2019). Only a few, cursory analyses have been conducted on the aspect of age in the DiDi corpus, investigating quantitatively the mean proportion of Italian, German and other language texts for younger vs. older writers (Frey & Glaznieks, 2018; Glaznieks & Frey, 2018), and qualitatively investigating multilingual phenomena and code-switching as well as the distribution of spelling variants for individual words (Glaznieks & Glück, 2019).

[169] The unequal distributions, combined with the hierarchical structure of the data (users who produced various texts of different text types) and the presence of non-standard language, are impeding issues for quantitative analysis and automatic processing.

18 Study design

Given the related research discussed in section 16, it is interesting to study age-related language in the CMC of German-speaking South Tyrolean residents, while making use of the methods developed in computational sociolinguistics.

The freely available DiDi Corpus of South Tyrolean CMC was designed to study age-related differences in CMC. The aforementioned study of Glaznieks and Frey (2018) focused on youth language in South Tyrolean Facebook posts and comments only, and showed that there are age-related differences in language and variety choice that can be observed in the DiDi corpus. However, the analyses were mostly restricted to simple descriptive analysis and qualitative explorations of the language without testing relations for statistical significance. The computational methods deriving from new data science trends, in particular the ones developed in the fields of age prediction and computational sociolinguistics, could potentially be utilized to approach the aspect of age-related language in the DiDi corpus[170] in more detail and with more methodological rigour.

This second corpus study therefore analyses age-related language in the DiDi corpus as a case study to evaluate if and how corpus linguistic approaches on available language resources can be enhanced by the use of data science and data mining methods.

Contrary to the first corpus study that investigated holistic grades in student essays in a broad and explorative way using a wide set of annotations, this study focuses on testing (linguistic) theories and hypotheses on the data in a targeted way, drawing from existing studies in the field. The study gives a multi-faceted view by studying known phenomena with new data and analysing a series of different aspects of the same data. While it contributes to our understanding of age-related language features in social media, it evaluates methodological aspects of data mining methods that could tackle general desiderata for corpus linguistic research. These comprise for example:

- the use of larger variables sets;
- the conduct of multifactorial analysis;
- a faster and less laborious preparation of variables, including the possibility to carry out analysis with different representations of variables and different operationalizations of concepts;
- the modelling of more complex relationships including a detailed interpretation of these relationships.

Strategies from author profiling (e.g. Raghunadha Reddy et al., 2016) and more specifically from the age prediction framework of computational sociolinguistics (see e.g. Nguyen, 2017; Simaki et al., 2016a) are used in order to facilitate an in-depth sociolinguistic investigation of linguistic features related to the writer's age. The study is based on findings of previous research in sociolinguistics and tests concrete hypotheses that were drawn from previous findings including already known linguistic correlates (e.g. those found for other languages) and reported potential confounders (e.g. gender or genre). Finally, the study aims at extending existing knowledge through in-detail model interpretation, including variable importance measures but also including monofactorial and combinatorial effects.

[170] For first approaches see also Frey and Glaznieks (2018).

18.1 Setup and research questions

In order to give an in-detail, multi-faceted overview of the observable traits of age-related language while probing the potential of the methodological toolset, the study tackles a series of research questions.

Section 19 tests a simple hypothesis, namely that language and variety choice as well as CMC-specific style differs between people who were born in a digital age (i.e. people born after 1980) and people who grew up in a non-digital environment. This hypothesis is based on the theory of the so-called *digital native*, who behaves substantially different from his older counterparts, the *digital immigrants,* in terms of communication and media consumption habits (Prensky, 2001, 2009). This theory is widely known in the study of media literacy, discourse, and CMC and has been approached by various researchers before (e.g. Bayne & Ross, 2011; Thornham & McFarlane, 2011). However, the concept of the *'digital native'* is barely researched in linguistics and deserves further attention[171]. Concretely, I ask the following question:

1. Can we use linguistic features of language choice, variety choice and CMC-specific style to classify *digital natives* and *digital immigrants* and thus establish empirically that there is a difference between the two generations?

Section 20 extends the research by considering more complex aspects of age-related language. Instead of the binary division of digital natives and digital immigrants studied in section 19, section 20 investigates a classification model and its predictions for three age groups. It evaluates the predictive performance of the models for different age-groups, to see if there is sufficient evidence in the data for a more fine-grained analysis. Moreover, it focuses on the German texts in the corpus and investigates individual effects of linguistic traits and age predictions. It first tries to separate the effect of individual language phenomena (non-standardness and variety choice) by comparing predictive performance of different variable sets. After that, it extends the set of features used based on previous research in age prediction. The research questions addressed in this section are:

2. Using the same aspects of language use as in the previous experiments (choice of language and variety, CMC style markers) and accounting for a possible interaction effect with the text type, can we classify three age groups (young people, middle-aged people and older people) with sufficient accuracy to expect non-random predictions?
3. How does the variety choice compare to other features of non-standardness when predicting author age for German Facebook texts? Are both features equally relevant to the prediction model or do they even complement each other with additional information?
4. How do the prediction results for German texts compare to models that use further features proposed for age prediction (i.e. punctuation, capitalization, etc.) and which factors are important for discrimination between the three age groups?

Section 21 then zooms in to examine the linguistic traits of the individual age groups. A particular focus is put on older writers. The prediction of three age groups allows to compare the

[171] See Frey and Glaznieks (2018) for a first analysis of the language of digital natives on the DiDi corpus.

language of elderly people with younger age groups using interpretation techniques on the models built in section 20. Section 21 therefore asks:

5. How does the language of elderly people in the DiDi corpus compare to the other groups?
    - Can we find language features that characterize older writers?
    - Where do older writers differ from a) young writers and b) middle-aged writers?

Finally, section 22 questions the chronological operationalization of age and evaluates different alternative age concepts that have been proposed in the literature on age-related language. In particular it tests the appropriateness of the chronological age compared with a *social age* (Eckert, 1997; Nguyen et al., 2013) and three possible operationalisations of a *digital age* (Glaznieks & Stemle, 2014), by asking:

6. Can we evaluate different operationalizations of age on the basis of social media texts?
    - How do the models' predictive performances change for different operationalizations of the age variable?
    - Which operationalizations give the best prediction results based on the language features of interest?
    - Does the *digital age* explain language use better than other age operationalisations?

18.2 Corpus subsets for analysis

All analyses are based on the DiDi Corpus of South Tyrolean CMC (see section 17). However, as the study focuses on German first language speakers, only the texts produced by users who declared German as their only first language (105 users) are used.

In order to reduce user bias introduced by the high variance in post, comment and messaging frequency, the corpus is subsampled to contain a maximum of 300 randomly chosen texts per user. For the first two experiments all texts of a user are considered as possible candidates for the random sample (i.e. `all_language`). However, research questions 3-8 focus on more detailed, language-dependent features of the text and only address the German texts in the sample (i.e. de).

The `all_language` corpus subset contains a total of 14 346 texts from 105 German L1 speakers. The de corpus subset on the other hand contains 11 821 texts. The age distributions for texts and users are displayed in Figure 69 and Figure 70.

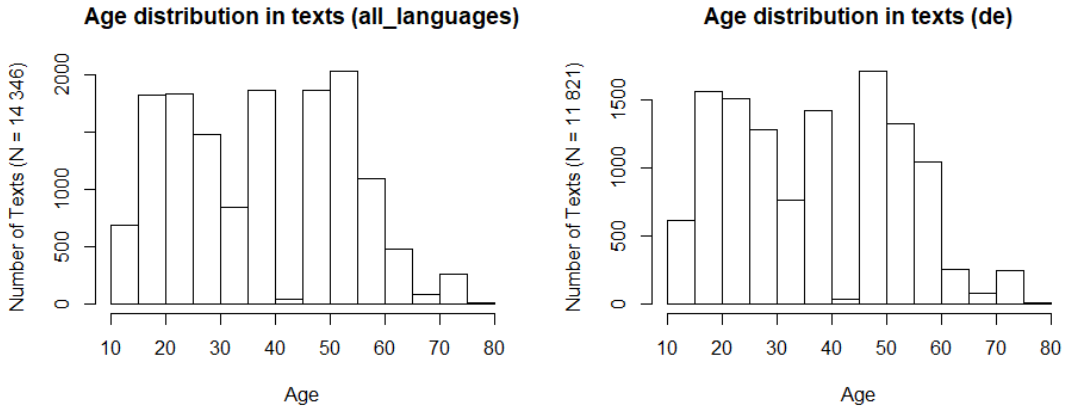


Figure 69: Age distribution in the texts of the corpus subsets.

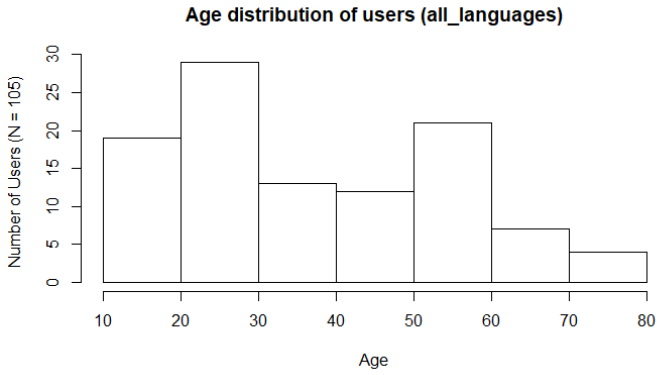


Figure 70: Age distribution of the users in the corpus subsets.

The texts are not uniformly spread over the three text types: posts (status updates), comments (responses to status updates) and chat messages (private 1-to-N communication) (see Figure 71 and Figure 72). An interaction effect with the text type is considered for the main predictor variables language choice, variety choice and ratio of CMC style markers in the analyses. Furthermore, the gender-related differences and the difference in text frequency per user is considered when reasonable.

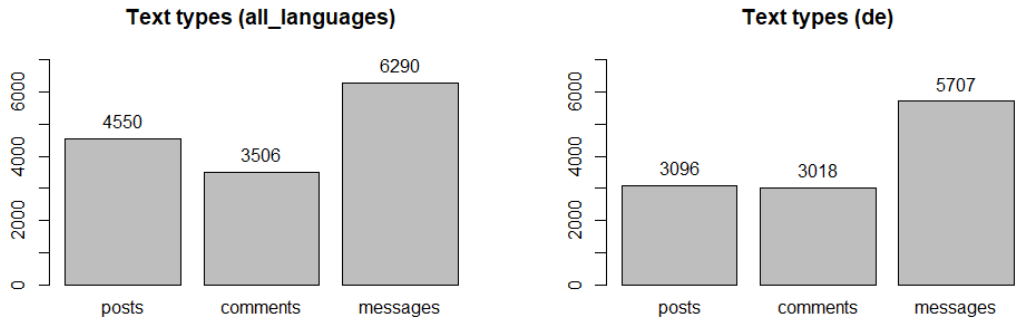


Figure 71: Texts per text type in the corpus subsets.

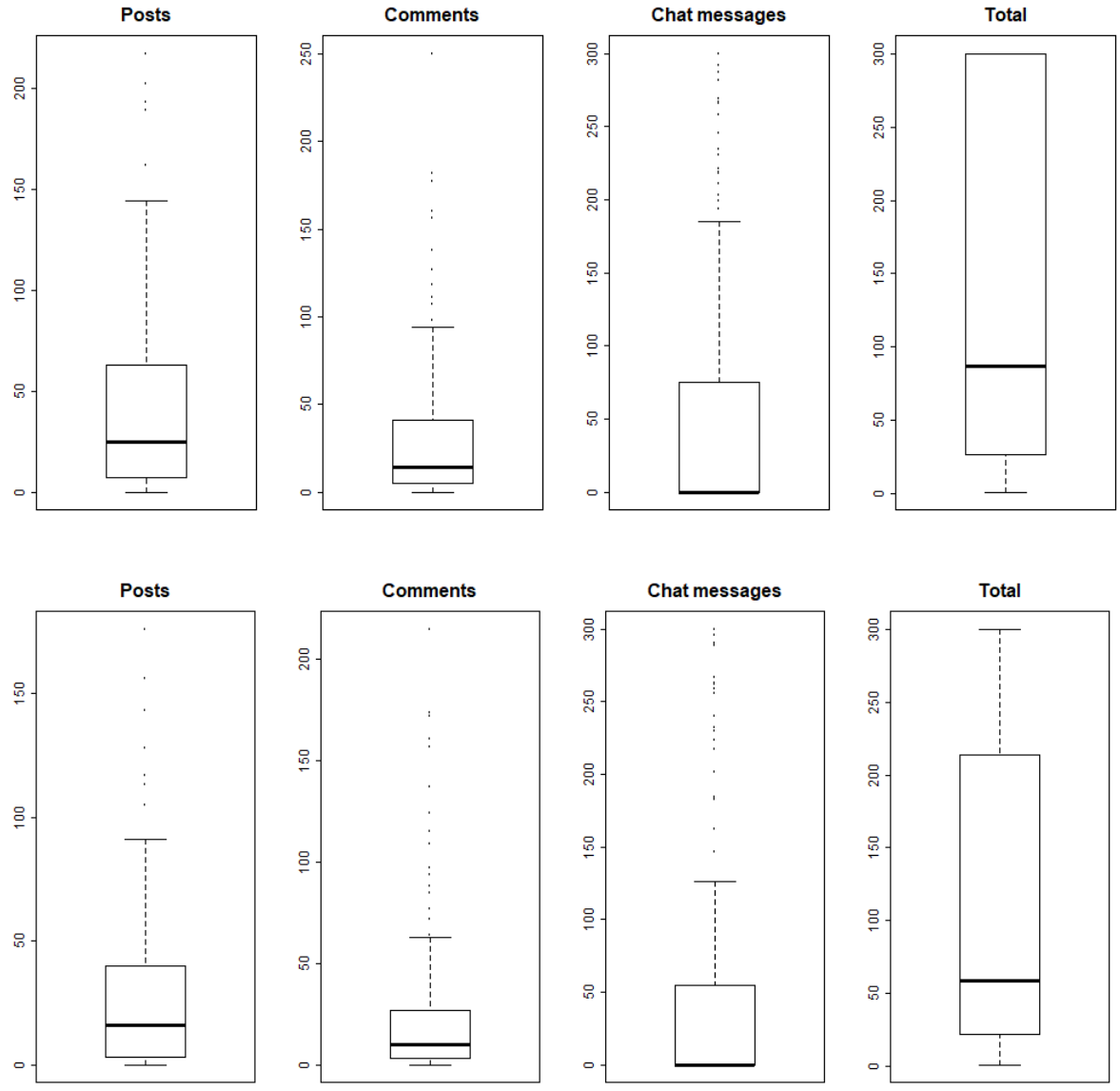


Figure 72: Dispersion of posts, comments, chat messages and total amount of messages per use in all_languages (above) and de (below).

18.3 Methodology

Contrary to the explorative approach in the first corpus study of this thesis, where a high number of potentially related features were investigated, the current study focuses on a limited set of research questions, which are based on existing studies. The research questions are investigated using predictive modelling and machine learning (in particular methods used for age prediction tasks in computational sociolinguistics).

Section 19 is modelled as a binary classification task, while the other experiments are based on multi-class classification. All models are trained and evaluated on three different algorithms:

1) generalized linear models (glm)
2) conditional inference trees (ctree)
3) random forest models (randomForest)

The feature sets used for the study depend on the specific research question (see descriptions in the respective sections) and are designed to give interpretable insights based on the models. Possible confounding factors (e.g. text type, gender and texting frequency of the user) are controlled whenever reasonable. Interaction terms are inserted manually to all predictors for glm and modelled implicitly by adding them as additional predictors in the decision tree-based models ctree and randomForest[172].

After training the classifiers using the methodology laid out by Daelemans et al. (1997), their validity is evaluated by comparing their predictive performance (accuracy) with a baseline that would be achieved in the case that the dependent and independent variables were unrelated (i.e. majority baseline, see section 4.3.3.2). Subsequently, the different (valid) models are compared with each other, evaluating their predictive and explanatory power.

All models are trained and evaluated in R using `glm` (`glmer`), `multinom` (`nnet`), `lme4`, `ctree` (`partykit`) and `randomForest` and initialized with default parameters. I used 10-fold CV to estimate model accuracies for the glm and the ctree models, and OOB error for random forest models (see section 4.3.3.2 on model evaluation). Model performances are compared to each other (or to the baseline) with a binomial test for the distribution of prediction accuracy using a confidence level of alpha = 0.01 (cf. Gries, 2013).

Finally, the model internals are interpreted using variable importance measures for the random forest model, effects plots for the glm model and tree visualizations for the ctree model. For glm models, the regression coefficients, z-values and p-values for the individual predictor levels are observed and droppable predictors are inspected using the `drop1` function[173]. Moreover, for section 4 the *open-vocabulary approach* proposed by Schwartz et al. (2013) is used to explore the previously untouched lexical dimension of age-related language.

[172] Although Gries (Forthc.) points out possible problems with this approach, the method worked well for this data. Prediction results for models that explicitly integrated interaction terms also for the conditional inference tree models (as proposed by Gries, 2019) did not yield better results.
[173] The function reports on the predictors which cause the model performance to drop significantly if they are left out.

19 Testing a simple hypothesis: Digital natives vs. digital immigrants (RQ 1)

Research question 1 discusses how to test an existing (linguistic) theory with data science methods. It takes up on Prensky's theory of the *digital native* (Prensky, 2001, 2009), who is believed to show a completely different behaviour in communication than his older counterpart the d*igital immigrant* (see also Bayne & Ross, 2011; Prensky, 2009; Thornham & McFarlane, 2011). The digital native is a person who was born and raised in the digital era, growing up with computers and other digital devices in his or her environment. Prensky claimed that just the presence of digital media during childhood changes peoples' brains and attitudes, made visible in different language use and communication habits (Prensky, 2001) and an increased digital literacy (Prensky, 2009). The theory gained a lot of media attention and was soon afterwards discussed and frequently also criticized by various researchers from sociology, pedagogy, discourse studies and others. However, linguistic studies on the topic are rare. In Frey and Glaznieks (2018) a brief attempt at a corpus linguistic investigation of the theory was made, exploring the differences in the DiDi corpus by comparing group means and predicting digital natives with a simple decision tree classifier, whose prediction performance is known to be low on observational data because of overfitting (see section 3.5.3). Although the authors found differences in the language use, the analysis was preliminary and did not account for known interaction effects.

This work extends the study of Frey and Glaznieks (2018). It compares different types of state-of-the-art models used for confirmatory settings in corpus linguistics and data science, adds additional analyses, and controls for known confounders.

For a first approach, a chronological splitting criterion at the birth year is used to operationalize the dependent variable. I define a digital native as a person who was born after the year 1980, as in other studies which investigated Prensky's theory. Consequently, everyone born before or in 1980 is considered a digital immigrant (the distribution for this operationalization can be seen in Figure 73).

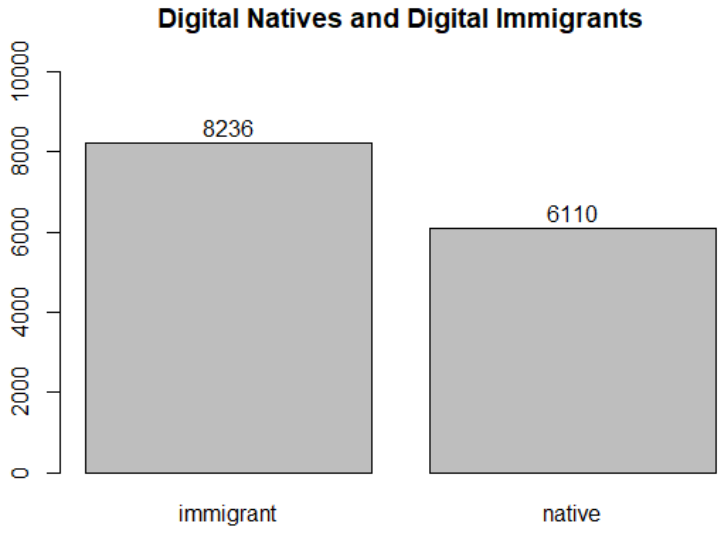


Figure 73: Distribution of digital natives and digital immigrants in the all_languages corpus.

Three variables are considered as predictor variables to measure differences in communication behaviour between digital natives and digital immigrants in the first experiment. In a second step the analysis is extended with additional variables measuring the non-standardness of the texts.

Language choice

The language chosen for a social media post can be a sign of digital literacy, as it can be used for explicit audience design (Androutsopoulos, 2014; Tagg & Seargeant, 2014), conscious identity management (Androutsopoulos et al., 2013; Bolander, 2017; Leppänen, 2012; Schreiber, 2015), and affiliation to group standards (Morel & Pekarek-Doehler, 2013; Stæhr, 2015). Studies on *networked multilingualism* (Androutsopoulos, 2013a) and digital *multilingual repertoires* (Jonsson & Muhonen, 2014) have thus frequently been related to young users, even in non-multilingual areas. I therefore consider this variable to be relevant for the study of age-related language in South Tyrolean social media.

In this study, language choice is analysed using a simplified version of the language annotations in the DiDi corpus. The corpus was labelled semi-automatically with the prevalent language of each text (Frey et al., 2015). The labels fell into one of the eight categories: German, Italian, English, Spanish, French, Portuguese, any other language, or International (i.e. non-language or non-identifiable words because of equal spelling of words in more than one language, e.g. emoticons, links, numbers, '*uh*', '*hi*', '*super*', etc.). They are based on a preliminary labelling done with the language identification tool langid.py (Lui & Baldwin, 2012) that was corrected when one of the following cases was true:

- The language identification tool did not exceed a threshold confidence of 0.8.
- The identified language was not one of the aforementioned languages.
- The text had less than 30 characters.

For the current study, Spanish, French and Portuguese texts as well as the residual category '*other languages'* are excluded as the total number of texts belonging to these categories was very low (< 1%).

The resulting distribution of languages in the `all_languages` corpus subset is shown in Figure 74.

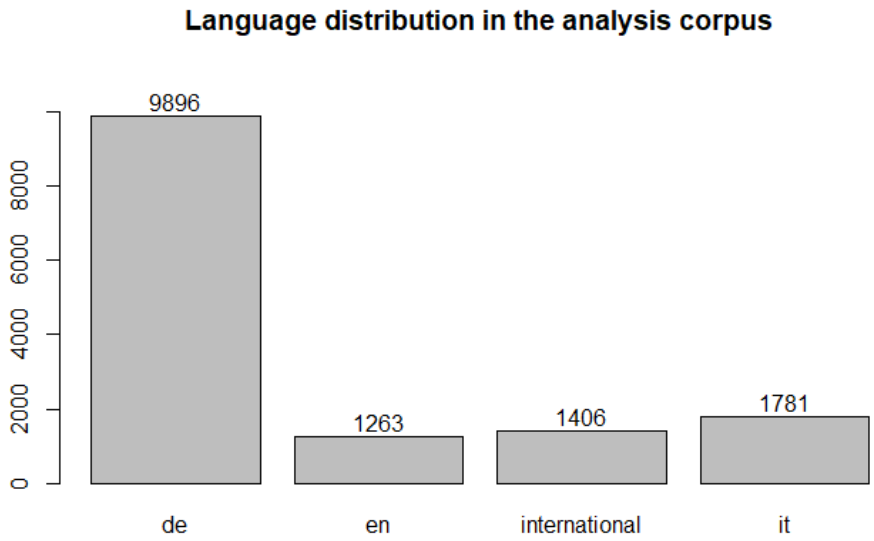


Figure 74: Distribution of languages in the all_languages corpus subset.

Variety choice

Previous studies have observed that dialectal texts are a prominent feature of South Tyrolean CMC (Glaznieks & Stemle, 2014) and SMS texting (Huber, 2013)[174]. The German part of the DiDi corpus contains a high proportion of dialectal texts. Around 40% of the texts show various signs of a South Tyrolean dialect variety (Frey et al., 2015), indicated by a high ratio of non-standard spellings due to phonological spelling of dialectal sounds, as well as by dialect lexemes (‘*sel*’, ‘*olm*’, ‘*hem*’, ‘*ingaling*’) without corresponding standard spelling and by typical character combinations (e.g. ‘*ua*’ as in *‘guat’*, ‘*oa*’ as in *‘koan’*, ‘*gg*’ as in ‘*brugg*’ or ‘*ea*’ as in *‘geaht’*). In comparison, about another third of the texts are clearly non-dialectal (texts that were longer than 30 characters but had almost no non-standard spellings[175]).

This study uses dialect annotations made for the German texts in the DiDi corpus in order to analyse the relationship between age and dialectal texts. As only the German texts contain an annotation for the dialect (i.e. language variety), the two variables can be joined, dividing the German texts into three categories: ‘*de dialect*’, ‘*de non-dialect*’ and ‘*de undef*’ (undefined) (see Figure 75). In the following, the variable names `language` and `dialect` refer to language and language variety distinction separately, while `variety` refers to the joined variable of languages and language varieties.

[174] Similar to other multilingual areas with high-prestige dialect varieties such as in Switzerland (Siebenhaar, 2008), Belgium (Peersman et al., 2016), or the Netherlands (Nguyen, 2017).
[175] The rest of the texts could not be classified straightforwardly and have thus been assigned to a third, mixed or ‘*undefined*’ variety.

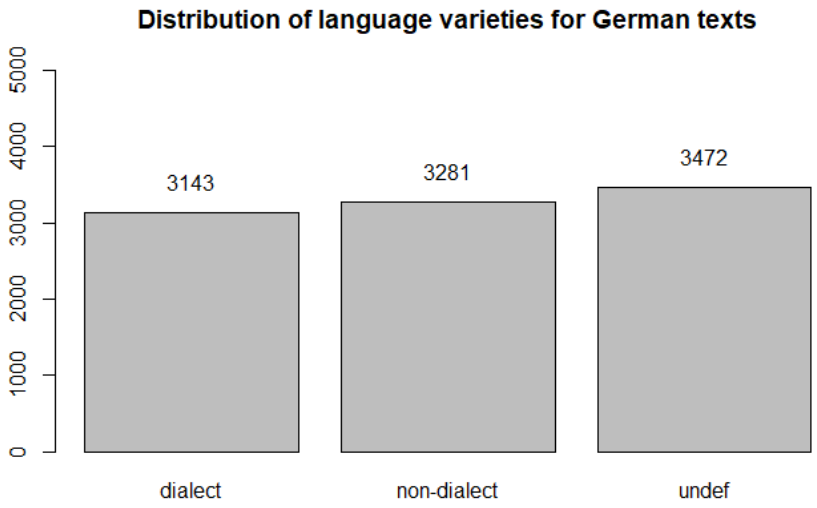


Figure 75: Distribution of language varieties for German text in the all_languages corpus subset

CMC-specific style

The use of CMC-specific style is one of the first and most researched topics in (age-related) CMC research. It is often stated that *netspeak* phenomena like emoticons, emojis, the use of hyperlinks and user-references (@ username), character flooding or iterated punctuation marks[176] are related to young people's changing media habits (Boyd, 2008; Haas et al., 2011). Peersman et al. (2016) for example investigate this relationship by explicitly analysing CMC style features as one type of non-standardness in CMC. While Hovy et al. (2015) and Danet and Herring (2007) found emoticons to be more present in younger writers of CMC, Spina noticed that younger writers also tend to replace punctuation marks more frequently than older users by the simple use of an emoticon.

As the DiDi corpus already contains annotations for CMC-specific style markers, the presence and ratio of CMC-specific tokens in the text is also used as a predictor to distinguish between digital natives and digital immigrants (see Figure 76 for the distribution of this variable).

[176] See e.g. Crystal (2011) or Storrer (2000) (for the German context) for discussions and examples of such phenomena.

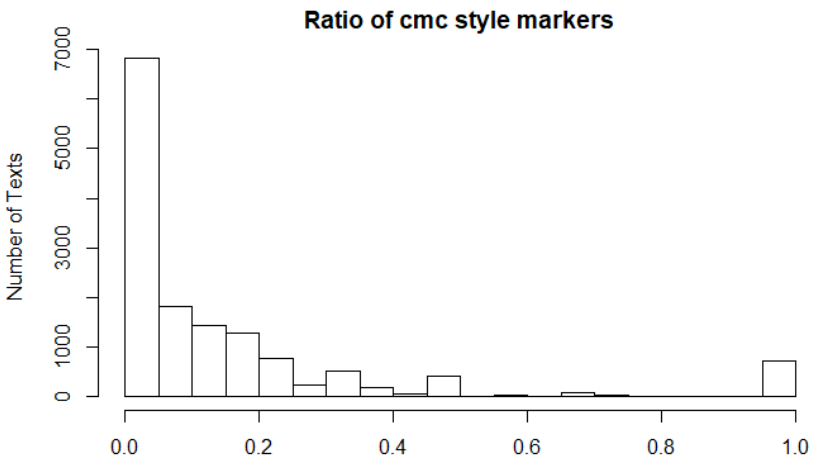


Figure 76: Histogram for ratio of CMC style markers.

Potential confounders

There is a significant association between digital natives and gender as well as between digital natives and the texting frequency (chi-squared test for independence, p-value < 0.05). However, in both cases, the actual effect size is very low (Cramer's V of 0.03 for gender and 0.05 for texting frequency). However, the association for the third expected confounder text type (post, comment or chat message) is significant and had an effect size of 0.23 for Cramer's V. This study thus concentrates on interactions between the main predictor variables and the text type and only controls the other confounders cursorily.

Using different algorithms for predictive modelling, I train binary classification systems to distinguish texts written by digital natives from those written by digital immigrants, considering possibly different effects for posts, comments and chat messages (`type`)[177]. The prediction results show an average prediction performance of 0.7 (see Table 52) and are significantly above the baseline of 0.57 (p-value < 0.01, binomial test) for each of the models.

Baseline	glm	ctree	randomForest
0.574	0.705	0.705	0.717

Table 52: Predicting digital natives.

The regression coefficients and significance of predictor variables for this updated model can be seen in Model output 12. We see that most of the predictor variables are significantly different from 0. Only for the use of Italian, international texts, and a not further classifiable

[177] Interaction effects for gender and texting frequency have been tested as well. However, none of the models achieved higher prediction performance by adding the gender variable, which is why the results reported here were calculated for the simpler models without gender interaction effects. A chi-squared test for droppable predictors on the glm model revealed that there is no significant effect for the interaction between the ratio of CMC style markers and the text type in the glm model. Consequently, the interaction term has been removed from the model.

variety of German (de undef) in comments is there either no effect or not enough evidence in the data.

```
glm(formula = native ~ variety + variety:type + cmc_ratio,
 family = binomial, data = digital_native)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1679  -0.9528  -0.4752   0.9341   2.5320

Coefficients:
                          Estimate Std. Error   z value   Pr(>|z|)
(Intercept)              -0.95567    0.06144   -15.554    < 2e-16 ***
de undef                  0.25883    0.09380     2.760    0.00579 **
de dialect                1.60573    0.11271    14.246    < 2e-16 ***
international             0.29336    0.11593     2.530    0.01139 *
en                        0.61749    0.09551     6.465   1.01e-10 ***
it                       -1.16871    0.14716    -7.942   1.99e-15 ***
cmc_ratio                 1.41260    0.09388    15.047    < 2e-16 ***
de non-dialect:comments   0.25412    0.10090     2.519    0.01178 *
de undef:comments         1.30017    0.09939    13.082    < 2e-16 ***
de dialect:comments       0.46942    0.12016     3.907   9.35e-05 ***
international:comments    1.26173    0.16918     7.458   8.79e-14 ***
en:comments               1.05330    0.19901     5.293   1.21e-07 ***
it:comments               0.31251    0.20603     1.517    0.12932
de non-dialect:messages -1.08105    0.10318   -10.478    < 2e-16 ***
de undef:messages         0.14246    0.08756     1.627    0.10375
de dialect:messages      -0.66039    0.10789    -6.121   9.30e-10 ***
international:messages    0.10431    0.12852     0.812    0.41701
en:messages               0.25414    0.12889     1.972    0.04863 *
it:messages              -1.03983    0.20712    -5.020   5.16e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19572  on 14345  degrees of freedom
Residual deviance: 16220  on 14327  degrees of freedom
AIC: 16258

Number of Fisher Scoring iterations: 5
```

Model output 12: Final glm model for predicting digital natives.

The chi-squared tests for droppable predictors (Model output 13) on the updated glm model shows that both the ratio of CMC style markers in the text and the language variety are significant predictors of digital natives. While there is no interaction between the text type and the relative amount of CMC style markers, the variety choice, however, is different depending on the type of the text.

```
Single term deletions

Model:
native ~ variety + variety:type + cmc_ratio

                  Df   Deviance     AIC      LRT    Pr(>Chi)
<none>                   16220   16258
cmc_ratio          1      16458   16494   238.01   < 2.2e-16 ***
variety:type      12      16951   16965   731.75   < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model output 13: Drop1 output for final glm model (digital natives)

Moreover, the effects plots (partial dependence plots) for the model show that the probability of the text being written by a digital native increases the higher the ratio of CMC style markers

is in the text (see Figure 77). The South Tyrolean dialect is, in general, a good indicator for a digital native writer in all text types (although the effect is strongest for comments and lowest for chat messages). Italian, on the contrary, indicates a digital immigrant writer (see Figure 78). Clearly non-dialectal texts on the other hand have a higher probability of being from a digital immigrant writer (especially for chat messages, while less so for the semi-public communication on the Facebook wall). For the other varieties, Figure 78 shows a clear difference for text types, reporting that comments written in English, undefined German or international language are indicating digital native writers. However, these effects have to be taken with a pinch of salt, as the regression coefficients for these predictor levels are not significant and the effect might be due to too little data.

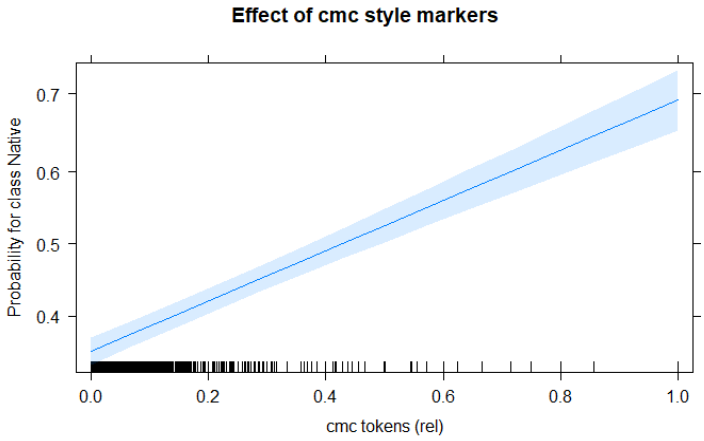


Figure 77: Variable effect for ratio of CMC style markers in text.

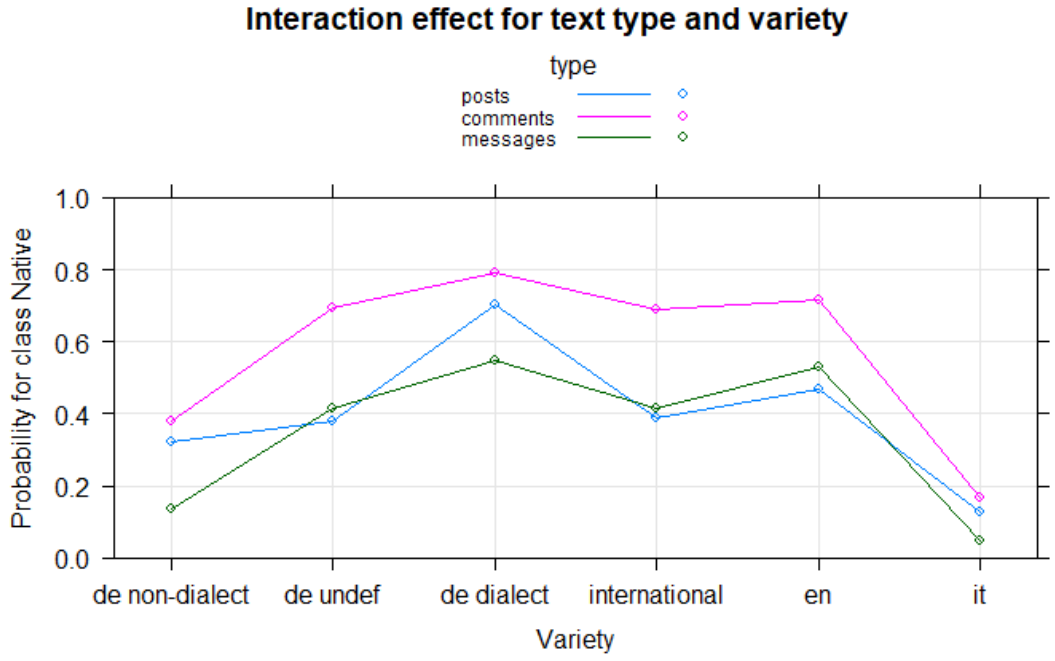


Figure 78: Interaction effects for text type interaction on the effect of language and language variety on the probability of predicting a digital native.

The ctree decision tree model on the other hand allows the internal tree structure to be interpreted. The decision tree visualization of the ctree model (see Figure 80) also shows that

German non-dialectal texts and Italian texts are predicted to be written by digital immigrants (unless they are comments written in standard-oriented German and containing a relatively high ratio of CMC style markers per token). The use of any other variety in comments is interpreted as a sign of digital natives. For posts, only the German dialect clearly indicates digital natives, while the ratio of CMC style markers is crucial for the systems predictions for the other varieties. Chat messages that are not written in standard-oriented German or Italian are predicted to be written by digital natives only if the ratio of CMC style markers is relatively high. Compared to the glm, the decision tree shows the effect of CMC style markers, varieties and text types in combination.

In order to interpret the random forest model, additional interpretation methods are needed. The interaction plot generated with the randomForestExplainer package for R shows variable combinations that occur most commonly at early splitting points of the individual trees (see Figure 79).

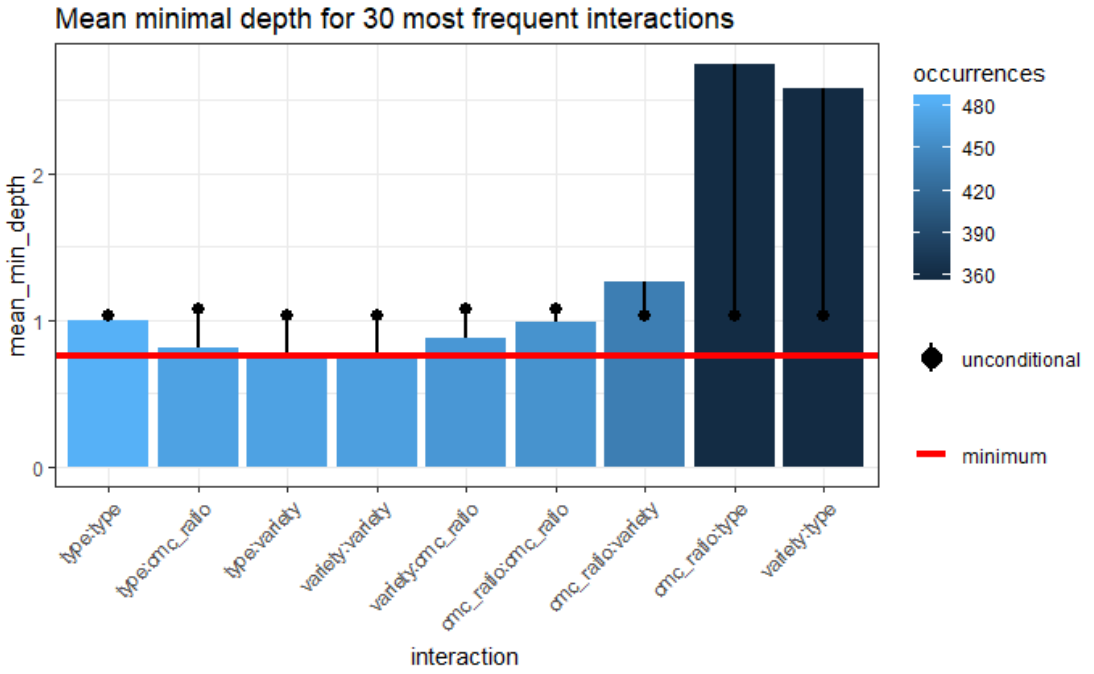


Figure 79: Combinatorial effects for predicting digital natives with a random forest model.

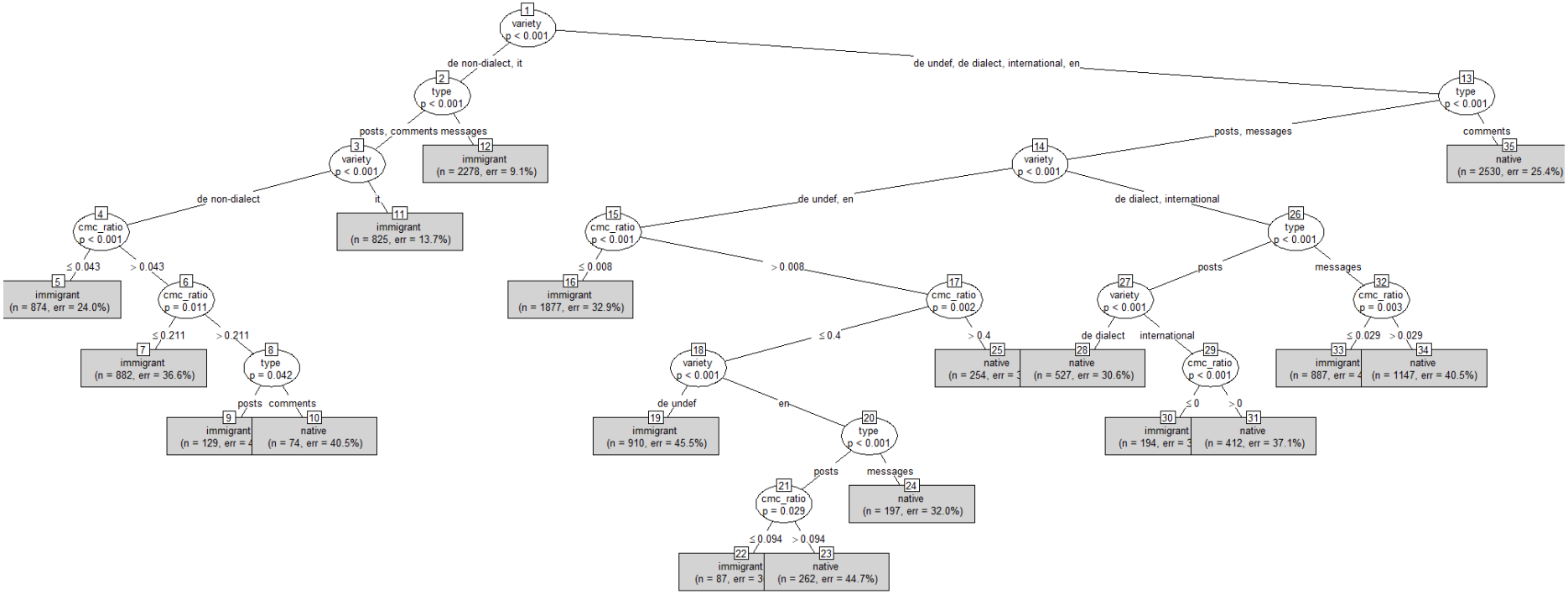


Figure 80: Decision tree visualization for classifying digital natives and digital immigrants.

20 More detailed analyses: Age prediction

In the following, the concept of age-related language in CMC is analysed in more detail. After the previous section showed that it is possible to predict digital natives and digital immigrants by their choice of language, variety, and their use of CMC-specific style, this section extends the analysis by predicting more than two age groups. For this purpose, the dataset has been split into three similarly-distributed writer age groups (see Figure 81):

1. Younger writers under 30
2. Writers between 30 and 49
3. Older writers from 50 years onwards

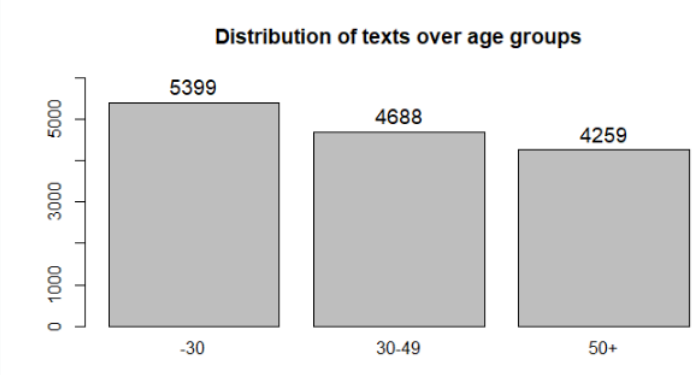


Figure 81: Distribution of texts over age groups.

The age categorization above is used for all the following experiments in this section. While section 20.1 evaluates if it is possible to distinguish older writers from younger ones automatically, section 20.2 evaluates how important non-standard spellings are for the prediction of age groups. In section 20.3, the age prediction experiments are elaborated by adding additional variables (specifically stylometry and unigram features) that have been used in other studies and evaluating whether the variables give further insights.

20.1 Predicting multiple age groups (RQ 2)

Contrary to the binary prediction performed in section 19, I now try to distinguish three age groups by training multi-class prediction models. The baseline for prediction accuracy on this dataset is 37.6% according to the majority class distribution. Table 53 shows 10-fold CV and OOB estimates for the glm, ctree and randomForest models for the multi-class prediction on the `all_languages` corpus subset. All models are significantly above the baseline (p-value < 0.001, binomial test) and are thus able to extract meaningful information from the variables.

Baseline	glm	ctree	randomForest
0.376	0.523	0.530	0.530

Table 53: Classification results for the prediction of three age groups using language and variety choice and CMC-specific style.

All hypothesised predictor variables, including their interaction with the text type, contribute significantly to the model performance for glm (see Model output 14).

```
Analysis of Deviance Table (Type III tests)

Response: age3groups
                  LR Chisq   Df   Pr(>Chisq)
variety             651.77   10    < 2.2e-16 ***
cmc_ratio           176.15    2    < 2.2e-16 ***
cmc_ratio:type       35.33    4    3.975e-07 ***
variety:type        765.48   20    < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model output 14: Type III Anova showing significance of factors in the glm model.

The results for predicting the three age groups on the basis of language choice, variety choice, and CMC-specific style are, in total, good enough to reject the hypothesis that the model's predictions are random (see section 20.1). However, the prediction accuracy for the group of older writers is visibly lower than for the other writers in all three models. The group accuracies reported in the confusion matrices of the three models (see Table 54) show that the accuracy rates correctly predicting the older group (i.e. precision) were lower than for the other groups (in average 38% compared to an average of 75% for the younger writers). Moreover, older writers have been confused with middle-aged and younger ones by the algorithm (i.e. recall) in more than half of the cases. In addition, the confusion matrices do not exclusively show misclassifications for adjacent age groups. Older writers are confused with younger writers and writers from the middle age group (and vice versa). We can therefore conclude that the linguistic difference in language and variety choice as well as in CMC-specific style is less clear between writers in their 30s and 40s and writers over 50.

glm				
target	*-30*	*30-49*	*50+*	*accuracy*
-30	3789	990	620	0.702
30-49	1525	2390	773	0.510
50+	1254	1741	1264	0.294

ctree				
target	*-30*	*30-49*	*50+*	*accuracy*
-30	4096	672	631	0.759
30-49	1690	1793	1205	0.382
50+	1398	1027	1834	0.431

randomForest				
target	*-30*	*30-49*	*50+*	*accuracy*
-30	4225	618	556	0.783
30-49	1883	1660	1145	0.354
50+	1514	1025	1720	0.404

Table 54: Error analysis for predicting three age groups.

20.2 Dialect or non-standardness? (RQ 3)

We observed in section 19 that the probability of predicting a digital native is higher for dialectal texts. Similarly, when inspecting the model's prediction probabilities for the three age groups, the model reports a higher probability that a dialectal text is classified to be written by a younger writer (see Figure 82).

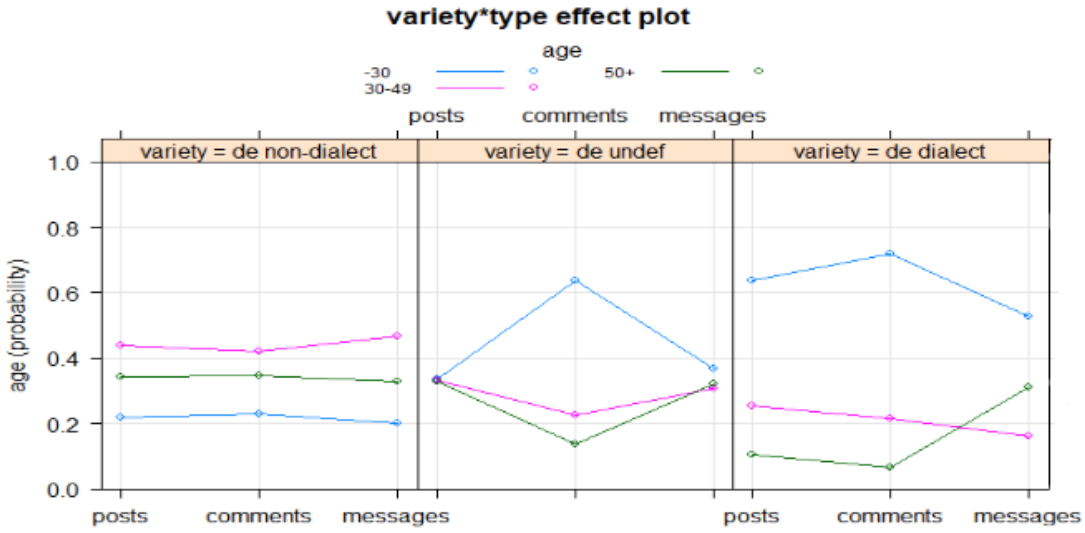


Figure 82: Effect of the chosen variety of German on the prediction probabilities for the three age groups.

However, it remains unclear whether the effect of the variety variable on the model's predictions, namely that dialectal texts have a higher probability of predicting a young writer, is related to non-standardness as an expression of individuality performed by younger and older writers as assumed by Peersman et al. (2016) or to the value of the regional variety in expressing local identities, as one could hypothesise according to theories of *translocality* (Leppänen, 2012; Leppänen et al., 2009). While previous studies have not differentiated between these two aspects, treating local varieties and standardness as equal (e.g. Hilte et al., 2018; Peersman et al., 2016), this section compares the effects of choosing the regional variety with the effect of more general variables of non-standardness on the predictive performance of an age prediction model that was trained on the '*de*' subset (only German texts) of the DiDi data.

To measure non-standardness, I thus use two further variables in addition to the variety variable: the presence of non-standard spellings in the data (`has_nonstd`) and the ratio of non-standard spellings to standard spellings of tokens in the text (`nonstd_ratio`)[178]. Contrary to the annotations for the South Tyrolean dialect, these variables measure non-standardness independently of whether it derives from dialectal spelling or other types of non-standardness (e.g. missing capitalization, wrong hyphenation and word compounding, non-codified abbreviations, character flooding or elision of characters and syllables) and is calculated using the annotations for standardized spellings provided for each out-of-dictionary (i.e. non-standard)

[178] Of course, the variables for non-standardness and dialect can overlap as dialectal texts often contain a certain amount of (dialectal) non-standard spellings to represent the sounds of the usually spoken variety.

token in the German texts in the corpus. Figure 83 shows the distribution of these variables in the de corpus subset.

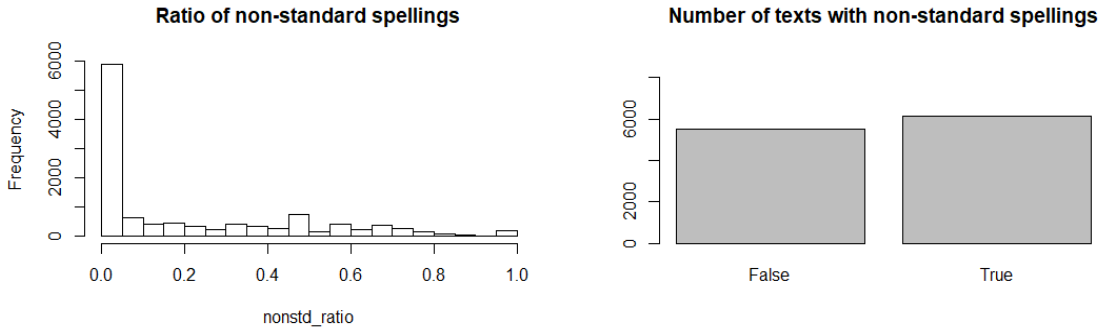


Figure 83: Distributions for ratio and presence of non-standard spellings in the de corpus subset.

The results in Table 55 compare the 10-fold CV/OOB estimates of prediction accuracy for the variable '*variety*' with those for the features of non-standardness. The ratio of CMC style markers that was previously found to be significant remains in the model. Additionally, interaction terms for the text type have been inserted for all predictor variables. A third set of models controls whether both the local dialect as well as non-standardness contribute individually to the successful prediction of the three age groups by testing if the overall accuracy of the original model increases significantly when the additional variables are added[179].

	glm	ctree	randomForest
dialect	0.499	0.497	0.504
non-standard	0.51*	0.522*	0.52*
dialect + non-standard	0.513	0.517	0.524

Table 55: Prediction performance for models with dialect, non-standardness, or both variables.

The results show that the features of non-standardness predict the age groups significantly better than the feature of variety choice (binomial test, p-value < 0.01 for all three models). Furthermore, adding both features to the model does not predict the age groups better than the non-standardness features alone.

The variable importances reported for the full randomForest model (dialect + non-standard) that give a ranking of variables according to the resulting average decrease in gini impurity also indicate that the ratio of non-standard spellings is more important than the variety choice, whereas simply the presence of non-standard spellings is rated least important (compare Figure 84).

[179] The baseline for this task was slightly higher than for the previous experiments on the *all_languages* corpus subset (0.388). However, all models had a 10-fold CV accuracy that was significantly above the baseline (binomial test, p-value < 0.001).

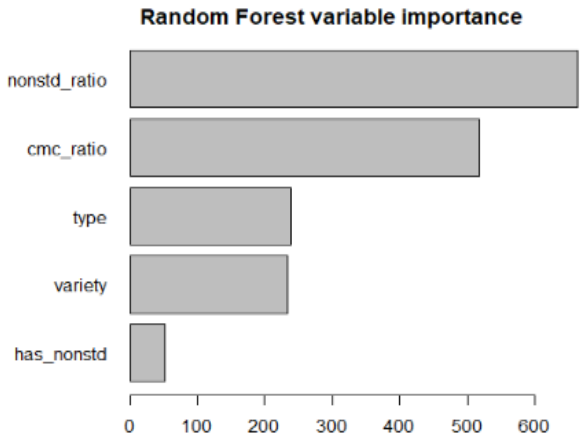


Figure 84: Random forest variable importance for model with dialect and non-standardness features.

The drop1 function output for the full glm model (dialect + non-standard) shows significant main effects for all the assumed variables including their interaction with the text type (see Model output 15) However, as stated before, the model did not achieve significantly better results when evaluated on 10-fold CV prediction accuracy.

```
Analysis of Deviance Table (Type III tests)

Response: age
                 LR Chisq Df Pr(>Chisq)
variety            43.150  4  9.632e-09 ***
cmc_ratio         202.762  2  < 2.2e-16 ***
has_nonstd         36.176  2  1.395e-08 ***
nonstd_ratio       14.706  2  0.0006406 ***
cmc_ratio:type     42.971  4  1.049e-08 ***
variety:type      102.243  8  < 2.2e-16 ***
has_nonstd:type   121.134  4  < 2.2e-16 ***
nonstd_ratio:type  49.664  4  4.243e-10 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model output 15: Significant predictors in the model based on dialect + non-standardness features.

The two effects plots in Figure 85 and Figure 86 show the variable effects for the choice of variety and for the ratio of non-standard spellings in the text for the full glm model. The graphs show a higher probability of predicting a younger writer under 30 for both dialectal texts as well as texts with higher ratios of non-standard spellings. While the division between younger writers and both other age groups is relatively clear for inherently dialogic text types such as comments and chat messages, it is not so clear-cut for wall posts. Given that theories of age-related language in spoken conversation repeatedly described that middle-aged people have a higher tendency to use prestige varieties of the language, while younger and older people have a higher tendency to use vernacular language, we would expect the highest probability of predicting a writer between 30 and 49 years for texts composed in a standard-oriented variety of German, while the probabilities of predicting an older writer over 50 or a younger writer under 30 should be lower. However, this cannot be seen in the models trained on the DiDi data. Here, the older writers over 50 act similar to the writers between 30 and 49 in the private chat messages in terms of variety choice. In the semi-public text types, they use even

less dialect than the writers between 30 and 49. The effects of the presence of non-standard spellings did not show any meaningful difference.

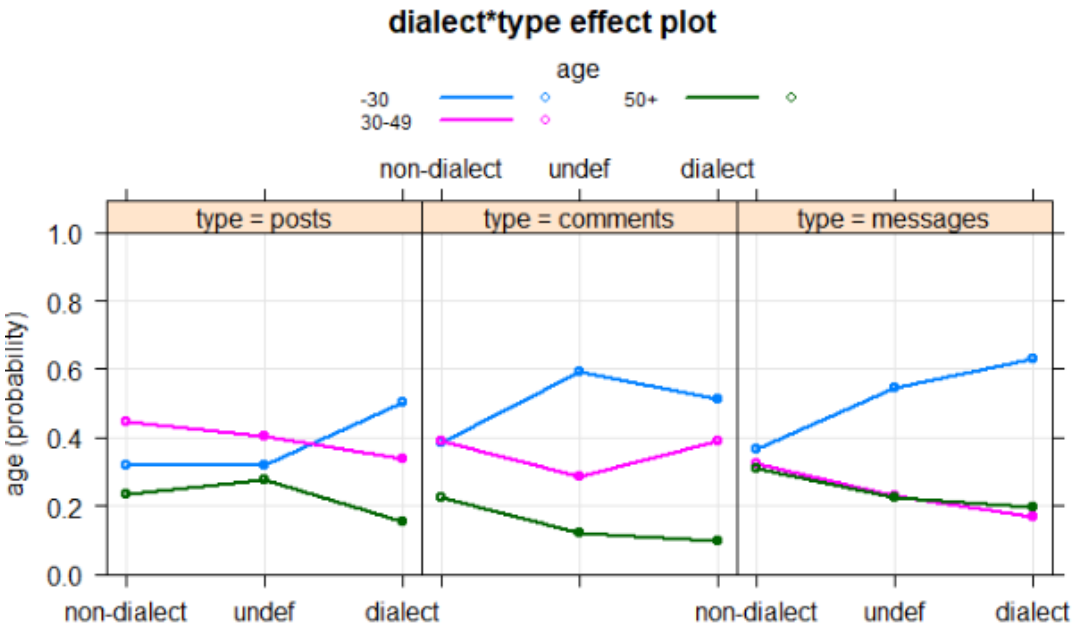


Figure 85: Variable effect of the dialect on predicting three age groups, including its interaction with the text type.

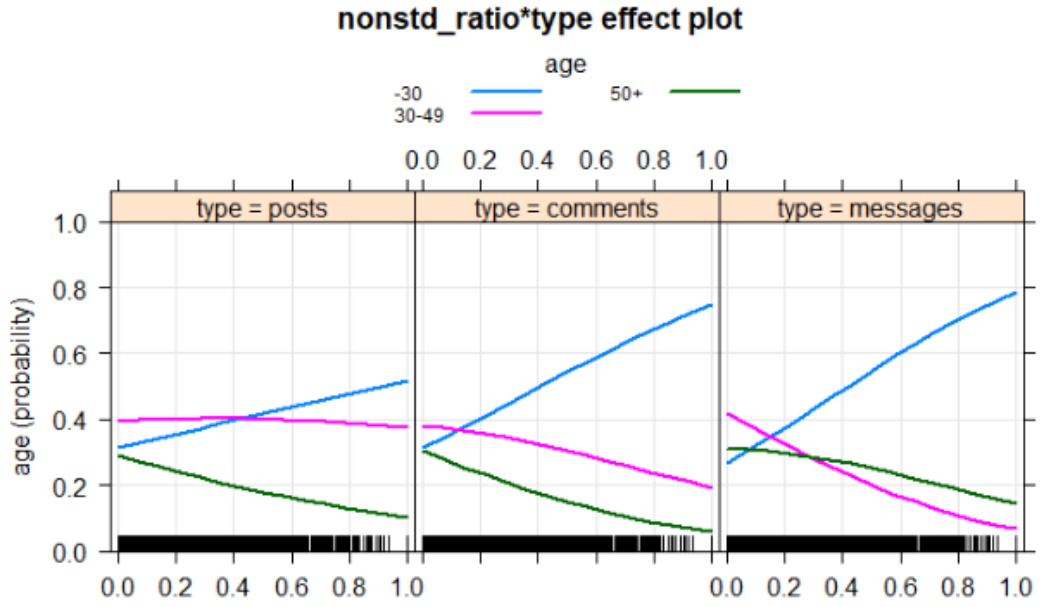


Figure 86: Variable effect of the ratio of non-standard spellings on predicting three age groups, including its interaction with the text type.

20.3 Adding more features (RQ 4)

The experiments in section 19 showed that it is possible to predict (with a prediction performance that exceeds the majority baseline) digital natives in the DiDi corpus based on their choice of language and language variety, and the ratio of CMC style markers in the text.

However, it also indicated that there might be further features of language use that are relevant to predicting age groups and describing age-related language using the corpus data.

In this section the German subset of texts ('*de'*) is used to explore further possible features that contribute to the prediction of the three age groups.

In line with similar research on English data, I test the predictive performance of stylometry features that are frequently mentioned to be relevant for age prediction or deemed relevant for the DiDi corpus (see Table 56).

Variable	**Description**
text_length	Number of tokens in the text
has_anonym anonym_ratio	Presence of anonymized content and ratio of anonymized content to all tokens in the text
has_nonstd nonstd_ratio	Presence of non-standard spellings and ratio of non-standard spellings to all tokens in the text
has_cmc cmc_ratio	Presence of CMC style markers and ratio of CMC style markers to all tokens in the text
has_all_caps	Presence of words written entirely in capitalized letters
has_low_caps	Presence of words written without capitalization although capitalization would be needed to match the standard spelling as well as the ratio of such words per token
has_at	Presence of direct address via @
has_character_flooding	Presence of character flooding in the text
has_emoticon	Presence of emoticons
has_emoji	Presence of emojis
has_hashtag	Presence of hashtags
has_hyperlink	Presence of URLs
has_newline newline_ratio	Presence of newline characters and ratio of newline characters to all tokens in the text
has_punct punct_ratio	Presence of punctuation and ratio of punctuation tokens that are not emoticons or other CMC style markers to all tokens in the text
punct_length_ratio	Average amount of punctuation characters per punctuation token
has_hapax_orig hapax_orig_ratio	Presence and ratio of unique spellings that occur in the text but not in any other text
has_hapax hapax_ratio	Presence and ratio of words that occur in the text but are specific to the writer and are not used by any other writer
has_user_hapax user_hapax_ratio	Presence and ratio of lemmas that occur in the text but are specific to the writer and are not used by any other writer

has_user_hapax_orig user_hapax_orig_ratio	Presence and ratio of spellings that occur in the text but are specific to the writer and are not used by any other writer

Table 56: Stylometry features.

The results displayed in Table 57 show that with the higher number of features, the random-forest classifier achieves better results than the glm and the ctree model (significant at alpha 0.001, binomial test). All models achieve a prediction accuracy above the baseline of a majority class classifier.

Baseline	glm	ctree	randomForest
0.388	0.549	0.536	0.586

Table 57: Classification results for the prediction of the three age groups with an extended feature set.

In addition, the CV/OOB estimates are significantly better than those from the previous experiment which were based on non-standardness, CMC-specific style, and variety choice (binomial test, p-value < 0.001), indicating that the additional features did contribute substantially to predicting the three age groups.

The most important features reported by the built-in variable importance measure for the random forest model are ranked below (Figure 87). The ratio of non-standard spellings (nonstd_ratio) is by far the most important feature for the random forest model. This corresponds to the observations made in the experiments regarding non-standardness and variety choice. The ratio of CMC style markers is ranked fourth, confirming what we observed before. At ranks two and three, however, are other features that were not considered in the previous analysis. The second most important feature for the relatively well-performing random forest model is the ratio of punctuation tokens in the text; the third is the text length, measured by the number of tokens. The fifth best feature is related to punctuation and indicates the average number of punctuation characters per punctuation token in the text. The corpus was tokenized with a CMC-specific tokenizer that would not split various punctuation marks which occurred next to each other without a space (except for those combinations that also exist in standard language). Therefore, the average number of punctuation characters per punctuation token gives a measure of how many times people used repeated punctuation marks like '*!!*', '*!?!*', … or similar[180]. Moreover, the ratio of spellings that are unique within the corpus or specific to the user score relatively high. These rankings thus indicate that there might be further relevant factors which distinguish between age groups in the data.

[180] Emoticons were not considered in this calculation.

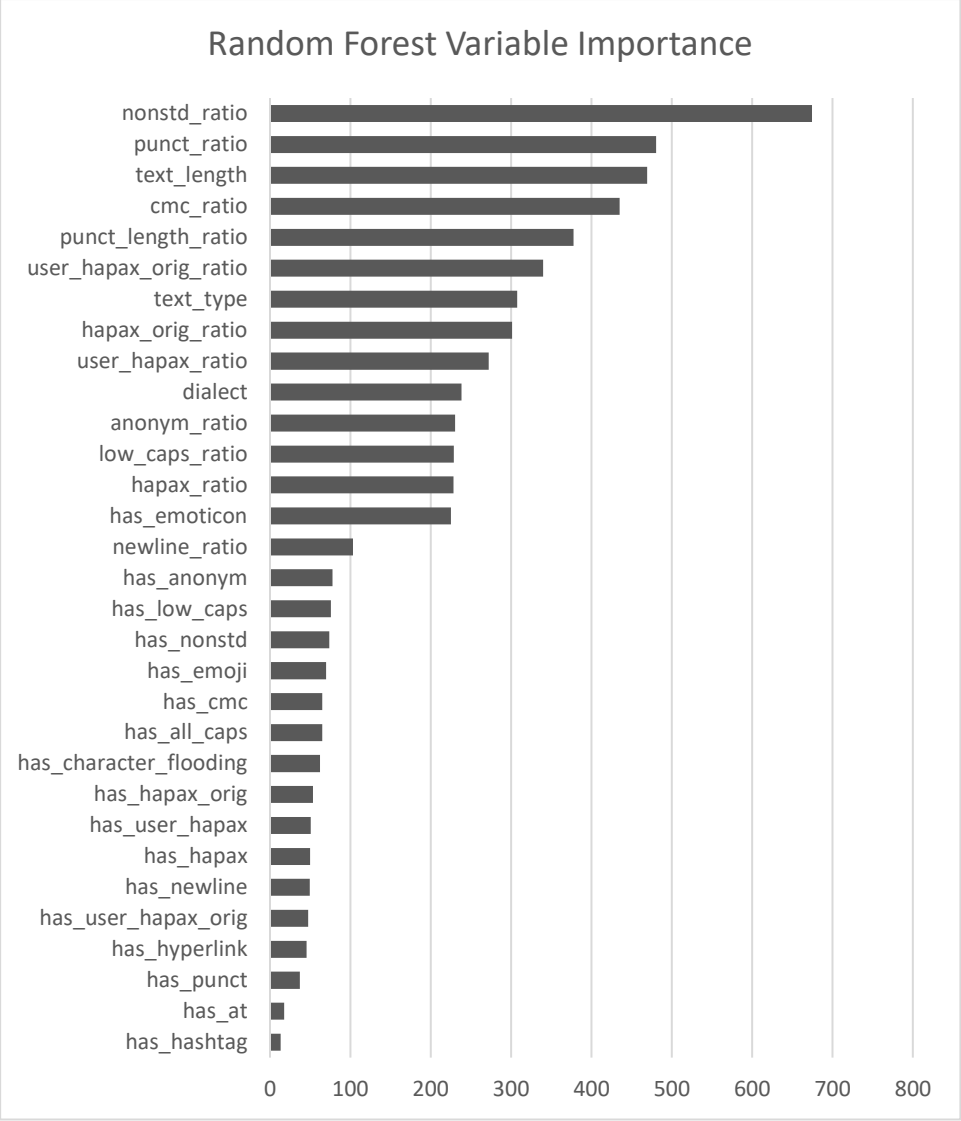


Figure 87: Random forest variable importance for extended feature set.

The glm model reports various significant factors including the previously highly-ranked `punctuation_length_ratio`, i.e. the average number of punctuation characters per punctuation token (analysis of deviance, see Model output 16).

However, some of the highly-ranked features from the random forest model did not significantly contribute to the model's performance in the glm, e.g. the text length and the ratio of punctuation tokens to all the tokens in the text.

```
Analysis of Deviance Table (Type III tests)

Response: age
                       LR Chisq Df Pr(>Chisq)
dialect                  111.36  4  < 2.2e-16 ***
type                     386.00  4  < 2.2e-16 ***
anonym_ratio              12.99  2  0.0015081 **
cmc_ratio                111.59  2  < 2.2e-16 ***
hapax_orig_ratio          18.31  2  0.0001056 ***
hapax_ratio                5.55  2  0.0622589 .
has_all_caps              12.29  2  0.0021427 **
has_anonym                37.74  2  6.386e-09 ***
has_at                    48.93  2  2.372e-11 ***
has_character_flooding    87.04  2  < 2.2e-16 ***
```

```
has_cmc                  25.94  2  2.329e-06 ***
has_emoji               276.22  2  < 2.2e-16 ***
has_emoticon            272.94  2  < 2.2e-16 ***
has_hapax                 0.86  2  0.6502771
has_hapax_orig            3.16  2  0.2054847
has_hashtag              39.08  2  3.269e-09 ***
has_hyperlink           101.85  2  < 2.2e-16 ***
has_low_caps             85.54  2  < 2.2e-16 ***
has_newline               9.46  2  0.0088150 **
has_norm                 17.96  2  0.0001259 ***
has_punct                14.65  2  0.0006577 ***
has_user_hapax            3.80  2  0.1493261
has_user_hapax_orig       5.66  2  0.0589172 .
low_caps_ratio           14.12  2  0.0008588 ***
newline_ratio             3.61  2  0.1646933
norm_ratio              199.97  2  < 2.2e-16 ***
punct_length_ratio      192.18  2  < 2.2e-16 ***
punct_ratio               4.32  2  0.1152938
text_length               3.42  2  0.1807444
user_hapax_orig_ratio    15.86  2  0.0003603 ***
user_hapax_ratio          2.96  2  0.2274263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model output 16: Type III Anova with variable significance for the extended feature set.

Through stepwise backward model selection for the glm model, features such as *text length*, *presence* and *ratio of newlines*, the *ratio of punctuation tokens* and some of the features regarding the presence and ratio of words in the text that occur only once in the corpus or only with one user have been removed, as they did not contribute significantly to the model performance (p-value < 0.01).

21 Interpreting the models to observe differences in language use (RQ 5)

Finally, this (sub)section aims to interpret the models built in order to describe the relations found between individual age groups and their language use. The variable importance measures reported above give a general idea about which factors are at play when predicting age based on the writers' language use. However, in order to compare the different age groups and describe the relationships found, it is also necessary to investigate the individual effects of the variables on the predictions.

By interpreting the best performing prediction models and their global and local variable and interaction effects, it is possible to shed light on the typical language use of writers who are 50+ in SNS, making use of the relatively well-represented older age groups in the DiDi corpus.

The models built in section 20 already showed that it is possible to distinguish not only younger writers (or digital natives) from the rest (see research question 1), but also to distinguish between younger, middle-aged and older writers solely by considering their choice of language and language variety and their use of CMC-specific style markers.

The variable effects reported for the glm model of this simple prediction task (Figure 88) show that the probability of predicting a writer in the younger age group increases with the ratio of CMC style markers, while it usually decreases for the other age groups. However, there is a

significant interaction with the text type. While the effect of CMC style markers on predicting the younger age group is clear for the semi-public posts and comments, it is not as clear for private chat messages.

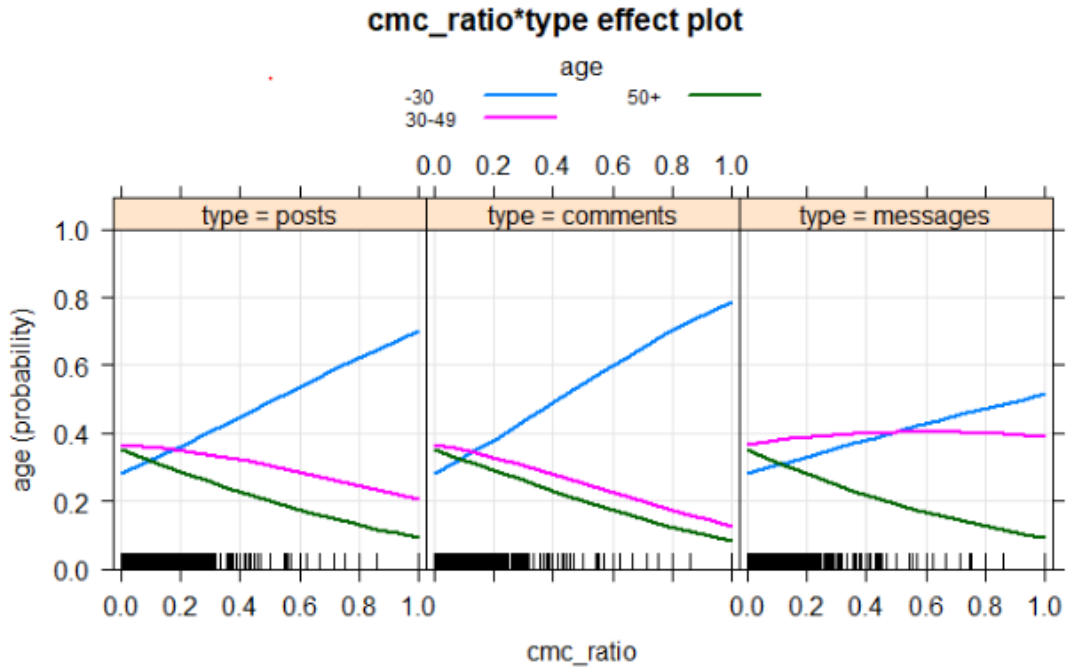


Figure 88: Effects plot for interaction between the ratio of CMC style markers to tokens in the text and the text type.

With regards to the chosen language and language variety, the effects also depend on the text type (see Figure 89). For younger users the prediction probability is generally higher for texts written in the South Tyrolean dialect. However, while users over 50 years of age seem to use less dialect in posts and comments (the semi-public communication modes on Facebook) the probability for predicting the 50+ age group for a dialectal posts is higher for chat messages. It seems that the older users do indeed use dialect in chat messages. Users in their 30s and 40s do not share this behaviour. The probability of the text being written by a writer between 30 and 49 is lowest for dialectal chat messages.

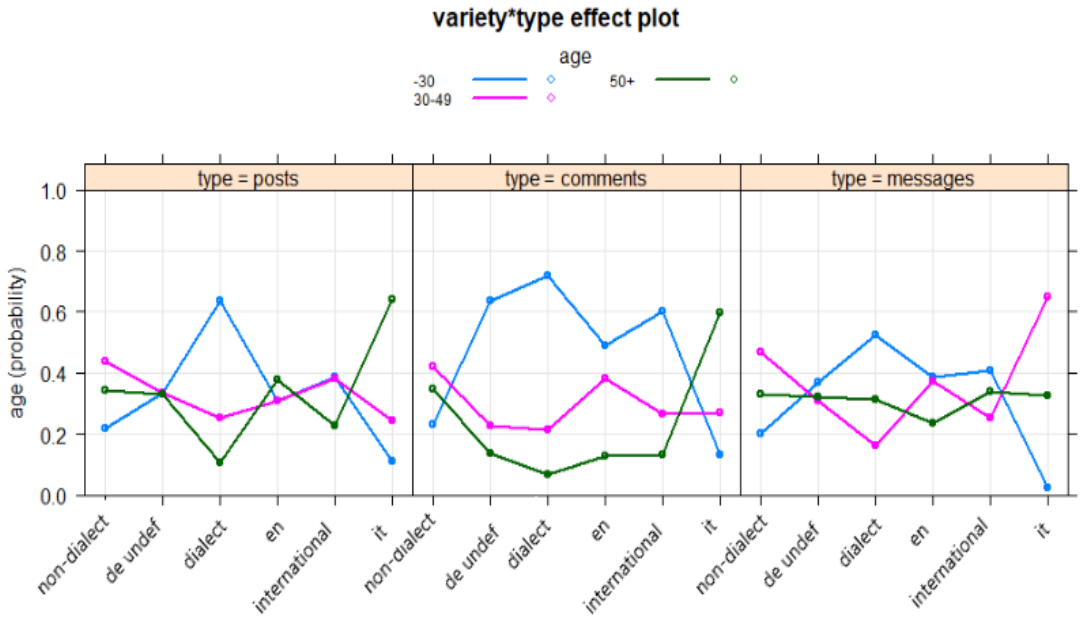


Figure 89: Effects plot for interaction of language and variety choice and text type.

Furthermore, there is an interesting difference in the effect of the Italian language on the prediction probabilities for semi-public and private chat communication. While writers between 30 and 49 use Italian in their private chat messages, the use of the second official language on the semi-public Facebook wall (posts and comments) is a sign of older users. The younger age group of under 30-year olds is rarely predicted for Italian texts.

In total the effects plots illustrate that the probability for predicting a young age under 30 is the lowest out of all three age groups for the two official standard varieties of South Tyrol (Italian and non-dialect German). In this case, the model would usually predict older ages instead.

This is also visible in the tree visualization for the ctree model (Figure 90). Figure 91 shows a slightly different but simpler ctree model. The continuous CMC style marker variable in the original ctree model (Figure 90) creates various splits on different thresholds of the variable. It thereby increases the complexity of tree models and makes them more difficult to read. For easier interpretation, I therefore simplified the CMC style marker variable from a continuous ratio variable to a binary variable indicating the presence or absence of a CMC style marker in the text. The resulting model yielded a lower prediction accuracy than the original model (53% accuracy on 10-fold CV instead of 53,8%), however, the performance decrease was not significant (paired *t*-test, p-value < 0.001) and the size of the tree could be limited from 45 nodes to 29 nodes. Figure 91 shows that standard-oriented German texts are only predicted to be written by under 30 year olds when CMC style markers are present and the text is a comment. The other languages (or varieties) are dominated by under 30 year olds, unless there are no CMC style markers at all (for posts and chat messages).

However, looking at the German text subset (‘*de*’) it is possible to add further detail to the observations. Apart from the effects of variety choice and CMC style marker ratio which are equally visible in the subset containing only German texts (see Figure 92), it is possible to identify further significant main effects through the models built with the extended features

set containing stylometry features[181] (cf. section 20.3). By investigating the variable effects for the variables in the glm model trained on these features (Figure 93), we can identify the following particularities in the data.

The use of various typical features of CMC, e.g. emojis, emoticons, non-standard spellings, or idiosyncratic spellings and uncommon word uses increases the probability that the model predicts a person under 30 years of age. In addition, content directly referencing figures and places in the person's life, which can be identified via the ratio of anonymised content in the texts, were treated as a sign of younger writers. Some features of CMC language, e.g. the various possibilities to link content and people with hashtags, @ signs or hyperlinks, and the ratio of non-capitalized nouns and sentence beginnings are, however, more indicative of a writer between 30 and 49 according to the prediction model. Only the iteration of punctuation characters within one token, e.g. '!!!', or '?!?' and the use of spellings that are unique in the corpus are treated exclusively as signs for older writers.

[181] Note that the features of variety choice, ratio of CMC style markers and the presence and ratio of non-standard spellings are also present in this feature set.

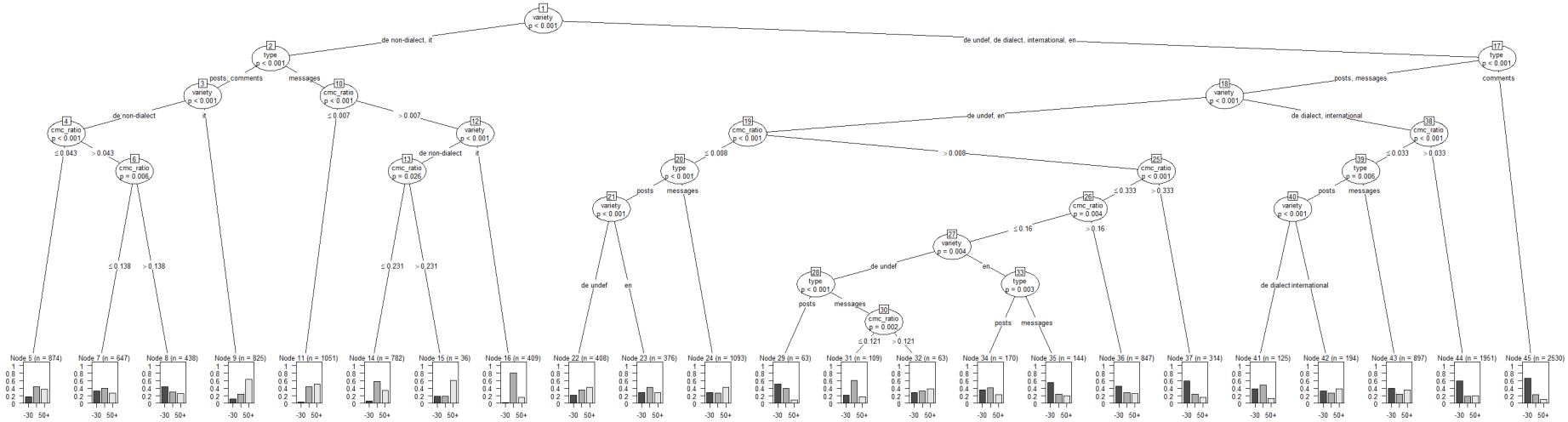


Figure 90: Decision tree visualization (conditional inference tree) for age prediction model.

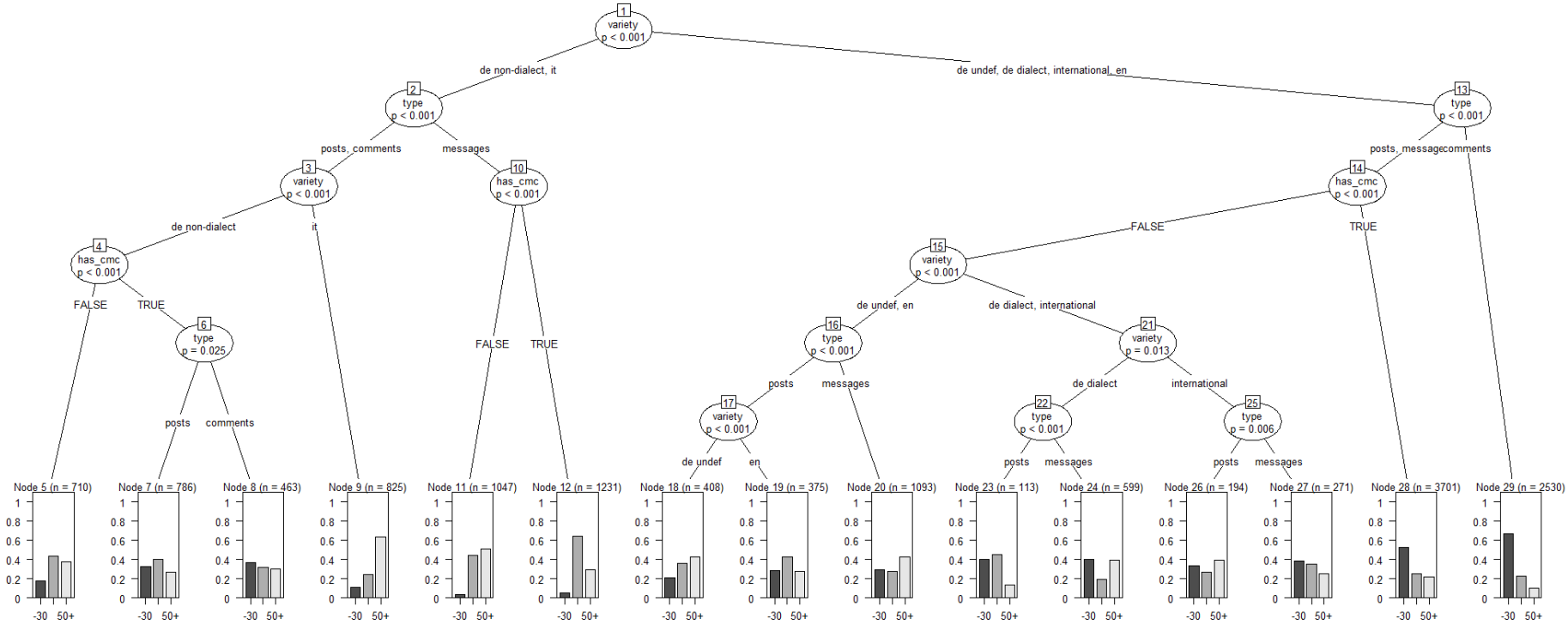


Figure 91: Decision tree visualization (conditional inference tree) for simplified age prediction model that only checks for presence of CMC style markers instead of ratio

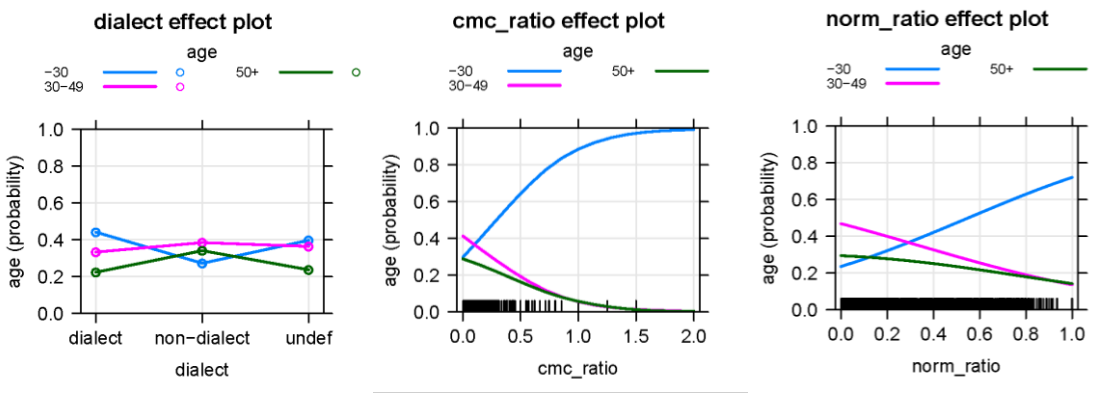
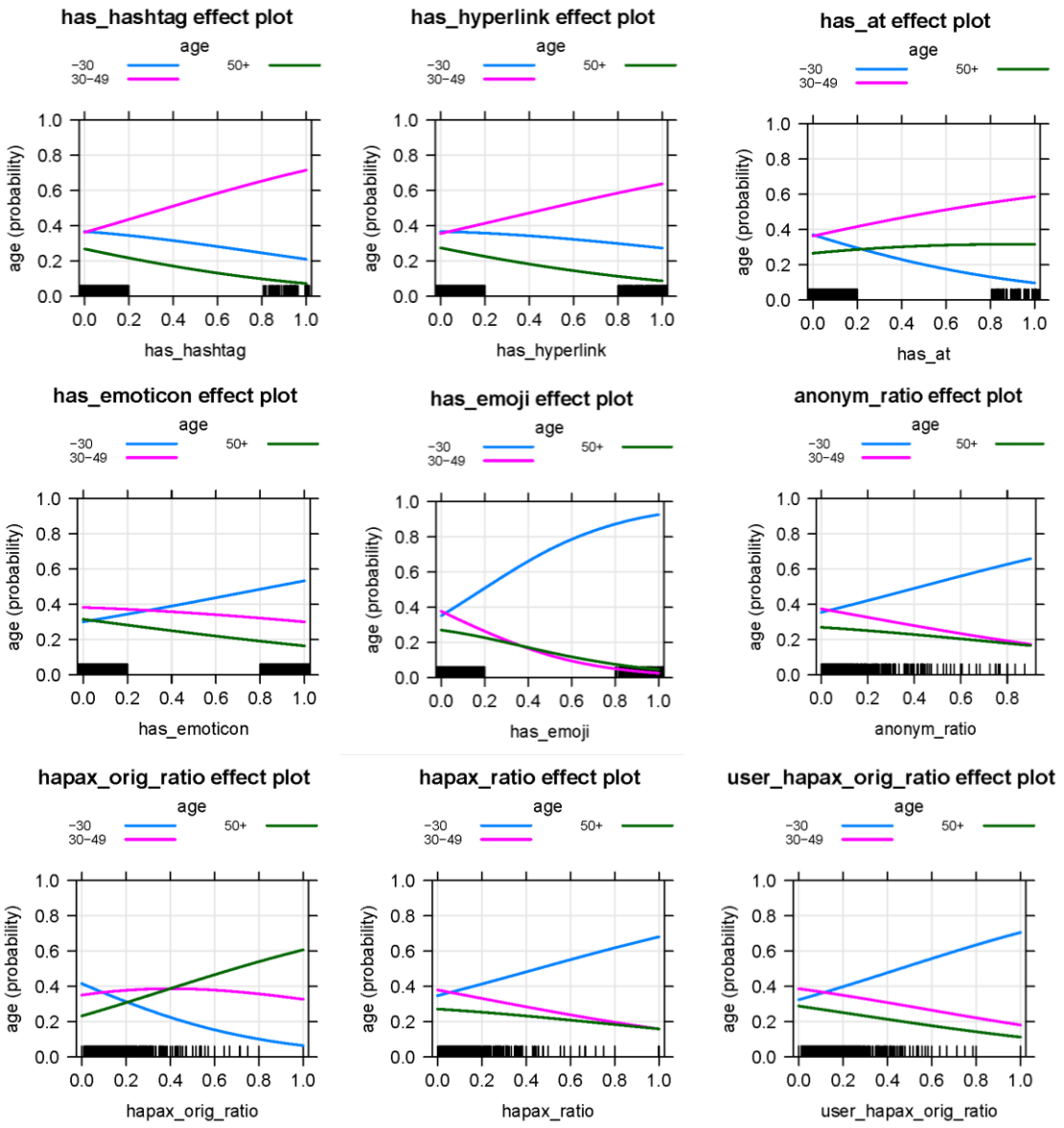


Figure 92: Variable effects for variety choice, ratio of CMC style markers and ratio of non-standardness spellings in the glm model with all stylometry features.



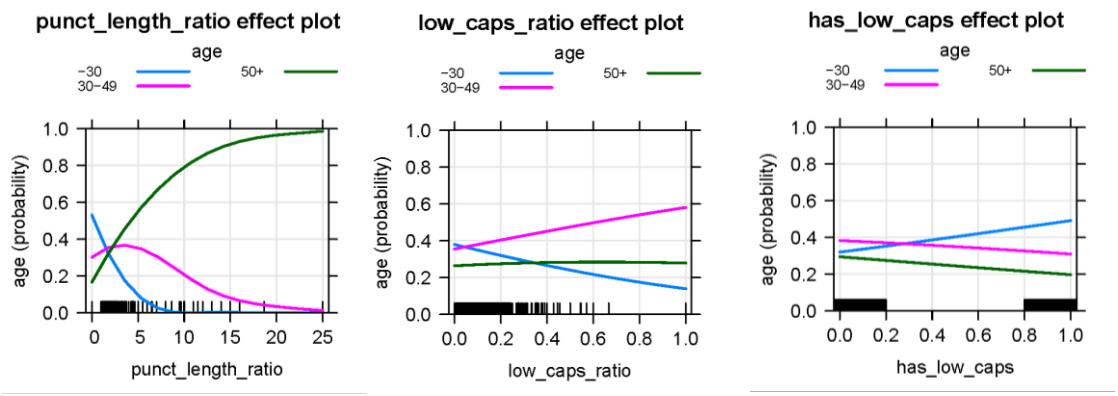


Figure 93: A selection of effects plots for stylometry features.

Moreover, here we can observe the variable effects for the ratio of punctuation tokens to tokens in the text as well as the text length (the two highest ranked features according to the mean gini impurity decrease in the random forest classifier) by plotting partial dependence plots for the random forest classifier[182]. The plots show the probability for predicting a younger writer is higher with a higher ratio of punctuation tokens, while lower ratios are a sign for writers over 30 or even older writers. Furthermore, the longer the text, the higher the probability is of predicting a writer between 30 and 49 in the random forest model (see Figure 94). The remaining factors did not show any effects that were not already shown by the glm model.

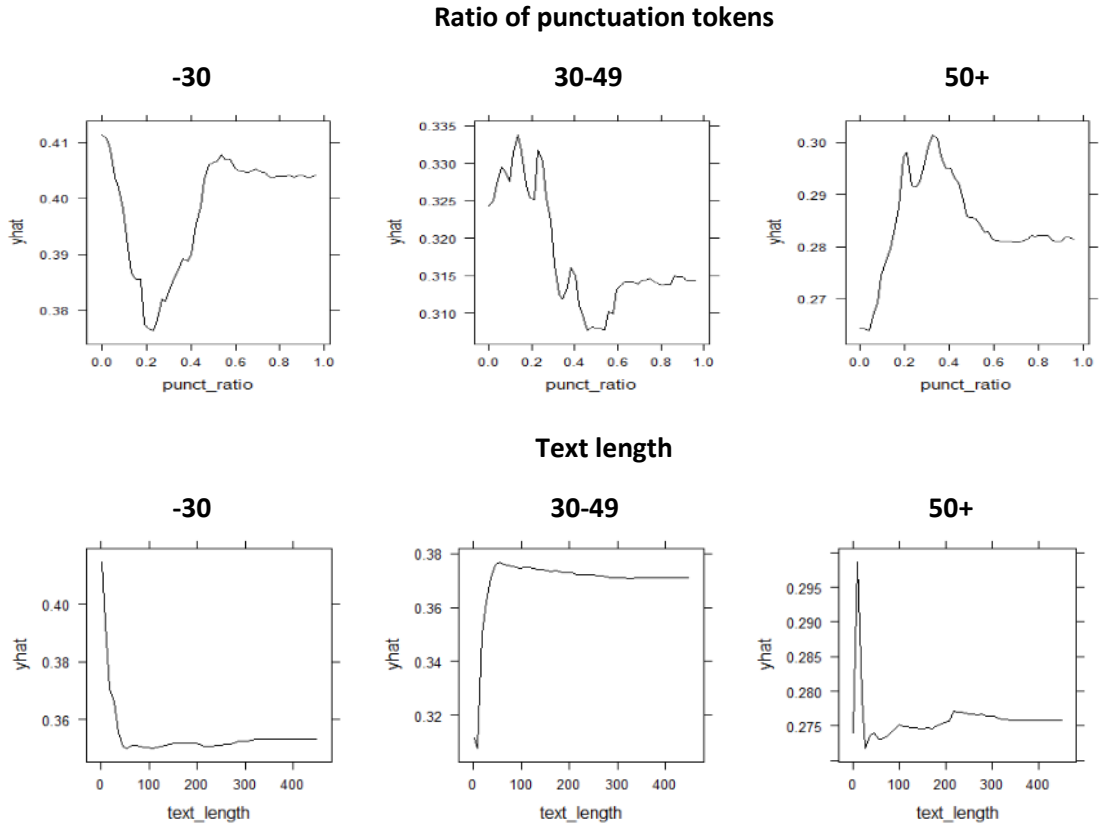


Figure 94: Partial dependence plots for punct_ratio and text_length.

[182] For all partial dependence plots the function `partial` from the `pdp` package for R was used.

Differences in vocabulary

To describe the language further, I apply an association measure study similar to the open-vocabulary approach proposed by Sap et al. (2014) to inspect the previously untouched content dimension of age-related language in the corpus via age-predictive lexica.

For this purpose, I extract a lexicon of every original word form (including standard and non-standard spellings, henceforward called *spellings*) in the corpus as well as a version were non-standard spellings have been unified (henceforward called *vocabulary*). I subsequently filtered all types that occurred at least 50 times in the corpus and that were used by more than one user. For these types I then calculate the Pearson product-moment correlation with their occurrence in a text of each of the three age groups using the Bonferroni correction and a confidence level of 99.9% to account for repeated testing (cf. Sap et al., 2014).

In order to evaluate the results numerically, the 20 highest correlations for the vocabulary and the spellings for the three age groups are listed in Table 58 and Table 59.

Correlated vocabulary

younger		**30-49**		**older**	
haha	0,111	einsetzen	0,113	nun	0,103
in	-0,097	Wohle	0,111	hier	0,103
die	-0,091	Leuten	0,106	mich	0,099
sel	0,089	aller	0,102	liebe	0,095
lei	0,088	Online	0,081	hallo	0,079
liebe	-0,087	Hallo	0,080	in	0,079
Hallo	-0,084	Grüße	0,079	Liebe	0,079
und	-0,083	Kommentar	0,076	wünsche	0,077
hier	-0,082	Neuer	0,074	unterschreibe	0,076
nur	-0,082	zum	0,073	da	0,074
an	-0,080	CD	0,072	und	0,073
zu	-0,078	um	0,061	sel	-0,070
hel	0,076	guten	0,059	Freude	0,070
zum	-0,076	Welt	0,059	Hi	0,069
Liebe	-0,074	haha	-0,057	dich	0,069
sehr	-0,074	an	0,057	gerade	0,069
für	-0,073	Mein	0,056	immer	0,067
immer	-0,073	die	0,055	dir	0,066
Grüße	-0,072	Südtirol	0,054	dort	0,066

Table 58: Twenty highest correlated vocabulary items per age group.

Correlated spellings

younger		**30-49**		**older**	
ich	-0,168	einsetzen	0,113	ich	0,162
i	0,155	Wohle	0,111	mich	0,138
xD	0,146	Leuten	0,110	liebe	0,130

auch	-0,140	aller	0,103	i	-0,121
das	-0,138	zum	0,083	hab	0,118
dr	0,138	Online	0,082	das	0,116
ist	-0,137	daß	0,080	da	0,115
jo	0,133	Hallo	0,079	dich	0,113
nt	0,127	Neuer	0,074	nicht	0,110
nicht	-0,127	lg	0,073	nun	0,102
ja	-0,125	die	0,069	ja	0,102
die	-0,119	guten	0,069	hier	0,102
mich	-0,113	Kommentar	0,067	a	-0,100
dann	-0,112	ned	0,067	dann	0,098
isch	0,111	nt	-0,067	<PersNE>[183]	0,096
haha	0,110	dr	-0,066	gut	0,096
mal	-0,109	um	0,063	auch	0,095
da	-0,108	Welt	0,063	grad	0,093
ein	-0,106	fuer	0,062	dir	0,092

Table 59: Twenty highest correlated spellings per age group.

Figure 95, Figure 96, and Figure 97 show positively correlated vocabulary for the three age groups. The strength of correlation is indicated by the size of the words. The upper word cloud in each set gives significantly related words for the vocabulary in general, using counts for words in their standardized form. The lower figures show the spellings that were correlated with 50+ writers. These word clouds make it possible to evaluate the vocabulary in terms of the semantic content of texts as well as the spellings (standard and non-standard) used. They allow for hypotheses on the expressed communicative functions of the medium and patterns of non-standard use.

Comparing the word clouds for older users with those of the other age groups, we can assume that writers over 50 use Facebook predominantly for phatic communication and to connect with others (e.g. to make plans). Contrarily, writers in their 30s and 40s seem to be more concerned with expressing their opinions and doing explicit facework by showing that they are sociable that they take social responsibility, and that they are situated in a local social network (references to South Tyrolean, family and leisure time activities). The vocabulary of under 30 year olds on the other hand shows a casual use of the medium, with less content words and more function words (interjections, adverbs, etc.) and semantically less charged verbs (e.g. *muas*, *tuasch*, *hosch*), a high amount of dialectal spellings and dialect-specific lexemes as well as emoticons.

In terms of non-standard spellings, the older age groups are characterised by forms known from colloquial writing such as the elision of the unaccented e in high-frequency words (e.g. *grad* instead of *gerade*, *hab* instead of *habe*). Compared to this, the correlated non-standard spellings of writers between 30 and 49 are due to typing conventions (e.g. *fuer* instead of *für*), formerly valid norms (*daß* instead of *dass*) and some highly frequent dialectal spellings (*ned*, *haint*). The correlated non-standard spellings for the younger writers under 30 are almost exclusively dialectal spellings.

[183] <PersNE> is used in the corpus to anonymise names of people.

Young writers under 30

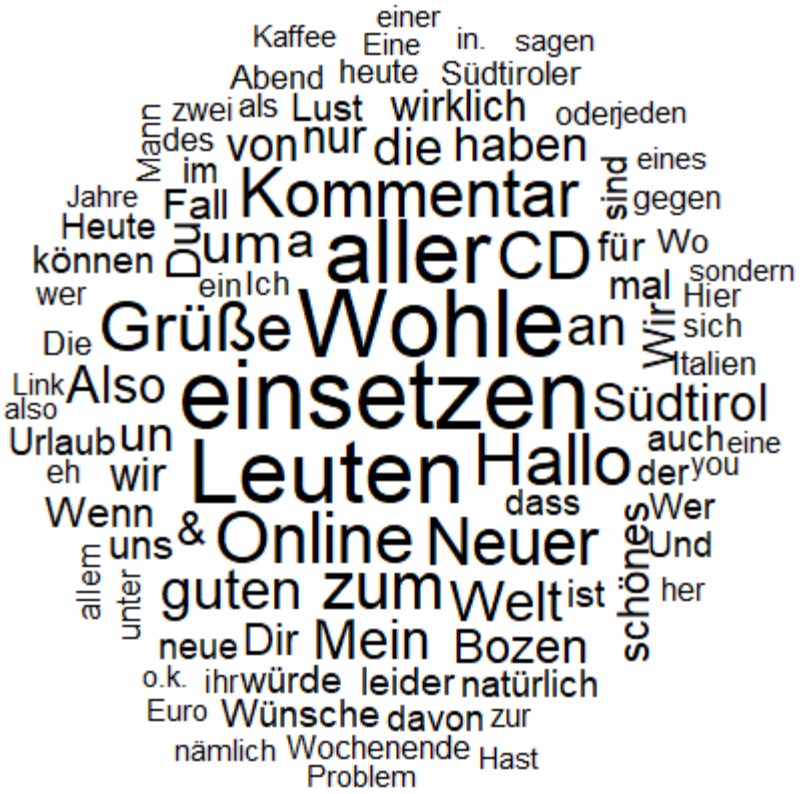
Vocabulary

Spellings

Figure 95: Word clouds for vocabulary items and spellings that are significantly correlated with writers under 30.

Middle age group (30-49)

Vocabulary



Spellings

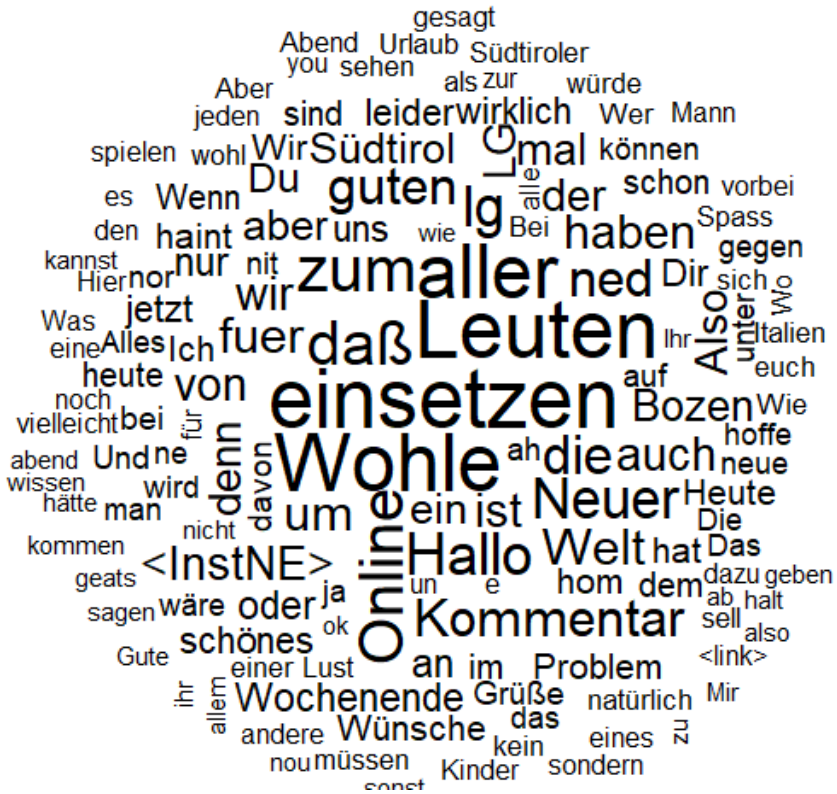


Figure 96: Word clouds for vocabulary items and spellings that are significantly correlated with writers between 30 and 49. <InstNE> signals anonymised content that refers to institutions.

Older writers over 50

Vocabulary

Spellings

Figure 97: Word clouds for vocabulary items and spellings that are significantly correlated with writers over 50. <PersNE> signals anonymised content that refers to people.

22 Evaluating different operationalisations of age (RQ 6)

In this section, I evaluate different possible operationalisations of age in order to extend the chronological definition of age in years since birth, often criticised as too simplistic (Eckert, 1997; Glaznieks & Stemle, 2014; Nguyen et al., 2014). Chronological age is compared with other operationalisations that were possible through the socio-demographic data of the users in the DiDi corpus.

Chronological age

Chronological age in the form of the number of years since birth is the most common operationalization for age. For comparability, the age groups used for the previous experiments are also used in this analysis.

Social age

Social age is defined by a person's affiliation to social age cohorts such as students, employees or retired people, and does not necessarily coincide with chronological age.

Digital age

In addition to chronologically and socially motivated operationalizations of the age variable, Glaznieks and Stemle (2014) proposed a digital age, defined as a person's experience with the digital world or with the digital communication medium. As the authors do not further specify their intended construction of the digital age variable, I use the metadata available in the corpus to construct and compare three possible measures for digital age:

a) a person's *exposure to the internet*, given by the amount of years he or she has actively used the internet;
b) a person's *usage frequency of the medium*, given by the frequency of his or her social media use for social networking sites;
c) a person's *experience with digital media*, as a composite variable of exposure to the internet and usage frequency of SNS.

To measure internet exposure, I use the corresponding metadata annotation available as count of years and build three groups. Users with long-term internet exposure have had at least ten years of active use of the internet. Users with less than five years of active internet use are classified as having had short-term internet exposure. Everyone else (5-9 years) is classified as having had medium-term internet exposure.[184]

To measure the usage frequency of the medium, I grouped together people who indicated in the questionnaire that they use SNS at least once a day as frequent users, users than indicated that they use SNS at least once every week as moderate users and users that indicated less than once a week as rare users.

The third measure of digital age combines both previous user indications. Users with long-term internet exposure and high usage frequency were classified as highly experienced. Users with either low

[184] This variable can be in some cases counter-indicative to the digital native – digital immigrant perception of younger people being more proficient with the web, as this operationalization favours older people that had more change of longer-term exposure to the internet.

usage frequency or short-term internet exposure were classified as inexperienced. Everyone else was classified as moderately experienced.

Table 60 to Table 62 show the distribution of texts according to the writer's chronological age, social age and the three operationalizations of digital age.

Chronological Age	
-30 years	5399
30-49 years	4688
50+ yeas	4259

Table 60: Distribution of texts for chronological age.

Social Age	
school	1395
university	2678
work life	6256
retired	525
NA	3429

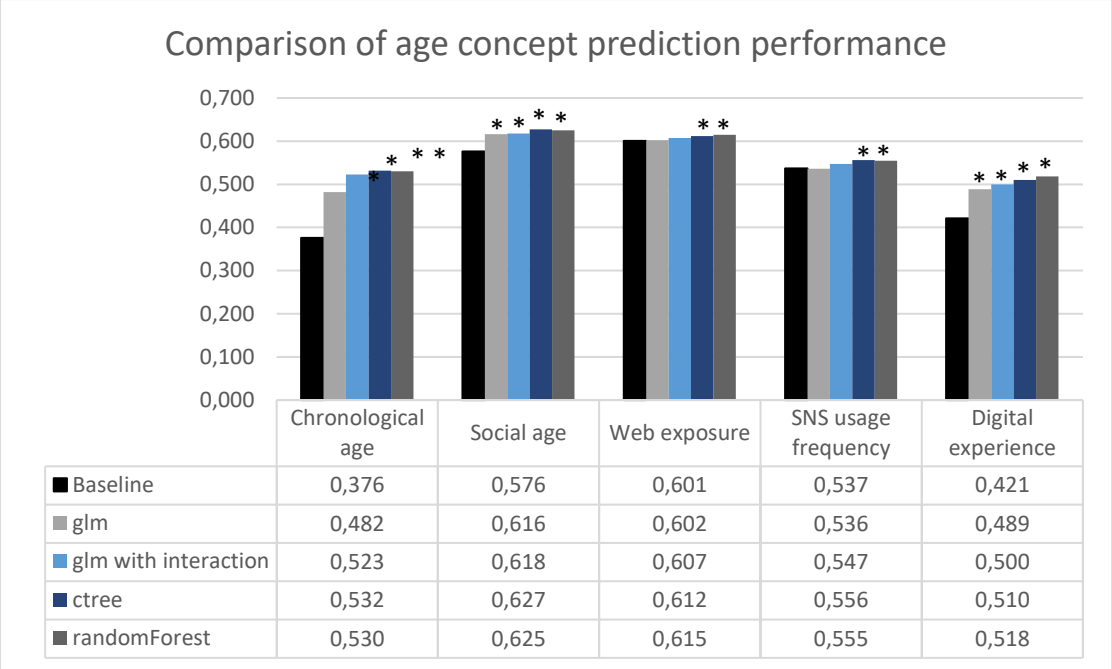
Table 61: Distribution of texts for social age.

Digital Age					
Web Exposure		**Usage Frequency**		**Experience**	
short-term	1598	rare	2201	low	2201
medium-term	4085	moderate	3995	moderate	5543
long -term	8568	high	7183	high	5635
NA	95	NA	967	NA	967

Table 62: Distribution of texts for the three operationalizations of digital age.

In order to evaluate the different operationalizations of the age variable, I train four different prediction models and compare their results to a) the baseline for the current operationalization and b) the other operationalizations. For comparable results I use the minimal variable set used within the first prediction approach for digital natives and digital immigrants, i.e. language choice, language variety choice and the ratio of CMC style markers including possible interactions with the text type.

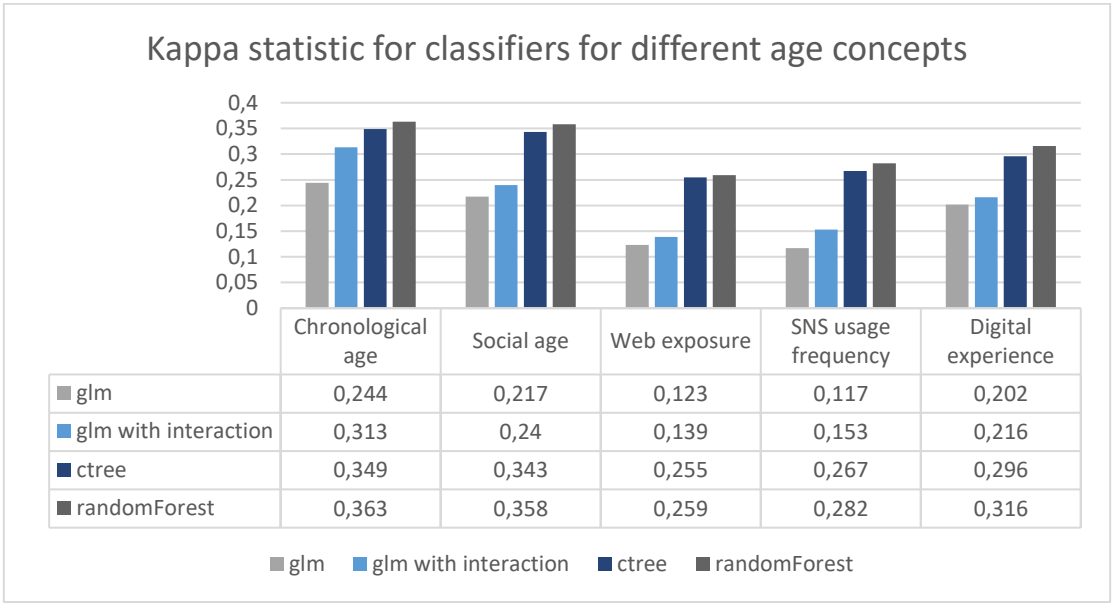
The test results for these experiments can be seen in Figure 98. It is worth mentioning that all models performed significantly better than the majority baseline when using tree-based models (ctree or randomForest). The glm models for web exposure measured in years of active internet use and SNS usage frequency were not significantly better than the baseline models for these age concepts.



	Chronological age	Social age	Web exposure	SNS usage frequency	Digital experience
Baseline	0,376	0,576	0,601	0,537	0,421
glm	0,482	0,616	0,602	0,536	0,489
glm with interaction	0,523	0,618	0,607	0,547	0,500
ctree	0,532	0,627	0,612	0,556	0,510
randomForest	0,530	0,625	0,615	0,555	0,518

Figure 98: Comparison of 10-fold CV accuracy results for different age concepts.

The figure shows that the classification problems had varying baselines, making direct comparisons difficult as the absolute performance and the increase over the baseline differ[185]. I therefore use Cohen's Kappa to compare the performances after correcting for different baselines according to Ben-David (2008).



	Chronological age	Social age	Web exposure	SNS usage frequency	Digital experience
glm	0,244	0,217	0,123	0,117	0,202
glm with interaction	0,313	0,24	0,139	0,153	0,216
ctree	0,349	0,343	0,255	0,267	0,296
randomForest	0,363	0,358	0,259	0,282	0,316

Figure 99: Comparing prediction performance for different age concepts using Cohen's Kappa to account for differing baselines.

[185] Although the models trained to distinguish chronological age show the highest accuracy increase over the baseline, predicting the age operationalizations for social age, web exposure and SNS usage frequency has higher absolute accuracies.

The table in Figure 99 shows the Kappa values for the classification experiments on the different age concepts. The graph shows that, given the features used, chronological age could be modelled best, followed by social age, for which the results were only slightly worse than for chronological age in the tree-based models. However, the operationalisations for digital age, measuring exposure to the web, and/or usage frequency of SNS achieved worse results.

23 Summary and discussion

The experiments gave a number of interesting insights into age-related features of language use in semi-public and private SNS communication in South Tyrol that complement our previous knowledge on the field.

The study addressed Prensky's theory of digital natives, i.e. people born and raised in a digital environment who behave and communicate differently to their older counterparts, the digital immigrants. The analysis showed that South Tyrolean Facebook users who were born after the year 1980 and therefore grew up with digital media do indeed differ in terms of language choice, variety choice, and the use of CMC style markers employed in SNS. They use more dialect, less standard-oriented language, and more CMC style markers.

While the prediction of two age groups (digital natives and digital immigrants) was possible with a prediction accuracy of about 0.7, the prediction of three age groups was slightly more difficult. When predicting three age groups with the same features of language choice, variety choice, and ratio of CMC style markers, the prediction accuracy decreased to ca. 53%. The accuracy was furthermore not equally good for all age groups. The identification of older writers was difficult for all observed models (generalized linear model, conditional inference tree and random forest), with an average recall of ca. 0.56 (i.e. the percentage of successfully identified older writers in the corpus) and an average precision of ca. 0.38 (i.e. the percentage of correctly classified instances of older writers among the instances that were predicted to be written by older users). The group of older writers that has so far barely been analysed in CMC language due to a lack of data (cf. Nguyen et al., 2014; Peersman et al., 2016) was therefore also difficult to interpret in this dataset, despite having a sufficiently large number of texts written by older writers over 50 years of age. At least for this simple set of predictor variables, discriminating older writers over 50 from writers between 30 and 49 and younger ones under 30 was not very reliable.

Considering previous research that observed U-shaped distributions for vernacular language and non-standardness over a person's life span (e.g. Chambers, 2003; Downes, 1998; Labov, 2001; Peersman et al., 2016), it was interesting to divide between the non-standardness of a text and dialectal texts as both can have different functions in CMC. While dialectal texts are referred to as being a tool to express one's local identity in a global communication setting (Androutsopoulos, 2013b; Kytölä, 2016), vernacular registers and thus non-standard language is expected to be used when a person does not want or need to adapt to community norms (cf. Chambers, 2003; Peersman et al., 2016). Using the dialect as exclusive marker for non-standardness and vernacular registers is, however, problematic in the South Tyrolean context, as it is known as the unmarked variety in most forms of informal language use (Eichinger, 2001). The study therefore tried to distinguish between choosing the dialect to compose texts and using features of non-standardness in general. Through the analysis, the study could show that prediction models achieved better results when using the more general features of non-standardness compared to the variety choice as predictor variables. Besides, using both features in

the model did not yield significantly better predictions, hinting to a conceptual difference between dialectal writing and non-standardness for discriminating author age in CMC.

However, by extending the feature set to include other stylometric features that have been used for age prediction in other languages (cf. Flekova et al., 2016; Simaki, Aravantinou, et al., 2017) it was possible to identify further language features indicative of age in the German SNS texts from South Tyrol, reaching an overall prediction performance of 58.6% accuracy for the prediction of three age groups, which was the highest in this study.

The **interpretation** of the simpler model including all languages through partial dependence plots and marginal effects plots shows a higher ratio of non-standard-spellings and CMC style markers for younger writers under 30 as well as a higher use of dialectal texts. The expected higher use of non-standard varieties of older writers when compared to the writers between 30 and 49 was, however, only visible for older people's private chat communication and could not be found in the more public wall posts and comments. It seems that writing in non-standard varieties in a public and thus more formal situation is not considered by older writers, although they do show their competence and willingness to write in dialectal form in chat messages. However, the writers between 30 and 49 showed, as expected, the highest association with standard-oriented texts and the lowest association with dialectal texts and non-standard spellings (although not very different from the older writers). It seems as if the main difference between middle aged and older writers is their attitude towards (semi-) public and private written language. The writers between 30 and 49 clearly prefer standard-oriented German or Italian (i.e. one of the codified and acknowledged languages of the territory) in their private communication, while their public posts and comments are more open towards other varieties. Furthermore, compared to the other age groups, the increased use of the second language Italian among writers over 50 in public wall communication is salient.

Regarding the question of whether the observed relationship between digital natives and language use is generation-specific, i.e. obtained during a person's life span and carried on until he or she dies, or age-specific, i.e. preferably used during a certain age, the synthetic age cohorts in the data do not allow us to make any conclusions. However, the observed decrease of the use of Italian and increase of the use of English for younger generations might suggest a societal change in South Tyroleans' everyday language use, rather than an age-preferential use that will change once people get older.

With regard to the **stylometry features explored in German texts** only, the average number of characters in a punctuation token is particularly relevant for predicting older writers over 50. Emoticons and Emojis are highly predictive for younger writers under 30 as well, alongside the ratio of words that are unique in the corpus or not used by other users. Writers between 30 and 49 had a higher ratio of uncapitalized nouns and sentence beginnings than other age groups. In terms of CMC-specific style they furthermore used linking strategies provided by the platform more often in their texts. This finding strikes me as of general interest as it goes against the trend of CMC-style markers being a sign of younger writers. However, while consistently writing in lower case is a relatively common stylistic decision in CMC and is therefore less often seen as a norm violation, the use of linking strategies could also be interpreted as a need to be socially recognized and connected by linking people (@), contents (hashtags) and web-resources (URLs) to the text. The findings of these experiments could therefore support the Adolescent Peak Principle, the observation that people's use of non-standard language including local dialect varieties peaks in their adolescence and declines as they get older, until it reaches its lowest point around the age of 45.

The **language of older users** was described as similar to middle age group in their use of non-standard language and CMC style markers, but in terms of hashtags, hyperlinks and @mentions they were

indeed closer to the younger writers under 30 than to the writer between 30 and 49. Older users seem to make a clear difference between public or semi-public communication settings and private communication, whereas this difference is not as visible in other age groups. Their vocabulary use suggests that they use Facebook predominantly for phatic communication and for connecting with others (e.g. to make plans, etc.), and that they employ non-standardness mostly in the form of colloquial writing that also occurs in other forms of written language (e.g. colloquial elision of graphs).

Finally, an evaluation of prediction performance given different proposed age concepts also provided interesting insights. Chronological age, measured as groups of younger, middle-aged and older writers, and social age, measured in life stages, can be learned relatively well by learning algorithms that use language choice, variety choice, and the ratio of CMC style markers in the text. However, the operationalizations for digital age (i.e. a person's experience with the medium measured as web exposure, SNS usage frequency, and a combined measure of both) were less easy to predict by these language features according to a classifier comparison based on Cohen's Kappa values for baseline-corrected prediction performances of four different classification methods.

24 Conclusion

This second case study analysed age-related language in SNS writing in South Tyrol. It used the openly available DiDi Corpus of South Tyrolean CMC to observe linguistic differences in older and younger writers in detail. The study discussed previous research in the field and laid out the current understanding of age-related language in CMC as well as possible factors and interactions, before it addressed a number of pending research questions with data mining strategies known from the age-prediction framework of computational sociolinguistics (Nguyen et al., 2016).

While the study adds new knowledge to the understanding of local language use in South Tyrol, the contributions of this study to the analysis of age-related language in social media are manifold. This study is, so far, the first work to perform age prediction experiments as a data mining technique to analyse age-related language with a prevalently German language corpus. It tested a broad set of hypotheses regarding age-related language in general as well as specifically in social media, and thus contributed by giving a thorough account of the topic based on a single dataset that contained various text types. The large number of texts written by writers over 50 in the corpus made it possible to put a special focus on this older age group and to compare the language used by older writers with writers in the middle and younger age group, something which has not been possible in other datasets before. Moreover, other studies usually were constrained to approximating age using unreliable clues in the texts or the user profiles, whereas the socio-demographic metadata available for this corpus allowed this study to use reliable age labels. It furthermore gave the possibility to evaluate chronological as well as social and experience-related operationalisations of age, in order to give a more detailed picture of the various aspects at play.

In future research it would be interesting to investigate if the acceptance of non-standard language and other CMC style markers in this written but informal social media genres changes over time. This study showed that older writers make a clear difference between public and non-public communication, using less dialectal and non-standard language in public genres than younger writers. However, as people get more accustomed to the medium, such behaviour could change in future generations. Given the fact that the used data is from the year 2013 it could be interesting to compare results with a newer dataset. Future research could also explore the content-dimension of the used language in

more detail, using e.g. topic modelling or semantic approaches to text analysis and results from the here presented informal but written genres could be compared with South Tyrolean spoken communication. Another objective would be to investigate the difference between age-specific and generation-specific features of language use. For example, regarding the results for language choice of younger and older writers that showed a higher use of English but less Italian among younger writers and the opposite for older writers, age-related explanations as well as generation-related explanations can be found. With the given data and methodologies, it was, however, not possible to investigate this aspect.

Part III

Evaluation

A critical discussion of data-driven analysis methods for corpus repurposing

25 Methodological discussion of corpus study 1

25.1 Aim of the study

The first case study explored holistic grades as a measure of text quality in argumentative student essays of first language writers based on three different feature sets that were provided or extracted automatically for an available corpus of medium size.

In particular, metadata features from a text analysis questionnaire, frequency lists from multi-level error annotations, and indices (measures) of linguistic complexity were investigated in terms of

- how strongly they are related to holistic text quality judgments (monofactorial relations)
- how well they are suited to explain them (descriptive or explanatory modelling)
- how well they can, when combined together, predict the holistic grades (predictive modelling).

For all three feature sets monofactorial and multifactorial methods were used and compared in order to show the potential of predictive modelling for corpus linguistics. Furthermore, both text classification and regression models were used for building and interpreting predictive models, illustrating a broad set of possible methods and strategies.

The experiments illustrated a series of possible strategies and choices when exploring a corpus with data mining through predictive modelling and machine learning. They discussed:

- the use of different learning algorithms for building predictive models, referring mainly to standard implementations in their default configuration as provided in common data mining or machine learning software.
- the use of different task definitions for building predictive models (classification and regression, binary an multi-class classification).
- the use of different feature types that represent typical data types in corpus linguistics:
    - categorical features from a text analysis questionnaire
    - ordinally transformed features from a text analysis questionnaire
    - raw, normalised, capped and binarized error frequencies retrieved from a hierarchical error annotation scheme
    - continuous measures of linguistic complexity (absolute or standardized)
- the use of different feature selection methods and different missing value treatments.
- the resulting effect on predictive performance and model interpretability of the aspects above.

The experiments also showed how to evaluate models, how to compare them statistically and how to select one of a series of models for in-detail interpretation, finding a trade-off between interpretability and predictive performance as defined by the heuristic of *Occam's razor*, i.e. the simplest model that does not perform significantly worse in prediction than other tested models.

In a data-driven perspective, the experiments started with a naïve approach using almost untouched data and operationalizations. The original model for the task of predicting the holistic grades thus used an unrefined feature sets and all five grade levels for prediction. In order to reduce the complexity of this original model (or raising its predictive performance), various methods were tested, reducing the complexity of the feature set, class labels, learning algorithms, etc. However, the experiments aimed at integrating necessary complexity in case it increased the correctness of the model (e.g. when non-

independent datapoints or non-linear relationships are present). Furthermore, the experiments addressed questions of how to deal with hierarchy in the data or the features (mixed-effects models, hierarchical regression); how to deal with outliers and missing values, how to address multicollinearity and how to inspect biases and confounders.

In order to evaluate the interpretability of a selected model, the experiments used a set of interpretation methods, including methods for intrinsically interpretable models and methods for post-hoc black box interpretation of complex models.

25.2 Results

The experiments showed beneficial but also problematic aspects of predictive modelling and data science methods for the linguistic analysis of holistic grades.

25.2.1 First explorations & basics of data science

The first set of experiments in section 10 illustrated basic strategies and methods for data mining using standard algorithms with standard configurations for text classification in an easy-to-use data mining software with graphical user interface (WEKA) as well as standard methods for regression modelling with the statistical programming language R.

First, the effect of different problem definitions, different algorithms, different feature selection methods and variable treatments on the predictive performance of predictive models was discussed.

The experiments showed that tasks with more balanced classes or less class levels are naturally easier to solve for the learning algorithms resulting in higher prediction performance. These models are also easier to interpret for humans.

The comparison of different algorithms showed that some learning algorithms, known for producing simpler, intrinsically interpretable models have usually less predictive power, especially when many features are involved. The tree ensemble method '*random forest*' was most robust to noisy data in the experiments and yielded good predictive performance for classification tasks without changing default parameters. Other black box models such as support vector machines and multilayer perceptron neural networks, were significantly slower to train, and less robust towards untidy data or missing parameter optimization, resulting in lower predictive performance than the one for the random forest model. However, the more complex black box models, and especially the random forest method, worked better with the original, unchanged and complex feature set. Refining the feature set (feature selection, missing value treatment, feature transformations, etc.) did not result in better models but lowered the prediction performance significantly in most cases.

The classification experiments thus showed that the best performing models were actually complex models with almost unchanged data trained on robust black box learning algorithms. These models were, however, barely interpretable with the methods and model outputs provided by the WEKA data mining software, where they were trained in. While the simpler, but not so performant models allowed some investigation of the model internals (regarding global feature importance and feature effects), the black box models with the highest prediction performances only allowed to systematically compare prediction results for different feature subsets or to investigating feature ranks and confusion matrices.

Finally, this section also explored linear regression models by reformulating both holistic grades and the ordinal questionnaire items into numerical scales. Aiming to explain the variance in the dependent

variable (no prediction) with the predictor variables from the text analysis questionnaire, stepwise model selection approaches for mixed-effects models were performed with the statistical software R. While the transformation of the data into ordinal scales is controversial, the same approach with other regression models (e.g. *generalized linear mixed-effects models* with categorical predictor and descriptor variables) was unfeasible due to the complexity of the problem.

In total, the classification approaches as well as the regression models showed that the results for such multifactorial methods differed clearly from the ones obtained by monofactorial methods. Predictive performance and thus generalizability is not necessarily given by the features with the highest monofactorial correlation, but by features that complement each other contributing in combination to creating well-performing models that can explain the variance in the holistic grades.

25.2.2 Dealing with issues of observational data

Section 11 addressed issues of data distribution and representation by analysing frequency data from manual annotations as often performed in corpus linguistic studies. It investigated linguistic accuracy in the form of relative error frequency and error presence in the essays.

Although frequency lists for words or other linguistic phenomena like errors are frequently used types of features in corpus linguistic studies, they pose some problems when used in data mining scenarios.

- They are often biased by confounders like text length or individual variation that need to be accounted for in the analysis.
- They usually lead to sparse feature sets, i.e. feature sets where only a few observations per feature have actual values and the rest are zero counts, which is a problem for many learning algorithms.
- They are prone to contain outliers that can introduce substantial bias in the results of most predictive models.
- They often have very skewed distributions, which can reduce the performance of predictive models.

Additionally to the very low error frequencies in this type of data (L1 essays written in last grade of secondary school), the bivariate correlation between form errors and holistic grade is very low for this dataset and the trained classifiers and regression models did not reach the baseline accuracy needed to establish that the model picked up important structures in the data. This indicates that the error frequencies had no or very little influence on the holistic grades assigned by the annotators, and neither transformations for better variable distributions (removing outlier values or reducing skewedness by variable binarization) nor the selection of subsets of features to reduce complexity (via automatic or manual feature selection methods) helped to achieve significantly better prediction results.

With regards to sparsity of the data, aggregating error types can help to accumulate evidence for a broader category of errors and reduce the time and effort needed for annotation. The total number of lexical errors for example is one of the most relevant features for explaining the holistic grade.[186] However, it can also disguise the effect of individual features by diluting it with other unrelated features. For example, only few orthography errors are correlated with the holistic grade, and the linear regression model showed that only errors regarding word separation and compounding contribute significantly to explaining the variance in the dependent variable, leading to a lower correlation and no significant effect on the dependent variable for the aggregated total of orthography errors.

[186] Probably because the comprehensibility of the text might seriously decrease with the number of misused lexical items.

Unfortunately, such effects cannot be anticipated prior to in-detail analysis. Furthermore, the theoretical validity of aggregating or splitting variables needs to be considered, whenever such measures are taken.

The analysed dataset was very small and there might be more evidence in bigger datasets allowing to build a classifier or regression model that is able to surpass the majority baseline, demonstrating that the machine can learn from the data. Given the fact that error annotations are usually obtained laboriously by manual annotation, the chance of obtaining big enough datasets is, however, relatively low and, considering the low correlations, such endeavour might not be worth an extended study.

To sum up, only few significant and weak monofactorial correlations could be found between form violations and the holistic grade. The variables are strongly biased, skewed and theoretically problematic, because of the deficiency-based perspective to linguistic accuracy. Outliers made the analysis additionally difficult. Although the experiments based on the error annotations showed differences in the insights gained from monofactorial and multifactorial methods, the use of multifactorial methods was difficult because of the hierarchical, multicollinear, barely relevant and overly complex data as model assumptions might be violated, regression coefficients might be incorrect and model selection techniques might fail. Furthermore, the features were hardly relevant for the prediction of holistic text quality ratings in this corpus. The various possibilities to transform, aggregate and split the variables partly helped to decrease the complexity of the model and/or build models that perform better (when explaining the variance in the data). They also allowed for a more detailed monofactorial analysis. However, some transformations that are often suggested in the literature, hardly made a difference in prediction performance if more complex black box models like tree ensembles are used. The interpretation of multifactorial models was therefore only possible via explanatory models that did not ensure a baseline of predictive performance.

25.2.3 Interpretation of complex feature sets

Section 12 finally focused on the interpretability of complex models. The experiments showed various difficulties for model interpretation, when complex modelling tasks are attempted.

Monofactorial methods only allow superficial analysis of relationships in complex feature sets. However, the interpretation of complex predictive models depends on various factors.

A number of experiments showed that for the task of predicting five grade levels of holistic text quality based on a large set of linguistic complexity measures, intrinsically interpretable model typologies such as decision trees or regression models that are usually preferred when it comes to explanatory settings are reaching their limits. The experiments showed that theoretically interpretable decision trees get big very fast and only few (continuous or multi-class categorial) variables are needed to make them incomprehensible. Furthermore, regression coefficients for theoretically interpretable regression models are too diluted by multicollinearity, and complex mixed-effects models with manual model selection processes are difficult to build and to interpret (especially for multinomial logistic regression). While generalized linear models (including logistic regression for classification problems) are often easier to interpret with continuous numerical features and certainly easier to conduct for regression problems if mixed-effects models are used, decision trees are less complex, when categorical values are used, as subtrees cannot be split repeatedly on different thresholds of the continuous variables. However, in both cases practical interpretability as well as the achieved predictive performance prohibited the use of such models for data-driven analysis.

A second set of experiments tested post-hoc black box interpretation methods for this complex modelling task, starting with simple classifier comparison and feature engineering/feature ranking

approaches and ending with interpretation methods for extracting global and local feature importance, feature effect and interaction effects for complex models. While the first two approaches used the standard setup of classifiers already used in section 10, the last part trained customized black box models with random forests and neural networks and used the Python SHAP library for extracting and visualizing Shapley values that can inform about global and local feature importance, effect and interactions in trained models.

The systematic classifier comparison and ablation study showed the importance of individual subsets of features, providing some general insights. The feature ranking showed relevance estimates for some individual features, with limited interpretability as no feature effects could be derived. In the model-agnostic post-hoc interpretation of the customized random forest and neural network models done with the SHAP library the interpretation of the most salient features was feasible, whereas the interpretation of less important features was partly possible but often hampered by the small data size.

The feature effects reported for the less important features in the neural network models often concerned only few individual predictions and therefore did not allow their interpretation. Moreover, the effect direction was not always clear, as collinear features can have strong but contradictory effects. The random forest model on the other hand allowed for broader generalizations over more than just one or two observations. However, it was marked by less predictive power and very few variables that overruled all other effects. The important variables were furthermore all from very similar categories and therefore did not add any additional information to the human understanding of holistic grades[187]. Moreover, the type of the model conditions the types of relationships learned. The graphs are hardly interpretable without a basic understanding of the learning algorithm. We saw that two equally well-performing models do not need to make use of the same features and can therefore also lead to completely different interpretations, given the same dataset. Only the interpretation of more than one well-performing model can thus give less biased results.

In total, we can conclude that the interpretation of complex models with unfiltered, collinear and noisy feature sets like the linguistic complexity measures used in this study needs to be taken with caution. Although the models reported good predictive performance and picked up relevant information that helps to assign holistic grades, the actual gained insights were limited. Methods for post-hoc black box interpretation of complex models like Shapley values that are by now widely acknowledged in the explainable artificial intelligence and interpretable machine learning community, can help to interpret model internals, but they strongly depend on the concrete model explained, which is why the interpretation of various, probable diverging but well-performing models should be considered.

25.3 Summary and conclusion

The analysis of text quality in student essays is a many-faceted problem in linguistics. Complexity is introduced by the operationalization of the text quality concept (in this case a holistic grade on a 5-point Likert scale), by the selection of features to investigate (where many different features have been proposed in previous research), by the nature of the investigated features (distribution, data types), and by additional bias of confounding actors (e.g. annotators). In this context, the chosen methods influence strongly which results and insights can be drawn from the data. While monofactorial analysis like correlation analysis are unreliable when possibly biased, observational data as in language corpora are used (as shown in the experiments but also pointed out for example by Gries,

[187] Although the algorithm did of course learn additional information.

2015d, 2015a, 2015b), the use of explanatory statistical modelling or predictive modelling can facilitate more complex analyses. However, the chosen methods influence the type and amount of insights that one can gain from the data. They define how much predictive performance the built model has (indicating thus how well it learns from the data and how well it generalizes to new data) but also how interpretable it will be.

While predictive models trained with simple intrinsically interpretable learning algorithms have been built and interpreted successfully for relatively simple problems (e.g. in Bernaisch et al., 2014, who modelled dative alternation with conditional inference trees using up to ten features), the models that perform well on complex modelling tasks like the exploration of 5-point scale holistic grades of text quality in a medium-sized corpus of argumentative student essays are complex as well. This comes with a series of shortcomings. First and foremost a sound theoretical background knowledge and practical skills are needed to build such complex models, and to refine them with feature engineering and other data science. Second, complex models are difficult to interpret no matter which learning algorithm is used. More often than not they need further post-hoc interpretation methods to investigate global and local model behaviour and to interpret feature importance, feature effects and interactions, which in turn need practical skills and theoretical background knowledge. Third, complex models need enough data to give insights on other features than the most strongly related ones.

26 Methodological discussion of corpus study 2

26.1 Aim of the study

The second case study used data mining strategies in order to test linguistic theories while widening the scope of variables, addressing more abstract research questions and comparing different operationalizations. Computational methods of data science were used for the extraction and preparation of variables, for training and evaluating predictive machine learning models and for interpreting the built models.

In terms of the choice of predictor variables, more variables than the amount usually used for confirmatory studies were addressed, respecting previous knowledge on related factors and possible confounders as well as tackling less concrete operationalizations for descriptor variables. In terms of the retrieval and preparation of variables, the use of previously investigated automatically retrievable linguistic features as well as testing different operationalizations with the same data was attempted. Analytical methods and tools, accounting for hierarchical data, as well as for repeated tests were used. Finally, in terms of interpretation and explanation, variable importance, variable effects and interactions were investigated to explain group differences globally as well as individual predictions locally (using effects plots, interaction plots, tree visualizations and variable importance measures).

By using the popular analysis frameworks of age prediction, the results could be related to similar studies on other languages. The age prediction scheme of computational linguistics thus offered the possibility to explore the DiDi corpus of South Tyrolean CMC (Frey et al., 2015), designed for the analysis of age-specific language, from a data-driven perspective.

26.2 Results

The use of predictive modelling techniques led to methodological advantages compared to simpler approaches without predictive modelling techniques.

First of all, the use of multifactorial methods based on predictive modelling made it possible to investigate the main effects of different linguistic features that are predictive of age in social networking communication in South Tyrol as well as to differentiate the effects for different text types produced in this medium, by analysing interaction effects. The hierarchically structured dataset representing different text types was used to point out genre differences that depend on the age of the writer. It therefore allowed for a much more fine-grained and correct investigation of digital text genres, while using the same data resource.

The study also made use of a wider set of predictor variables than usually used in confirmatory studies, respecting previous knowledge about confounding factors or other predictors. Through the extraction of stylometric features used in computational sociolinguistics, the exploration of further important features was feasible, leading to a richer understanding of social dynamics and the resulting linguistic behaviour.

Age group-related language behaviour was compared using effects plots, interaction plots as well as word clouds of correlated vocabulary items and spelling variants. Finally, different operationalisations of the social variable age were compared through evaluating Cohen's kappa for prediction accuracy of models trained on different target variables, which allowed to question the typical numerical measurement of age.

26.3 Summary and conclusion

In general, the experiments showed that the operationalization of a descriptor or predictor variable (e.g. age splits for digital natives and digital immigrants, non-standardness measured by dialectal texts or by incorrect spellings, etc.), the inclusion of control variables (e.g. text type or gender), the benchmarks set (different baselines, etc.) as well as the used methods (e.g. different algorithms, different interpretation techniques) can substantially change the conclusions we draw from the same dataset on the same questions. It is therefore reasonable to base the analysis not only on one individual methodological setup, but to make sure that the conclusions are not caused by individual methodological choices. The study thus consistently used and interpreted more than one model architecture. This comparison of results for different models, built with different learning algorithms, different descriptors or feature sets and interpreted with different interpretation methods, is an important step that is often neglected in confirmatory settings. Studies in applied linguistics often exclusively use one model type, most often a form of regression modelling. However, this work showed:

- that even with few predictors generalized linear models can perform worse than other models;
- that it is necessary to include interaction effects explicitly (as done for the text type interaction in the data) as every combinatorial effect that is not added explicitly will be neglected[188];
- that the conclusions drawn from interpreting regression models might differ from the ones drawn from other model architectures;

The emerging field of computational sociolinguistics counterbalances these shortcomings to a certain extend by making use of NLP methods for feature extraction and classifier comparison with different

[188] For further discussion see also (Gries, 2013, 2015c).

algorithms, but they usually don't perform in-depth model interpretation of variable importance and effects that includes direction and magnitude of effects for individual classes and for interactions. However, such interpretation would be needed for confirmatory analyses setups as employed usually in applied linguistics.

The use of computational methods makes it feasible to repeat analyses with intentionally changed parameters and thus helps to get a more profound understanding of a theory. Class definitions can be changed easily by extracting different versions from the base data. Further features can be added via text mining techniques from NLP. Moreover, through model-specific and model-agnostic interpretation methods, e.g. the visualization of tree models or partial dependence plots and variable importance measures for random forests, confirmatory analysis with predictive modelling is no longer restricted to the use of hitherto preferred (generalized linear) regression models. that is not only due to single methodological choices.

However, repeated testing of different hypotheses might enrich our understanding of a problem, but it also bares some methodological problems. Statistical tests always contain small probabilities of being incorrect, hence, doing many tests within one study increases the possibility that some of the results are incorrect (see also the notion of "data dredging", cf. Smith & Ebrahim, 2002).

27 Summary

While computational methods have always been an integral part of corpus linguistics in the stages of corpus collection, compilation and annotation, as well as in the retrieval of frequency lists, collocations and concordances, one of the main promises of data science and text mining for corpus linguistics can be located in the methods used for quantitative analysis.

The majority of recent quantitative studies in corpus linguistics includes simple statistical comparisons and monofactorial analysis (see section 2.2.2)[189]. However, such monofactorial research has been shown to encode methodological shortcomings that can lead to misguiding results (Gries, 2018; Paquot & Plonsky, 2017). This is why many scholars demand more rigour in quantitative corpus analysis, asking:

- to control overdispersion, multicollinearity, confounding and interactions among predictor variables;
- to address issues of hierarchical, multi-level data that can naturally occur through individual variation of writers that contribute in different quantities to language corpora;
- to guarantee objectivity in annotation and analysis (Kitchin, 2014; Moretti, 2013).

Furthermore, currently pending research questions are becoming more complex and advanced analysis methods including multifactorial research designs are needed to conduct research which:

- adopts a wider scope (e.g. cross-domain, cross-lingual analysis);
- uses more data (resulting in infeasibility of manual approaches);

[189] Although there are still some works using frequency lists, concordances and collocations without any statistical testing or control mechanism.

- observes complex, combinatorial (i.e. multifactorial) and non-linear relationships between linguistic phenomena and certain factors.

These are common requirements for present time research studies in applied linguistics (cf. de Marneffe & Potts, 2017; Desagulier, 2018; Gries, 2018; Kitchin, 2014; Paquot & Plonsky, 2017; Phakiti et al., 2018).

Finally, with the constantly increasing availability of language resources, the possibilities for corpus repurposing, i.e. the re-utilization and exploitation of time-consumingly created corpora becomes more and more important (de Marneffe & Potts, 2017; Kitchin & Lauriault, 2015; Kupietz et al., 2018). Therefore, the aims are:

- to explore corpora for patterns and relations using existing or automatic annotations;
- to do further analysis on the data, testing new hypothesis that were not envisioned before.

One possibility to deal with these issues is to make use of predictive modelling and machine learning when analysing a corpus. Using traditional statistical modelling techniques and machine learning methods, explanatory or even predictive models can be built that are able to do multifactorial analysis that accounts for combinatorial effects, known confounders and biases, addresses hierarchical data, can be applied to big datasets, for a wider scope and more complex problems, making use of datasets that are already available.

However, in order to use the built models, two prerequisites need to be met. First, the model's prediction performance needs to exceed the threshold of randomicity. And second, the model needs to be interpretable. Both of these requirements are genuinely related to the complexity of the model. While in many cases, more complex models perform better in prediction tasks, hence, have greater generalizability over the training data, they also tend to be less interpretable.

The complexity of a model, however, depends on many factors:

- the task itself (this includes how concrete or abstract the predicted variable is, but also how much we already know about the problem)
- the learning algorithm chosen to build the predictive model
- the levels and distribution of the dependent variable
- the number, type and distribution of features used in the modelling task including possible transformations made on them
- the type of relationship modelled (e.g. linear, non-linear, log-transformed)
- the hyperparameters and configurations of the learning algorithm (e.g. tree pruning, number of layers in neural networks, type of kernel in support vector machines)

Usually the simpler models are preferred for data mining purposes that aim to draw insights from the data[190]. However, there are reasons to build and interpret more complex models.

- The model is better at describing the structures and relationships in the data. This can be assumed when the prediction performance is significantly higher than for other simpler models.

[190] And often complex model architectures are not even considered in the modelling step excluding the fact that these models could potentially explain the relationships better.

- The simpler model is methodologically incorrect or at least dubious. This is the case when the model violates model assumptions, or multicollinearity distorts regression coefficients, etc.

A trade-off between the two forces needs to be found, by selecting the best performing, methodologically correct and still interpretable model, i.e. the model with the lowest complexity that a) exceeds the threshold of randomicity and b) does not perform significantly worse than other possible models. However, if more complex model architectures are considered, the model with the lowest complexity might not be simple enough to be interpreted directly but post-hoc black box interpretation methods are needed to interpret the model. Up to very recently these methods were primarily done through systematic model comparisons and ablation studies, but the last years showed significant advances in the interpretability of black box models coming from interpretable machine learning and explainable artificial intelligence. Strategies and methods to utilize even deep neural networks (and thus very complex model architectures) while being able to explain their predictions are being developed that can substantially change the selection of methods and model architectures tackled with data mining. These methods aim to individuate related variables, their importance and type, direction and magnitude of effect as well as possible interactions in the data.

However, post-hoc black box interpretation methods also have their drawbacks. The current methods and tools are often still experimental, available implementations often don't work out-of-the-box and need adaptations and troubleshooting to integrate them in the analysis process. Evaluations of the robustness and validity of some methods are still missing and/or showed problems on certain types of data.

Furthermore, the available data is a critical point in the interpretability of the models. The noise and redundancy in the data can limit the amount of insights that can be drawn from the models (inflated regression coefficients, contradicting variable effects, etc.). As much as the lack of data in general that can hinder to observe relationships in higher resolution. The evidence in the data might not be enough to see more than the main predictor variables that have already been identified with other simpler methods. Furthermore, the use of existing annotations bares the risk of adopting issues of reliability of the data and methodological soundness of operationalizations that have been made by other people and – most of the time – for different purposes.

28 Conclusion and future outlook

At this moment of time, the added value of data science methods with predictive modelling and machine learning for corpus repurposing is thus still questionable when it comes to repurposing language corpora of medium size and non-English language that have not been built for exactly the same analysis purpose.

Corpus study one showed interesting approaches for exploring a high number of relevant features through predictive modelling. Although model performances could be raised through more complex models, few additional insights could be gained, besides the ones that were already known. Corpus study two showed a detailed analysis of a series of related research questions using predictive modelling and data science methods, revealing that methodological rigour, however, requires the use of more complex modelling techniques even for those analysis that do not a priori assume a data-driven approach.

Nevertheless, research in traditional statistical modelling, text mining, machine learning and artificial intelligence continues, and data sources and available methods keep increasing. It is very likely that the presented methods might still become more interesting for corpus linguistics in future time as all current developments point to it.

The trend in machine learning research and artificial intelligence certainly goes towards building ever more powerful, increasingly complex models with *more* abstract concepts (or constructs) being modelled using *more* complex non-linear computations on *more* input variables trained on *much more* data. It is strongly biased towards the predictive use of the models and completely neglected explanatory data mining usages for many years[191]. This led to the extensive use and further development of:

- methods that only work well with big data (which automatically excludes most of the meticulously hand-crafted, linguistically annotated and enriched corpora);
- uninterpretable features (like character n-grams that barely allow to derive insights, even though a list of important discriminative features could be derived);
- complex, non-interpretable black-box models (that do not care about the type of relationship existing between predictor and descriptor variables);
- model evaluation metrics that focus on predictive power but completely neglect interpretability and explanatory power.

Therefore, most of the recently developed methods fall under the category of not intrinsically interpretable black box models. Recent machine learning methods are prevalently types of deep neural networks or, especially in data science, tree ensembles. Some of the most popular methods with particular relevance for natural language processing and text mining are for example variants of recurrent neural networks like *bi-directional long-short-term memory networks*, *generative adversarial networks*, *transformer networks* and networks based on *attention mechanisms*, *multi-task* or *transfer learning*, etc. At first glance, they thus do not seem reasonable for corpus analysis and statistical inference.

However, text mining and data mining approaches are on the rise in many different fields (e.g. business intelligence, biomedicine, social sciences) and new methods are developed at a fast pace. Additionally, the machine learning and artificial intelligence community also experiences a recent need for making the used methods interpretable, so that potential errors can be found, and automatic systems

[191] This development might be changing in future times though.

are trusted (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016b). In particular when deploying predictive systems in real life settings, such as risk prediction for health care decisions or insurance rate calculations, or for cyber security, it is necessary to be able to explain automatically made decisions in order to gain trust and ensure fairness (or avoid systematic bias) (Lepri et al., 2017; Loukina et al., 2019; Sun et al., 2019). Therefore, interpretability and explainability of machine learning models gain more importance also in those communities and methods that can be used for model inspection are being researched (Adadi & Berrada, 2018; Guidotti et al., 2019; Molnar, 2018b)[192].

Apart from the trend towards explainable AI and interpretable machine learning, there are further trends that might become relevant for data science in corpus linguistics. These include emerging techniques to:

- learn from less data;
- artificially produce counter examples for predictions in order to help refining created models;
- do automatic machine learning and automatic feature engineering.

Techniques to learn from less data
After the last years have been spent primarily creating scalable systems that perform well with high amounts of data, recent research in machine learning now investigates how to learn also from less data, the main two strategies being (1) synthesizing new data and (2) transferring a model trained for one task or domain to another (i.e. transfer learning or multi-task learning).

Techniques to artificially produce counter examples for predictions in order to refine created models
A very promising and fast emerging area of machine learning consists of the automatic generation of counter examples for predictions of a predictive system. By generating close, adversarial examples, the learning process of the actual predictive system can be enhanced leading to cyclic, mutually informing networks with ever more precision (i.e. generative adversarial networks) (Goodfellow et al., 2014; Kos et al., 2018). These networks might become relevant also for data mining purposes, as they allow the detection and interpretation of examples and counter examples (Hovy, 2016; Jia & Liang, 2017).

Techniques to do automatic machine learning and automatic feature engineering
Finally, the many approaches to automating the individual steps of text mining might help future corpus linguists to perform systematic model training experiments while needing less technical know-how about the actual implementation of the methods (Feurer et al., 2015; Kanter & Veeramachaneni, 2015; Kotthoff et al., 2017). To date, a lot of background knowledge is still needed to choose the learning algorithm for the task, optimize its parameter settings, and train it with the best feature combination. Work on automatically building and evaluating predictive models can thus help less knowledgeable researchers to apply machine learning methods for their own purposes.

[192] For black box interpretation research specific to NLP see for example Linzen et al. (2019) or Alishahi et al. (2019).

29 Bibliography

Abadi, M./ Barham, P./ Chen, J./ Chen, Z./ Davis, A./ Dean, J./ Devin, M./ Ghemawat, S./ Irving, G./ Isard, M. (2016): Tensorflow: A system for large-scale machine learning. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.

Abbasi, A./ Chen, H. (2007): Categorization and analysis of text in computer mediated communication archives using visualization. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, pp. 11–18.

Abbasi, A./ Chen, H. (2008): CyberGate: a design framework and system for text analysis of computer-mediated communication. *Mis Quarterly*, pp. 811–837.

Abel, A. (2007): Werkstattbericht über das Projekt "Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung." *Linguistik online*, *32* (3/07).

Abel, A./ Glaznieks, A. (Forthc.): Textqualität in sozialen Medien. In: *IDS-Jahrbuch 2019*.

Abel, A./ Glaznieks, A. (2015): Wo Sprachkompetenzforschung auf Varietätenlinguistik trifft: Empirische Befunde aus dem Varietäten-Lernerkorpus "KoKo." In: Lenz, A./ Ahlers, T./ Glauninger, M. (Eds.): *Dimensionen des Deutschen in Österreich. Variation und Varietäten im sozialen Kontext*. Frankfurt a. M.: Peter Lang, pp. 257–282.

Abel, A./ Glaznieks, A. (2017): *KoKo: Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenz bei Lernenden im deutschen Sprachraum – ein Ergebnisbericht*. Bolzano/Bozen, Italy: Eurac Research.

Abel, A./ Glaznieks, A./ Nicolas, L./ Stemle, E. (2016): An extended version of the KoKo German L1 Learner corpus. In: *Proceedings of the Third Italian Conference on Computational Linguistics, CLIC-it 2016*. Naples, Italy: Accademia University Press.

Abel, A./ Glaznieks, A./ Nicolas, L./ Stemle, E. W. (2014): KoKo: An L1 Learner Corpus for German. In: Calzolari, N./ Choukri, K./ Declerck, T./ Loftsson, H./ Maegaard, B./ Mariani, J./ Moreno, A./ Odijk, J./ Stelios, P. (Eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 2414–2421.

Abel, A./ Vettori, C./ Forer, D. (2012): Learning the Neighbour's Language: The Many Challenges in Achieving a Real Multilingual Society. The Case of Second Language Acquisition in the Minority–Majority Context of South Tyrol. In: *European Yearbook of Minority Issues* (Vol. 9). Leiden, Netherlands: Brill Academic Publishers, pp. 2271–2303.

Adadi, A./ Berrada, M. (2018): Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, pp. 52138–52160.

Agarwal, R./ Dhar, V. (2014): Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25* (3), pp. 443–448.

Agrawal, R./ Srikant, R. (1995): Mining sequential patterns. In: *Proceedings of the 11th International Conference on Data Engineering*. Washington, DC, USA: IEEE, pp. 3–14.

Alabbas, W./ Al-Khateeb, H. M./ Mansour, A. (2016): Arabic text classification methods: Systematic literature review of primary studies. In: *Proceedings of the 4th IEEE International Colloquium on Information Science and Technology (CiSt)*. IEEE, pp. 361–367.

Alishahi, A./ Chrupała, G./ Linzen, T. (2019): Analyzing and Interpreting Neural Networks for NLP: A

Report on the First BlackboxNLP Workshop. *Natural Language Engineering*, *25* (4), pp. 543–557.

Alpaydin, E. (2014): *Introduction to machine learning*. Cambridge, Massachusetts / London: MIT press.

Alvarez-Melis, D./ Jaakkola, T. S. (2018): On the Robustness of Interpretability Methods. In: *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, pp. 66–71.

Androutsopoulos, J. (2003): Spaß und Stil im Netz: eine ethnografisch-textanalytische Perspektive. In: Klemm, M./ Jakobs, E. (Eds.): *Das Vergnügen in und an den Medien*. Frankfurt a. M., pp. 223–248.

Androutsopoulos, J. (2006): Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, *10* (4), pp. 419–438.

Androutsopoulos, J. (2013a): Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, *19* (2), pp. 185–205.

Androutsopoulos, J. (2013b): Participatory Culture and Metalinguistic Discourse: Performing and Negotiating German Dialects on YouTube. In: Tannen, D./ Trester, A. M. (Eds.): *Discourse 2.0: Language & New Media*. Georgetown: Georgetown University Press, pp. 47–71.

Androutsopoulos, J. (2014): Languaging when contexts collapse: Audience design in social networking. *Discourse, Context and Media*, *4–5*, pp. 62–73.

Androutsopoulos, J./ Georgakopoulou, A. (Eds.) (2003): *Discourse constructions of youth identities*. Amsterdam/Philadelphia: John Benjamins.

Androutsopoulos, J./ Hsieh, Y. F./ Kouzina, J./ Şahin, R. (2013): Vernetzte Mehrsprachigkeit auf Facebook : Drei Hamburger Fallstudien. In: Redder, A./ Pauli, J./ Kießling, R./ Bührig, K./ Brehmer, B./ Breckner, I./ Androutsopoulos, J. (Eds.): *Mehrsprachige Kommunikation in der Stadt - Das Beispiel Hamburg*. Münster: Waxmann, pp. 161–198.

Anstein, S. (2013): *Computational approaches to the comparison of regional variety corpora: prototyping a semi-automatic system for German*. PhD dissertation, Universität Stuttgart.

Anthony, L. (2013): A critical look at software tools in corpus linguistics. *Linguistic Research*, *30* (2), pp. 141–161.

Apley, D. W. (2016): Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Retrieved from https://arxiv.org/abs/1612.08468.

Arakawa, Y./ Kameda, A./ Aizawa, A./ Suzuki, T. (2014): Adding Twitter-specific features to stylistic features for classifying tweets by user type and number of retweets. *Journal of the Association for Information Science and Technology*, *65* (7), pp. 1416–1423.

Argamon, S./ Koppel, M./ Pennebaker, J. W./ Schler, J. (2009): Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52* (2), pp. 119–123.

Arras, L./ Horn, F./ Montavon, G./ Müller, K.-R./ Samek, W. (2017): "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLoS ONE*, *12* (8), p. article e0181142.

Ashok, V. G./ Feng, S./ Choi, Y. (2013): Success with style: Using writing style to predict the success of novels. *Poetry*, *580* (9), pp. 1753–1764.

ASTAT (2016): South Tyrol in figures. (Gobbi, G./ Thurner, B., Eds.). Bozen/Bolzano: Autonomous Province of South Tyrol. Provincial Statistics Institute - ASTAT.

Augst, G./ Disselhoff, K./ Henrich, A./ Pohl, T./ Völzing, P.-L. (2007): *Text-Sorten-Kompetenz: eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter*. Frankfurt a. M.: Peter Lang.

Baayen, R. H./ Cutler, A. (2005): Data mining at the intersection of psychology and linguistics. *Twenty-first century psycholinguistics: Four cornerstones*, pp. 69–83.

Baker, P. (2010): *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Baker, R./ Inventado, P. S. (2014): Educational data mining and learning analytics. In: *Learning analytics*. Springer, pp. 61–75.

Ball, A. (2012): *Review of data management lifecycle models*. Bath, UK: University of Bath, IDMRC.

Bamman, D./ Eisenstein, J./ Schnoebelen, T. (2014): Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18* (2), pp. 135–160.

Baron, A./ Rayson, P./ Greenwood, P./ Walkerdine, J./ Rashid, A. (2012): Children Online: A survey of child language and CMC corpora. *International Journal of Corpus Linguistics*, *17* (4), pp. 443–481.

Baroni, M. (2009): Distributions in text. In: Lüdeling, A./ Kytö, M. (Eds.): *Corpus linguistics: An international handbook* (Vol. 2). Berlin: Mouton de Gruyter, pp. 803–821.

Baroni, M./ Bernardini, S. (2006): A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21* (3), pp. 259–274.

Baroni, M./ Evert, S. (2009): Statistical methods for corpus exploitation. In: Lüdeling, A./ Kytö, M. (Eds.): *Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin* (Vol. 2). Mouton de Gruyter, pp. 777–803.

Barth, D./ Kapatsinski, V. (2018): Evaluating logistic mixed-effects models of corpus-linguistic data. In: *Mixed-Effects Regression Models in Linguistics*. Cham: Springer, pp. 99–116.

Bartz, T./ Pölitz, C./ Morik, K./ Storrer, A. (2015): Using Data Mining and the CLARIN Infrastructure to Extend Corpus-based Linguistic Research. In: *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*. Linköping University Electronic Press, pp. 1–13.

Barua, A./ Thomas, S. W./ Hassan, A. E. (2014): What are developers talking about? An analysis of topics and trends in stack overflow. *Empirical Software Engineering*, *19* (3), pp. 619–654.

Barzilay, R./ Lapata, M. (2008): Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34* (1), pp. 1–34.

Bayne, S./ Ross, J. (2011): Digital Native and Digital Immigrant Discourses. In: Land, R./ Bayne, S. (Eds.): *Digital Difference: Perspectives on Online Learning*, pp. 159–169.

Becker-Mrotzek, M./ Böttcher, I. (2006): *Schreibkompetenz entwickeln und beurteilen: Praxishandbuch für die Sekundarstufe I und II*. Cornelsen Scriptor.

Ben-David, A. (2008): Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications*, *34* (2), pp. 825–832.

Bereiter, C. (1980): Development in writing. *Cognitive processes in writing*, pp. 73–93.

Berkling, K./ Fay, J./ Ghayoomi, M./ Hein, K./ Lavalley, R./ Linhuber, L./ Stüker, S. (2014): A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification. In: *Proceedings of the Ninth International Conference on Language Resources and*

Evaluation (LREC '14). Reykjavik, Iceland: European Language Resources Association, pp. 1212–1217.

Berland, M./ Baker, R. S./ Blikstein, P. (2014): Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19* (1–2), pp. 205–220.

Bernaisch, T./ Gries, S. T./ Mukherjee, J. (2014): The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide: A Journal of Varieties of English*, *35* (1), pp. 7–31.

Biber, D. (1993a): Representativeness in corpus design. *Literary and linguistic computing*, *8* (4), pp. 243–257.

Biber, D. (1993b): The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, *26* (5–6), pp. 331–345.

Biber, D. (2014): Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14* (1), pp. 7–34.

Biber, D./ Jones, J. K. (2009): Quantitative methods in corpus linguistics. In: Lüdeling, A./ Kytö, M. (Eds.): *Corpus linguistics: An international handbook* (Vol. 2). Berlin: Mouton de Gruyter, pp. 1286–1304.

Biber, D./ Reppen, R. (2015): *The Cambridge Handbook of English Corpus Linguistics*. Cambridge, UK: Cambridge University Press.

Biecek, P. (2018): DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research*, *19* (1), pp. 3245–3249.

Bien, J./ Tibshirani, R. (2011): Prototype selection for interpretable classification. *The Annals of Applied Statistics*, *5* (4), pp. 2403–2424.

Blei, D. M./ Lafferty, J. D. (2009): Topic models. In: Srivastava, A. N./ Sahami, M. (Eds.): *Text Mining*. New York: Chapman and Hall/CRC, pp. 101–124.

Blei, D. M./ Ng, A. Y./ Jordan, M. I. (2003): Latent dirichlet allocation. *Journal of machine Learning research*, *3* (Jan), pp. 993–1022.

Bolander, B. (2017): Language and identity on Facebook. In: Thorne, S. L./ May, S. (Eds.): *Language, Education and Technology. Encyclopedia of Language and Education*. Springer International Publishing.

Bonvin, A./ Lambelet, A. (2017): Algorithmic and subjective measures of lexical diversity in bilingual written corpora: A discussion. *Corela*, (HS-21), p. article 4843.

Boyd, D. (2008): Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pp. 119–142.

Boyd, D./ Crawford, K. (2012): Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15* (5), pp. 662–679.

Breiman, L. (2001a): Random forests. *Machine learning*, *45* (1), pp. 5–32.

Breiman, L. (2001b): Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16* (3), pp. 199–231.

Breiman, L./ Friedman, J./ Stone, C./ Olshen, R. A. (1984): *Classification and regression trees*. Boca

Ration, London, New York, Washington DC: Chapman and Hall/CRC.

Breindl, E./ Volodina, A./ Waßner, U. H. (2014): *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer* (Vol. 13). Berlin, München, Boston: Walter de Gruyter.

Brinker, K. (2010): *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden. Bearbeitet von Sandra Ausborn-Brinker* (7. Aufl.). Berlin, Germany: Erich Schmidt.

Britton, M. (2019): VINE: Visualizing Statistical Interactions in Black Box Models. Retrieved from https://arxiv.org/abs/1904.00561.

Brown, G. W. (1984): Discriminant analysis. *American journal of diseases of children (1911)*, *138* (4), pp. 395–400.

Brysbaert, M./ Keuleers, E./ New, B. (2011): Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, *2*, p. article 27.

Bubenhofer, N. (2009): *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs-und Kulturanalyse*. Berlin/ New York: de Gruyter.

Bucher, T. (2012): Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New media & society*, *14* (7), pp. 1164–1180.

Burnap, P./ Rana, O. F./ Avis, N./ Williams, M./ Housley, W./ Edwards, A./ Morgan, J./ Sloan, L. (2015): Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*, pp. 96–108.

Buttery, P./ Caines, A./ Tono, Y./ Kawaguchi, Y./ Minegishi, M. (2012): Normalising frequency counts to account for "opportunity of use" in learner corpora. *Developmental and crosslinguistic perspectives in learner corpus research*, pp. 187–204.

Bykh, S. (2017): *"Is There Choice in Non-Native Voice?" Linguistic Feature Engineering and a Variationist Perspective in Automatic Native Language Identification*. PhD dissertation, Eberhard Karls Universität Tübingen.

Bykh, S./ Meurers, D. (2016): Advancing Linguistic Features and Insights by Label-informed Feature Grouping: An Exploration in the Context of Native Language Identification. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 739–749.

Byrd, K./ Mansurov, A./ Baysal, O. (2016): Mining twitter data for influenza detection and surveillance. In: *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*. ACM, pp. 43–49.

Caliskan, A./ Bryson, J. J./ Narayanan, A. (2017): Semantics derived automatically from language corpora contain human-like biases. *Science*, *356* (6334), pp. 183–186.

Cao, Q./ Duan, W./ Gan, Q. (2011): Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50* (2), pp. 511–521.

Casalicchio, G./ Molnar, C./ Bischl, B. (2018): Visualizing the feature importance for black box models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–670.

Chambers, J. (2003): *Sociolinguistic theory: linguistic variation and its social significance*. Malden: Blackwell.

Chandrashekar, G./ Sahin, F. (2014): A survey on feature selection methods. *Computers & Electrical Engineering*, *40* (1), pp. 16–28.

Chen, C. (2006): CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, *57* (3), pp. 359–377.

Chen, H./ Chiang, R. H. L./ Storey, V. C. (2012): Business intelligence and analytics: From big data to big impact. *MIS quarterly*, *36* (Sp4), pp. 1165–1188.

Chen, X./ Meurers, D. (2016): CTAP: A web-based tool supporting automatic complexity analysis. In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, Japan: Association for Computational Linguistics, pp. 113–119.

Cheshire, J. (2005): Age and Generation-Specific Use of Language. *Sociolinguistics: An international handbook of the science of language and society*, (2nd Edition), pp. 1552–1563.

Chinkina, M./ Meurers, D. (2016): Linguistically Aware Information Retrieval: Providing Input Enrichment for Second Language Learners. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, California, US: Association for Computational Linguistics, pp. 188–198.

Cimino, A./ Dell'Orletta, F./ Venturi, G./ Montemagni, S. (2013): Linguistic profiling based on general–purpose features and native language identification. In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 207–215.

Cohen, W. W. (1995): Fast effective rule induction. In: *Machine Learning Proceedings 1995*. Elsevier, pp. 115–123.

Corti, L./ Van den Eynden, V./ Bishop, L./ Woollard, M. (2019): *Managing and sharing research data: a guide to good practice*. Los Angeles et al.: SAGE.

Coulthard, M./ Johnson, A./ Wright, D. (2016): *An Introduction to Forensic Linguistics: Language in Evidence* (2nd ed.). London/New York: Routledge.

Council of Europe: *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Coupland, J./ Coupland, N./ Giles, H./ Henwood, K. (1991): Formulating age: Dimensions of age identity in elderly talk. *Discourse Processes*, *14* (1), pp. 87–106.

Coupland, N. (1997): Language, ageing and ageism: A project for applied linguistics? *International Journal of Applied Linguistics*, *7* (1), pp. 26–48.

Cresci, S./ Tesconi, M./ Cimino, A./ Dell'Orletta, F. (2015): A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In: *Proceedings of the 24th International Conference on World Wide Web*. ACM, pp. 1195–1200.

Cristianini, N./ Shawe-Taylor, J. (2000): *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Crocker, M. W./ Demberg, V./ Teich, E. (2015): Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, *30* (1), pp. 77–81.

Croft, W. B./ Lafferty, J. (Eds.) (2003): *Language modeling for information retrieval* (Vol. 13). Dordrecht: Springer Science & Business Media.

Crossley, S. A./ Greenfield, J./ McNamara, D. S. (2008): Assessing text readability using cognitively

based indices. *Tesol Quarterly*, *42* (3), pp. 475–493.

Crossley, S. A./ Kyle, K./ Allen, L. K./ Guo, L./ McNamara, D. S. (2014): Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluationa. *Journal of Writing Assessment*, *7* (1).

Crossley, S. A./ Kyle, K./ McNamara, D. S. (2016): The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, *48* (4), pp. 1227–1237.

Crossley, S. A./ Kyle, K./ Varner, L./ McNamara, D. S. (2014): The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In: *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 300–303.

Crossley, S. A./ McNamara, D. (2010): Cohesion, coherence, and expert evaluations of writing proficiency. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32), pp. 984–989.

Crossley, S. A./ McNamara, D. (2011a): Text coherence and judgments of essay quality: Models of quality and coherence. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33), pp. 1236–1241.

Crossley, S. A./ McNamara, D. S. (2009): Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18* (2), pp. 119–135.

Crossley, S. A./ McNamara, D. S. (2011b): Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21* (2–3), pp. 170–191.

Crossley, S. A./ McNamara, D. S. (2016): Say More and Be More Coherent: How Text Elaboration and Cohesion Can Increase Writing Quality. *Journal of Writing Research*, *7* (3), pp. 351–370.

Crossley, S. A./ Roscoe, R./ McNamara, D. S. (2011): Predicting human scores of essay quality using computational indices of linguistic and textual features. In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 438–440.

Crossley, S. A./ Roscoe, R./ McNamara, D. S. (2014): What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, *31* (2), pp. 184–214.

Crossley, S. A./ Varner, L. K./ Roscoe, R. D./ McNamara, D. S. (2013): Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 269–278.

Crossley, S. A./ Weston, J. L./ McLain Sullivan, S. T./ McNamara, D. S. (2011): The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28* (3), pp. 282–311.

Crystal, D. (2001): *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2011): *Internet linguistics: A student guide*. London/New York: Routledge.

Daelemans, W. (2013): Explanation in computational stylometry. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 451–462.

Daelemans, W./ Berck, P./ Gillis, S. (1997): Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica*, *31* (1–2), pp. 57–76.

Daelemans, W./ Hoste, V. (2002): Evaluation of machine learning methods for natural language processing tasks. In: *Proceedings of the 3rd International conference on Language Resources and Evaluation (LREC 2002)*. European Language Resources Association (ELRA), pp. 755–760.

Daelemans, W./ Zavrel, J./ Van Der Sloot, K./ Van Den Bosch, A. (2007): *Timbl: Tilburg memory-based learner. Version 6.3. Reference Guide*. *ILK Technical Report - ILK 10-01* (Vol. 6).

Dalip, D. H./ Gonçalves, M. A./ Cristo, M./ Calado, P. (2013): Exploiting user feedback to learn to rank answers in Q&A forums: A case study with stack overflow. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 543–552.

Danescu-Niculescu-Mizil, C./ West, R./ Jurafsky, D./ Potts, C. (2013): No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In: *Proceedings of the 22nd international conference on World Wide Web*. Rio de Janeiro, Brazil, pp. 307–318.

Danet, B./ Herring, S. C. (2007): *The multilingual Internet: Language, culture, and communication online*. Oxford: Oxford University Press.

Davidson, T./ Warmsley, D./ Macy, M./ Weber, I. (2017): Automated hate speech detection and the problem of offensive language. In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. Association for the Advancement of Artificial Intelligence, pp. 512–515.

de Marneffe, M.-C./ Potts, C. (2017): Developing linguistic theories using annotated corpora. In: Ide, N./ Pustejovsky, J. (Eds.): *Handbook of Linguistic Annotation*. Dordrecht: Springer, pp. 411–438.

Degaetano-Ortlieb, S. (2015): *Evaluative meaning in scientific writing: macro-and micro-analytic perspectives using data mining*. PhD dissertation, Universität des Saarlandes.

Degaetano-Ortlieb, S./ Fankhauser, P./ Kermes, H./ Lapshinova-Koltunski, E./ Ordan, N./ Teich, E. (2014): Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik, Iceland: European Language Resources Association, pp. 1327–1334.

Degaetano-Ortlieb, S./ Kermes, H./ Khamis, A./ Teich, E. (2019): An information-theoretic approach to modeling diachronic change in scientific English. *From Data to Evidence in English Language Research*, pp. 258–281.

Degaetano-Ortlieb, S./ Teich, E. (2016): Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 165–173.

Degaetano-Ortlieb, S./ Teich, E. (2017): Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 68–77.

Dell'Orletta, F./ Montemagni, S./ Venturi, G. (2011): Read-it: Assessing readability of Italian texts with a view to text simplification. In: *Proceedings of the second workshop on speech and language processing for assistive technologies*. Association for Computational Linguistics, pp. 73–83.

Demšar, J. (2006): Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, *7* (Jan), pp. 1–30.

Desagulier, G. (2014): Visualizing distances in a set of near synonyms: rather, quite, fairly, and pretty. In: Glynn, D./ Robinson, J. A. (Eds.): *Corpus Methods for Semantics: Quantitative studies in*

polysemy and synonymy, pp. 145–178.

Desagulier, G. (2017): *Corpus Linguistics and Statistics with R : Introduction to Quantitative Methods in Linguistics*. Cham: Springer.

Desagulier, G. (2018): Multifactorial Exploratory Approaches. *Preprint*. Retrieved from https://halshs.archives-ouvertes.fr/halshs-01926339/.

Deshors, S. C./ Gries, S. (Forthc.): Mandative subjuntive vs. should in world Englishes: A new take on an old alternation. *Corpora*, *15* (2).

DeVito, M. A. (2017): From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism*, *5* (6), pp. 753–773.

Dhar, V. (2013): Data science and prediction. *Communications of the ACM*, *56* (12), pp. 64–73.

Dikli, S. (2006): An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5* (1).

Don, A./ Zheleva, E./ Gregory, M./ Tarkan, S./ Auvil, L./ Clement, T./ Shneiderman, B./ Plaisant, C. (2007): Discovering interesting usage patterns in text collections: integrating text mining with visualization. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, pp. 213–222.

Dong, G./ Li, J. (1999): Efficient mining of emerging patterns: Discovering trends and differences. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 43–52.

Doran, D./ Schulz, S./ Besold, T. R. (2018): What does explainable AI really mean? A new conceptualization of perspectives. *CEUR Workshop Proceedings*, *2071*.

Dörnyei, Z. (2009): *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Doshi-Velez, F./ Kim, B. (2017): *Towards a rigorous science of interpretable machine learning*. Retrieved from https://arxiv.org/abs/1702.08608.

Došilović, F. K./ Brčić, M./ Hlupić, N. (2018): Explainable artificial intelligence: A survey. In: *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, pp. 210–215.

Downes, W. (1998): *Language and society*. Cambridge: Cambridge University Press.

Duden (2005): *Die Grammatik*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Duden (2006): *Die deutsche Rechtschreibung*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Dunham, M. H. (2006): *Data mining: Introductory and advanced topics*. Pearson Education.

Dunn, J. (2019): Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology. *Frontiers in Artificial Intelligence*, *2*.

Durán, P./ Malvern, D./ Richards, B./ Chipere, N. (2004): Developmental trends in lexical diversity. *Applied Linguistics*, *25* (2), pp. 220–242.

Dürscheid, C. (2005): Medien, Kommunikationsformen, kommunikative Gattungen. *Linguistik online*, *22* (1/05), pp. 3–16.

Dürscheid, C./ Wagner, F./ Brommer, S. (2010): *Wie Jugendliche schreiben. Schreibkompetenz und neue Medien*. Berlin/New York: de Gruyter.

East, M. (2009): Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing writing*, *14* (2), pp. 88–115.

Eckert, P. (1997): Age as a sociolinguistic variable. In: Coulmas, F. (Ed.): *The Handbook of Sociolinguistics*. Oxford, Malden: Blackwell, pp. 151–167.

Edwards, J. R./ Bagozzi, R. P. (2000): On the nature and direction of relationships between constructs and measures. *Psychological methods*, *5* (2), pp. 155–174.

Eichinger, L. M. (2001): Die soziolinguistische Situation der deutschen Sprachgruppe in Südtirol. In: Egger, K./ Lanthaler, F. (Eds.): *Die deutsche Sprache in Südtirol: Einheitssprache und regionale Vielfalt*. Wien/Bozen: Folio, pp. 121–136.

Eichinger, L. M. (2002): South Tyrol: German and Italian in a changing world. *Journal of Multilingual and Multicultural Development*, *23* (1–2), pp. 137–149.

Ellson, J./ Gansner, E./ Koutsofios, L./ North, S. C./ Woodhull, G. (2001): Graphviz—open source graph drawing tools. In: *International Symposium on Graph Drawing*. Springer, pp. 483–484.

Emani, C. K./ Cullot, N./ Nicolle, C. (2015): Understandable big data: A survey. *Computer science review*, *17*, pp. 70–81.

Estival, D./ Gaustad, T./ Pham, S. B./ Radford, W./ Hutchinson, B. (2007): Author profiling for English emails. In: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)*, pp. 263–272.

Farag, Y./ Yannakoudakis, H./ Briscoe, T. (2018): Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 263–271.

Faris, H./ Al-Zoubi, A. M./ Heidari, A. A./ Aljarah, I./ Mafarja, M./ Hassonah, M. A./ Fujita, H. (2019): An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks. *Information Fusion*, *48*, pp. 67–83.

Fayyad, U./ Piatetsky-Shapiro, G./ Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17* (3), pp. 37–54.

Feilke, H. (2010): Schriftliches Argumentieren zwischen Nähe und Distanz - am Beispiel wissenschaftlichen Schreibens. In: Àgel, V./ Hennig, M. (Eds.): *Nähe und Distanz im Kontext variationslinguistischer Forschung*. Berlin/New York: Walter de Gruyter, pp. 209–231.

Feldman, R./ Sanger, J. (2007): *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Ferguson, C. A. (1959): Diglossia. *Word*, *15*, pp. 325–340.

Feurer, M./ Klein, A./ Eggensperger, K./ Springenberg, J./ Blum, M./ Hutter, F. (2015): Efficient and robust automated machine learning. In: Cortes, C./ Lawrence, N. D./ Lee, D. D./ Sugiyama, M./ Garnett, R. (Eds.): *Advances in neural information processing systems 28 (NIPS)*. Curran Associates, pp. 2962–2970.

Finlay, S. (2014): *Predictive analytics, data mining and big data: Myths, misconceptions and methods*. Palgrave Macmillan.

Finn, A./ Kushmerick, N. (2006): Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, *57* (11), pp. 1506–1518.

Fischer, J. L. (2015): Social Influences on the Choice of a Linguistic Variant. *Word*, *14* (1), pp. 47–56.

Fisher, A./ Rudin, C./ Dominici, F. (2018): All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. Retrieved from http://arxiv.org/abs/1801.01489.

Flekova, L./ Preoţiuc-Pietro, D./ Ungar, L. (2016): Exploring Stylistic Variation with Age and Income on Twitter. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 313–319.

Flesch, R. (1948): A new readability yardstick. *Journal of applied psychology*, *32* (3), pp. 221–233.

Francis, W. N./ Kucera, H. (1967): *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Frank, E./ Witten, I. H. (1998): Generating accurate rule sets without global optimization. In: Shavlik, J. (Ed.): *Proceedings of the Fifteenth International Conference of Machine Learning*. San Francicso, USA: Morgan Kaufmann, pp. 144–151.

Freitas, A. A. (2014): Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, *15* (1), pp. 1–10.

Frey, J.-C. (2018, February 3): Pluriliteracy on Social Media. The Multilingual Practices of South Tyroleans on Facebook. *Presentation at Language, Identity and Education in Multilingual Contexts, 2-4.2.2018, Marino Institute of Education, Dublin, Ireland*. Dublin, Ireland.

Frey, J.-C./ Ebner, M./ Schön, M./ Taraghi, B. (2013): Social media usage at universities: How should it be done? In: *Proceedings of the 9th International Conference on Web Information Systems and Technologies (WEBIST 2013)*. Aachen, pp. 608–616.

Frey, J.-C./ Glaznieks, A. (2018): The myth of the Digital Native? Analysing language use of different generations in Facebook. In: Vandekerckhove, R./ Fišer, D./ Hilte, L. (Eds.): *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*. Antwerp, Belgium: University of Antwerp, pp. 41–44.

Frey, J.-C./ Glaznieks, A./ Stemle, E. W. (2015): The DiDi Corpus of South Tyrolean CMC Data. In: Beißwenger, M./ Zesch, T. (Eds.): *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media, University of Duisburg-Essen, September 28, 2015.* Essen, Germany: German Society for Computational Linguistics & Language Technology, pp. 1–6.

Frey, J.-C./ Glaznieks, A./ Stemle, E. W. (2016): The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In: Corazza, A./ Montemagni, S./ Seneraro, G. (Eds.): *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), 5-6 December 2016, Napoli*. Torino, Italy: Accademia University Press, pp. 157–161.

Frey, J.-C./ Glaznieks, A./ Stemle, E. W./ Glaznieks, A. (2014): Collecting language data of non-public social media profiles. In: Faaß, G./ Ruppenhofer, J. (Eds.): *Workshop Proceedings of the 12th Edition of the KONVENS Conference (2014)*. Hildesheim, Germany: Universitatsverlag Hildesheim, Germany, pp. 11–15.

Friedman, J. H. (2001): Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29* (5), pp. 1189–1232.

Friedman, J. H./ Popescu, B. E. (2008): Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2* (3), pp. 916–954.

Fuhrhop, N. (2005): *Orthographie*. Heidelberg: Universitätsverlag Winter.

Galasso, S. (2014): *Exploring Textual Cohesion Characteristics for German Readability Classification*. Bachelor's thesis, Eberhard-Karls-Universität Tübingen.

Geisser, S. (2017): *Predictive inference*. New York: Routledge.

Geoffrey Leech (1996): The state of the art in Corpus Linguistics. *English Corpus Linguistics*, (3), pp. 8–29.

Georgalou, M. (2015a): Beyond the Timeline: Constructing time and age identities on Facebook. *Discourse, Context and Media*, *9*, pp. 24–33.

Georgalou, M. (2015b): 'I make the rules on my Wall': Privacy and identity management practices on Facebook. *Discourse & Communication*, *9* (6), pp. 40–64.

Gibson, E. (2000): The dependency locality theory: A distance-based theory of linguistic complexity. In: Marantz, A./ Miyashita, Y./ O'Neil, W. (Eds.): *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*. Cambridge, MA: MIT Press, pp. 95–126.

Gilpin, L. H./ Bau, D./ Yuan, B. Z./ Bajwa, A./ Specter, M./ Kagal, L. (2018): Explaining explanations: An overview of interpretability of machine learning. In: *Proceedings of the 5th International Conference on Data Science and Advanced Analytics (DSAA 2018)*, pp. 80–89.

Glaznieks, A./ Frey, J.-C. (Forthc.): Das DiDi-Korpus. Internetbasierte Kommunikation aus Südtirol. In: *IDS-Jahrbuch 2019*. Berlin, Germany: De Gruyter.

Glaznieks, A./ Frey, J.-C. (2018): Dialekt als Norm? Zum Sprachgebrauch Südtiroler Jugendlicher auf Facebook. In: Ziegler, A. (Ed.): *Jugendsprachen. Aktuelle Perspektiven Internationaler Forschung*. Berlin, Germany: De Gruyter, pp. 859–890.

Glaznieks, A./ Glück, A. (2019): From the Valleys to the World Wide Web: Non-Standard Spellings on Social Network Sites. In: Stemle, E. W./ Wigham, C. R. (Eds.): *Building computer-mediated communication corpora for sociolinguistic analysis*. Clermont Auvergne: Clermont Auvergne University Publishing, pp. 12–27.

Glaznieks, A./ Stemle, E. W. (2014): Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *Journal for Language Technology and Computational Linguistics (JLCL)*, *29* (2), pp. 31–57.

Goldbloom, A. (2016, January 13): How to win a Kaggle competition. *Interview with CEO of Kaggle for import.io*. Retrieved from https://www.import.io/post/how-to-win-a-kaggle-competition/.

Goldstein, A./ Kapelner, A./ Bleich, J./ Pitkin, E. (2015): Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24* (1), pp. 44–65.

Goodfellow, I./ Bengio, Y./ Courville, A. (2016): *Deep learning*. Cambridge, MA: MIT press.

Goodfellow, I./ Pouget-Abadie, J./ Mirza, M./ Xu, B./ Warde-Farley, D./ Ozair, S./ Courville, A./ Bengio, Y. (2014): Generative adversarial nets. In: Ghahramani, Z./ Welling, M./ Cortes, C./ Lawrence, N. D./ Weinberger, K. Q. (Eds.): *Advances in neural information processing systems 27 (NIPS 2014)*. Curran Associates, pp. 2672–2680.

Goodman, B./ Flaxman, S. (2017): European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine*, *38* (3), pp. 50–57.

Goswami, S./ Sarkar, S./ Rustagi, M. (2009): Stylometric analysis of bloggers' age and gender. In: *Third International AAAI Conference on Weblogs and Social Media*, pp. 214–217.

Grabowski, J./ Becker-Mrotzek, M./ Knopp, M./ Jost, J./ Weinzierl, C. (2014): Comparing and combining different approaches to the assessment of text quality. In: Knorr, D./ Heine, C./ Engberg, I. (Eds.): *Methods in Writing Process Research*. Frankfurt a. M.: Peter Lang, pp. 147–165.

Graham, M./ Milanowski, A./ Miller, J. (2012): *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.* Center for Educator Compensation Reform.

Greenwell, B. M./ Boehmke, B. C./ Mccarthy, A. J. (2018): A Simple and Effective Model-Based Variable Importance Measure. Retrieved from https://arxiv.org/abs/1805.04755, pp. 1–27.

Gries, S. (Forthc.): On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*.

Gries, S. (2009): *Quantitative Corpus Linguistics with R*. New York/London: Routledge.

Gries, S. (2013): Statistics for Linguistics with R: A Practical Introduction. Berlin/Boston: De Gruyter Mouton.

Gries, S. (2015a): Quantitative designs and statistical techniques. In: Biber, D./ Reppen, R. (Eds.): *The Cambridge Handbook of English Corpus Linguistics*. Cambridge, UK: Cambridge University Press, pp. 50–71.

Gries, S. (2015b): Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics*, *16* (1), pp. 93–117.

Gries, S. (2015c): The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, *10* (1), pp. 95–125.

Gries, S. (2015d): The role of quantitative methods in cognitive linguistics. *Applications of Cognitive Linguistics*, *31* (ii), pp. 311–326.

Gries, S. (2018): On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1* (2), pp. 276–308.

Gries, S. T./ Deshors, S. C. (2014): Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9* (1), pp. 109–136.

Gries, S. T./ Durrant, P. (Forthc.): Analyzing co-occurrence data. *Practical handbook of corpus linguistics*.

Gries, S. T./ Newman, J. (2014): Creating and using corpora. *Research Methods in Linguistics*, *2*, pp. 257–287.

Gröndahl, T./ Asokan, N. (2019): Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *ACM Computing Surveys (CSUR)*, *52* (3).

Guerini, M./ Pepe, A./ Lepri, B. (2012): Do linguistic style and readability of scientific abstracts affect their virality? In: *Sixth International AAAI Conference on Weblogs and Social Media*, pp. 475–478.

Guidotti, R./ Monreale, A./ Ruggieri, S./ Pedreschi, D./ Turini, F./ Giannotti, F. (2018): Local Rule-Based Explanations of Black Box Decision Systems. Retrieved from http://arxiv.org/abs/1805.10820.

Guidotti, R./ Monreale, A./ Ruggieri, S./ Turini, F./ Pedreschi, D./ Giannotti, F. (2019): A Survey Of Methods For Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, *51* (5).

Gunning, R. (1968): *Technique of clear writing*. New York: McGraw-Hill.

Gurumoorthy, K. S./ Dhurandhar, A./ Cecchi, G. (2017): Protodash: fast interpretable prototype

selection. Retrieved from https://arxiv.org/abs/1707.01212.

Guyon, I./ Elisseeff, A. (2003): An introduction to variable and feature selection. *Journal of machine learning research*, *3*, pp. 1157–1182.

Haas, C./ Takayoshi, P./ Carr, B./ Hudson, K./ Pollock, R. (2011): Young people's everyday literacies: The language features of instant messaging. *Research in the Teaching of English*, *45* (4), pp. 378–404.

HaCohen-Kerner, Y./ Beck, H./ Yehudai, E./ Mughaz, D. (2010): Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin. *Applied Artificial Intelligence*, *24* (9), pp. 847–862.

Haig, B. D. (2018): An Abductive Theory of Scientific Method. In: *Method Matters in Psychology*. Cham: Springer, pp. 35–64.

Hall, M. A. (1998): *Correlation-based feature subset selection for machine learning*. PhD dissertation, University of Waikato.

Hall, M./ Frank, E./ Holmes, G./ Pfahringer, B./ Reutemann, P./ Witten, I. H. (2009): The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, *11* (1), pp. 10–18.

Han, J./ Kamber, M./ Pei, J. (2012): *Data mining: Concepts and techniques* (3rd ed.). Amsterdam et al.: Elsevier.

Hancke, J./ Meurers, D. (2013): Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research*, pp. 54–56.

Hancke, J./ Vajjala, S./ Meurers, D. (2012): Readability Classification for German using Lexical, Syntactic, and Morphological Features. In: *Proceedings of COLING 2012*. Mumbai, India, pp. 1063–1080.

Harish, B. S./ Guru, D. S./ Manjunath, S. (2010): Representation and classification of text documents: A brief review. *IJCA Special Issue on Recent Trends in Image Processing and Pattern recognition (RTIPPR, 2010)*, pp. 110–119.

Harsch, C./ Martin, G. (2012): Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17* (4), pp. 228–250.

Harsch, C./ Martin, G. (2013): Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20* (3), pp. 281–307.

Hashemi, M. R./ Babaii, E. (2013): Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal*, *97* (4), pp. 828–852.

Hastie, T./ Tibshirani, R./ Friedman, J. (2011): *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, Berlin.

Hayes, J. R./ Flower, L. S. (1986): Writing research and the writer. *American psychologist*, *41* (10), pp. 1106–1113.

He, W./ Zha, S./ Li, L. (2013): Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33* (3), pp. 464–472.

Hearst, M. A. (1999): Untangling text data mining. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pp. 3–10.

Hearst, M. A. (2003): Text data mining. In: Mitkov, R. (Ed.): *The Oxford Handbook of Computational*

Linguistics. Oxford, UK: Oxford University Press, pp. 616–628.

Hechtlinger, Y. (2016): Interpretation of prediction models using the input gradient. In: *29th Conference on Neural Information Processing Systems, Interpretable Machine Learning Workshop (NIPS 2016)*. Barcelona, Spain.

Heister, J./ Würzner, K.-M./ Bubenzer, J./ Pohl, E./ Hanneforth, T./ Geyken, A./ Kliegl, R. (2011): dlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*, pp. 10–20.

Herring, S. C./ Kapidzic, S. (2015): Teens, gender, and self-presentation in social media. In: *International Encyclopedia of Social and Behavioral Sciences* (2nd ed.). Oxford: Elsevier, pp. 1–16.

Hey, T./ Tansley, S./ Tolle, K. M. (2009): *Jim Gray on eScience: A transformed scientific method*. *Article based on the transcript of a talk given by Jim Gray to the NRC-CSTB*. Microsoft research.

Hilte, L./ Vandekerckhove, R./ Daelemans, W. (2016): Expressiveness in Flemish Online Teenage Talk: A Corpus-Based Analysis of Social and Medium-Related Linguistic Variation. In: *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia*, pp. 30–33.

Hilte, L./ Vandekerckhove, R./ Daelemans, W. (2017): Modeling non-standard language use in adolescents' CMC: the impact and interaction of age, gender and education. In: Wigham, C./ Stemle, E. (Eds.): *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17), 3-4 October 2017*. Bolzano, Italy: Eurac Research, pp. 11–15.

Hilte, L./ Vandekerckhove, R./ Daelemans, W. (2018): Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics*, *7* (1), pp. 2–25.

Hinrichs, E./ Krauwer, S. (2014): The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 1525–1531.

Holte, R. C. (1993): Very simple classification rules perform well on most commonly used datasets. *Machine learning*, *11* (1), pp. 63–90.

Holzinger, A./ Schantl, J./ Schroettner, M./ Seifert, C./ Verspoor, K. (2014): Biomedical text mining: State-of-the-art, open problems and future challenges. In: *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, pp. 271–300.

Honegger, M. (2018): Shedding Light on Black Box Machine Learning Algorithms: Development of an Axiomatic Framework to Assess the Quality of Methods that Explain Individual Predictions. Retrieved from http://arxiv.org/abs/1808.05054.

Horbach, A./ Poitz, J./ Palmer, A. (2015): Using shallow syntactic features to measure influences of L1 and proficiency level in EFL writings. In: *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015, Vilnius, 11th May, 2015*. Linköping University Electronic Press, pp. 21–34.

Hota, S. R./ Argamon, S./ Chung, R. (2007): Understanding the Linguistic Construction of Gender in Shakespeare via Text Mining. In: *Digital Humanities 2007. The 19th Joint International Conference of the Association for Computers and the Humanities, and the Assuciation for Literary and Linguistic Computing, University of Illinois*. Urbana-Champaign: University of Illinois, pp. 84–89.

Hothorn, T./ Hornik, K./ Zeileis, A. (2015): ctree: Conditional Inference Trees. *The Comprehensive R*

Archive Network. Vienna: R Foundation for Statistical Computing.

Housen, A./ Kuiken, F. (2009): Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, *30* (4), pp. 461–473.

Hovy, D. (2015): Demographic Factors Improve Classification Performance. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Laanguage Processing, Beijing, China, July 26-31*. Association for Computational Linguistics, pp. 752–762.

Hovy, D. (2016): The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews – An Adversarial Study. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 351–356.

Hovy, D./ Johannsen, A. (2016): Exploring language variation across Europe: A web-based tool for computational sociolinguistics. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2986–2989.

Hovy, D./ Johannsen, A./ Søgaard, A. (2015): User review sites as a resource for large-scale sociolinguistic studies. In: *Proceedings of the 24th international conference on World Wide Web*, pp. 452–461.

Hovy, D./ Spruit, S. L. (2016): The Social Impact of Natural Language Processing. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 591–598.

Huber, J. (2013): *Sprachliche Variation in der SMS-Kommunikation. Eine empirische Untersuchung von deutschsprachigen Schreiberinnen und Schreibern in Südtirol.* Bachelor's thesis, Freie Universität Bozen.

Huber, J./ Schwarz, C. (2017): SMS-Kommunikation im mehrsprachichten Raum. Schriftsprachliche Variation deutschsprachiger SMS-Nutzer/innen in Südtirol. *NETWORX. Die Online-Schriftenreihe des Projekts Sprache@Web*. mediensprache.net.

Humenberger, J. (2012): Wie fluchen und schimpfen Jugendliche? Eine Analyse französischer und italienischer Fluch-und Schimpfwörter. *Wiener Linguistische Gazette*, *76* (A), pp. 168–184.

Hunston, S. (2009): Corpus compilation collection strategies and design decisions. *Corpus linguistics: An international handbook*, *2*, pp. 154–168.

Hutto, C. J./ Gilbert, E. (2014): Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Eighth international AAAI conference on weblogs and social media*. Association for the Advancement of Artificial Intelligence, pp. 216–225.

Ida, M./ Morio, G./ Iwasa, K./ Tatsumi, T./ Yasui, T./ Fujita, K. (2019): Can You Give Me a Reason?: Argument-inducing Online Forum by Argument Mining. In: *Proceedings of the World Wide Web Conference (WWW'19)*. San Diego: ACM, pp. 3545–3549.

Ilisei, I.-N. (2012): *A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models*. PhD dissertation, University of Wolverhampton.

Iria, J. (2012): *Learning for text mining: Tackling the cost of feature and knowledge engineering.* PhD dissertation, University of Sheffield.

Ivaska, I./ Bernardini, S./ Ferraresi, A. (2018): Detecting traces of constrained communication: A corpus-driven approach to mapping of the intersection between learner language and translated language. In: Granger, S./ Lefer, M.-A./ Aguiar de Souza Penha Marion, L. (Eds.): *CECL Papers 1.*

Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition), Louvain-la-Neuve, 12-14 Semptember 2018. Louvain-la-Neuve, Belgium: Université catholique de Louvain, pp. 85–87.

Ivaska, I./ Siitonen, K. (2017): Learner language morphology as a window to crosslinguistic influences: A key structure analysis. *Nordic Journal of Linguistics*, *40* (2), pp. 225–253.

Jaech, A./ Zayats, V./ Fang, H./ Ostendorf, M./ Hajishirzi, H. (2015): Talking to the crowd: What do people react to in online discussions? In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 2026–2031.

James, G./ Witten, D./ Hastie, T./ Tibshirani, R. (2013): *An introduction to statistical learning*. New York et al.: Springer.

Jänicke, S./ Franzini, G./ Cheema, M. F./ Scheuermann, G. (2015): On close and distant reading in digital humanities: A survey and future challenges. In: *Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association*.

Jänicke, S./ Franzini, G./ Cheema, M. F./ Scheuermann, G. (2017): Visual text analysis in digital humanities. In: *Computer Graphics Forum* (Vol. 36). Wiley Online Library, pp. 226–250.

Japkowicz, N./ Shah, M. (2011): *Evaluating learning algorithms: a classification perspective*. Cambridge: Cambridge University Press.

Jarvis, S. (2002): Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19* (1), pp. 57–84.

Jarvis, S. (2011): Data mining with learner corpora. In: Meunier, F./ De Cock, S./ Gilquin, G./ Paquot, M. (Eds.): *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam/Philadelphia: John Benjamins, pp. 127–154.

Jarvis, S. (2012): The detection-based approach: An overview. *Approaching language transfer through text classification: Explorations in the detection-based approach*, pp. 1–33.

Jarvis, S./ Crossley, S. A. (2012): *Approaching Language Transfer Through Text Classification: Explorations in the Detection-Based Approach*. Clevedon/Buffalo/Toronto: Multilingual Matters.

Jechle, T. (1992): *Kommunikatives Schreiben: Prozeß und Entwicklung aus der Sicht kognitiver Schreibforschung*. Tübingen: Gunter Narr Verlag.

Jia, R./ Liang, P. (2017): Adversarial examples for evaluating reading comprehension systems. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2021–2031.

Jockers, M. L. (2013): *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Jockers, M. L./ Mimno, D. (2013): Significant themes in 19th-century literature. *Poetics*, *41* (6), pp. 750–769.

Johnson, K. (2011): *Quantitative methods in linguistics*. Oxford, Malden: Blackwell.

Jonsson, C./ Muhonen, A. (2014): Multilingual repertoires and the relocalization of manga in digital media. *Discourse, Context and Media*, *4–5*, pp. 87–100.

Jordan, M. I./ Mitchell, T. M. (2015): Machine learning: Trends, perspectives, and prospects. *Science*, *349* (6245), pp. 255–260.

Jung, Y./ Crossley, S./ McNamara, D. (2019): Predicting Second Language Writing Proficiency in Learner

Texts Using Computational Tools. *Journal of Asia TEFL*, *16* (1), pp. 37–52.

Kanter, J. M./ Veeramachaneni, K. (2015): Deep feature synthesis: Towards automating data science endeavors. In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE.

Karami, A./ Bennett, L. S./ He, X. (2018): Mining public opinion about economic issues: Twitter and the US presidential election. *International Journal of Strategic Decision Sciences (IJSDS)*, *9* (1), pp. 18–28.

Karlgren, J. (2004): The whys and wherefores for studying textual genre computationally. *Proceedings of the AAAI Fall Symposium of Style and Meaning in Language, Art and Music*.

Katz, G./ Shin, E. C. R./ Song, D. (2017): ExploreKit: Automatic feature generation and selection. *Proceedings of the International Conference on Data Mining, ICDM*, pp. 979–984.

Kelleher, J. D./ Tierney, B. (2018): *Data science*. Cambridge, MA: MIT Press.

Kessler, B./ Nunberg, G./ Schütze, H. (1997): Automatic detection of text genre. In: *35th Annual Meeting of the Assuciation for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, pp. 32–38.

Khurana, U./ Nargesian, F./ Samulowitz, H./ Khalil, E./ Turaga, D. (2016): Automating feature engineering. In: *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain: Curran Associates, pp. 1–2.

Khurana, U./ Turaga, D./ Samulowitz, H./ Parthasrathy, S. (2016): Cognito: Automated feature engineering for supervised learning. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 1304–1307.

Kiela, D. (2016): MMFeat: A Toolkit for Extracting Multi-Modal Features. In: *Proceedings of ACL-2016 System Demonstrations*, pp. 55–60.

Kilgarriff, A./ Grefenstette, G. (2003): Introduction to the special issue on the web as corpus. *Computational linguistics*, *29* (3), pp. 333–347.

Kim, B./ Khanna, R./ Koyejo, O. O. (2016): Examples are not enough, learn to criticize! criticism for interpretability. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. Curran Associates, pp. 2280–2288.

Kim, S.-M./ Hovy, E. (2006): Extracting opinions, opinion holders, and topics expressed in online news media text. In: *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. Association for Computational Linguistics, pp. 1–8.

Kitchin, R. (2013): Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, *3* (3), pp. 262–267.

Kitchin, R. (2014): Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1* (1), pp. 1–12.

Kitchin, R./ Lauriault, T. P. (2015): Small data in the era of big data. *GeoJournal*, *80* (4), pp. 463–475.

Koch, S./ John, M./ Wörner, M./ Müller, A./ Ertl, T. (2014): VarifocalReader—in-depth visual analysis of large text documents. *IEEE transactions on visualization and computer graphics*, *20* (12), pp. 1723–1732.

Koh, P. W./ Liang, P. (2017): Understanding Black-box Predictions via Influence Functions. In:

Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney, Australia: JMLR, pp. 1885–1894.

Kononenko, I./ Šimec, E./ Robnik-Šikonja, M. (1997): Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, *7* (1), pp. 39–55.

Koppel, M./ Schler, J./ Zigdon, K. (2005): Determining an author's native language by mining a text for errors. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, pp. 624–628.

Kos, J./ Fischer, I./ Song, D. (2018): Adversarial examples for generative models. In: *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, pp. 36–42.

Kosinski, M./ Wang, Y./ Lakkaraju, H./ Leskovec, J. (2016): Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21* (4), pp. 493–506.

Kotsiantis, S. B./ Zaharakis, I. D./ Pintelas, P. E. (2006): Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26* (3), pp. 159–190.

Kotsiantis, S./ Zaharakis, I./ Pintelas, P. (2007): Supervised machine learning: A review of classification techniques. *Informatica*, *31*, pp. 249–268.

Kotthoff, L./ Thornton, C./ Hoos, H. H./ Hutter, F./ Leyton-Brown, K. (2017): Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *The Journal of Machine Learning Research*, *18* (1), pp. 826–830.

Krawczyk, B. (2016): Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5* (4), pp. 221–232.

Krishnamoorthy, S. (2015): Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, *42* (7), pp. 3751–3759.

Kroeze, J. H./ Matthee, M. C./ Bothma, T. J. D. (2003): Differentiating data-and text-mining terminology. In: *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*. South African Institute for Computer Scientists and Information Technologists, pp. 93–101.

Kruse, N./ Reichardt, A./ Herrmann, M./ Heinzel, F./ Lipowsky, F. (2012): Zur Qualität von Kindertexten, Entwicklung eines Bewertungsinstruments in der Grundschule. *Didaktik Deutsch*, *18* (32), pp. 87–110.

Kuan, K. K. Y./ Hui, K.-L./ Prasarnphanich, P./ Lai, H.-Y. (2015): What makes a review voted? An empirical investigation of review voting in online review systems. *Journal of the Association for Information Systems*, *16* (1), pp. 48–71.

Kuhn, M. (2015): A Short Introduction to the caret Package. *R Reference Manual*. R Foundation for Statistical Computing.

Kupietz, M./ Diewald, N./ Fankhauser, P. (2018): How to Get the Computation Near the Data: Improving Data Accessibility to, and Reusability of Analysis Functions in Corpus Query Platforms. In: *Proceedings of the LREC 2018 Workshop "Challenges in the Management of Large Corpora" (CMLC-6)*. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 20–25.

Kyle, K./ Crossley, S. A. (2015): Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49* (4), pp. 757–786.

Kytölä, S. (2016): Translocality. In: Georgakopoulou, A./ Spilioti, T. (Eds.): *The Routledge Handbook of*

Language and Digital Communication. Abingdon: Routledge, pp. 371–388.

Labov, W. (1966): *The social stratification of English in New York city* (2nd ed.). Cambridge University Press.

Labov, W. (2001): *Principles of linguistic change* (Vol. 2). Malden, Oxford: Wiley-Blackwell.

Lage, I./ Chen, E./ He, J./ Narayanan, M./ Kim, B./ Gershman, S./ Doshi-Velez, F. (2018): *An Evaluation of the Human-Interpretability of Explanation*. *32nd Conference on Neural Information Processing Systems (NIPS 2018)*. Montreal, Canada.

Lage, I./ Ross, A. S./ Kim, B./ Gershman, S. J./ Doshi-Velez, F. (2018): Human-in-the-Loop Interpretability Prior. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*. Montréal, Canada: Curran Associates, pp. 1–13.

Lam, H. T./ Thiebaut, J.-M./ Sinn, M./ Chen, B./ Mai, T./ Alkan, O. (2017): One button machine for automating feature engineering in relational databases. Retrieved from http://arxiv.org/abs/1706.00327.

Lambiotte, R./ Kosinski, M. (2014): Tracking the digital footprints of personality. In: *Proceedings of the IEEE* (Vol. 102). IEEE, pp. 1934–1939.

Lamurias, A./ Couto, F. M. (2019): Text mining for bioinformatics using biomedical literature. *Encyclopedia of bioinformatics and computational biology*, *1*, pp. 602–611.

Landauer, T. K./ Foltz, P. W./ Laham, D. (1998): An introduction to latent semantic analysis. *Discourse processes*, *25* (2–3), pp. 259–284.

Lanthaler, F. (2001): Zwischenregister der deutschen Sprache Sudtirol. In: Egger, K./ Lanthaler, F. (Eds.): *Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt*. Wien/Bozen: Folio, pp. 137–152.

Lanthaler, F./ Saxalber, A. (1995): Die deutsche Standardsprache in Südtirol. In: Muhr, R./ Schrodt, R./ Wiesinger, P. (Eds.): *Österreichisches Deutsch. Linguistische, sozialpsychologische und sprachliche Aspekte einer nationalen Variante des Deutschen*. Wien: Hölder-Pichler-Tempsky, pp. 289–305.

Lazard, A. J./ Saffer, A. J./ Wilcox, G. B./ Chung, A. D./ Mackert, M. S./ Bernhardt, J. M. (2016): E-cigarette social media messages: A text mining analysis of marketing and consumer conversations on Twitter. *JMIR public health and surveillance*, *2* (2), p. article e171.

LeCun, Y./ Bengio, Y./ Hinton, G. (2015): Deep learning. *Nature*, *521*, pp. 436–444.

Leeper, T. J. (2018): Interpreting regression results using average marginal effects with R's margins. *R package reference manual*. R Foundation for Statistical Computing.

Lenci, A. (2008): Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20* (1), pp. 1–31.

Leppänen, S. (2007): Youth language in media contexts: Insights into the functions of English in Finland. *World Englishes*, *26* (2), pp. 149–169.

Leppänen, S. (2012): Linguistic and discursive heteroglossia on the translocal internet: The case of web writing. In: Sebba, M./ Mahootian, S./ Jonsson, C. (Eds.): *Language mixing and code-switching in writing: Approaches to mixed-language written discourse.* London: Routledge, pp. 233–254.

Leppänen, S./ Pitkänen-Huhta, A./ Piirainen-Marsh, A./ Nikula, T./ Peuronen, S. (2009): Young people's translocal new media uses: A multiperspective analysis of language choice and heteroglossia.

Journal of Computer-Mediated Communication, *14* (4), pp. 1080–1107.

Lepri, B./ Staiano, J./ Sangokoya, D./ Letouzé, E./ Oliver, N. (2017): The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. *Transparent Data Mining for Big and small Data*, pp. 3–24.

Lin, F.-R./ Hsieh, L.-S./ Chuang, F.-T. (2009): Discovering genres of online discussion threads via text mining. *Computers & Education*, *52* (2), pp. 481–495.

Linzen, T./ Chrupała, G./ Belinkov, Y./ Hupkes, D. (Eds.) (2019): *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics.

Lipizzi, C./ Iandoli, L./ Marquez, J. E. R. (2015): Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. *International Journal of Information Management*, *35* (4), pp. 490–503.

Lippi, M./ Torroni, P. (2016): Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, *16* (2), p. article 10.

Lipton, Z. C. (2017): The mythos of model interpretability. In: Kim, B./ Malioutov, D. M./ Varshney, K. R. (Eds.): *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, pp. 96–100.

Lisboa, P. J. G. (2013): Interpretability in Machine Learning – Principles and Practice. In: Masulli, F./ Pasi, G./ Yager, R. (Eds.): *Proceedings of the BT - Fuzzy Logic and Applications: 10th International Workshop, WILF 2013, Genoa, Italy, November 19-22*. Cham: Springer International Publishing, pp. 15–21.

Liu, S./ Wang, X./ Collins, C./ Dou, W./ Ouyang, F./ El-Assady, M./ Jiang, L./ Keim, D. (2019): Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Transactions on Visualization and Computer Graphics*, *25*, pp. 2482–2504.

Lobin, H. (2014): *Engelbarts Traum: Wie der Computer uns Lesen und Schreiben abnimmt*. Frankfurt a. M.: Campus Verlag.

Loper, E./ Bird, S. (2002): NLTK: The Natural Language Toolkit. In: *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 63–70.

Lops, P./ De Gemmis, M./ Semeraro, G. (2011): Content-based Recommender Rystems: State of the Art and Trends. In: Ricci, F./ Rokach, L./ Shapira, B./ Kantor, P. (Eds.): *Recommender systems handbook*. Boston, MA: Springer, pp. 73–105.

Loria, S./ Keen, P./ Honnibal, M./ Yankovsky, R./ Karesh, D./ Dempsey, E. (2014): Textblob: Simplified text processing.

Loukina, A./ Madnani, N./ Zechner, K. (2019): The many dimensions of algorithmic fairness in educational applications. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 1–10.

Louppe, G. (2014): *Understanding Random Forests: From Theory to Practice*. PhD dissertation, Université de Liège.

Louppe, G./ Wehenkel, L./ Sutera, A./ Geurts, P. (2013): Understanding variable importances in forests of randomized trees. *Neural Information Processing Systems*, pp. 1–9.

Lu, X./ Szymanski, B. K. (2018): Scalable prediction of global online media news virality. *IEEE Transactions on Computational Social Systems*, *5* (3), pp. 858–870.

Lucisano, P./ Piemontese, M. E. (1988): GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, *3* (31), pp. 110–124.

Lüdeling, A./ Kytö, M. (2008): *Corpus Linguistics: An International Handbook* (Vol. 1). Berlin, Germaniy: Mouton de Gruyter.

Lui, M./ Baldwin, T. (2012): langid. py: An off-the-shelf language identification tool. In: *Proceedings of the ACL 2012 System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 25–30.

Lundberg, S./ Lee, S.-I. (2017): A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing 30 (NIPS 2017)*. Curran Associates, pp. 4765–4774.

Lynch, G. (2009): *Computational Stylometry and Analysis of Style: A Study of Characterization in Playwrights*. Master's thesis, Trinity College Dublin.

Ma, Z./ Sun, A./ Cong, G. (2013): On predicting the popularity of newly emerging hashtags in Twitter. *Journal of the American Society for Information Science and Technology*, *64* (7), pp. 1399–1410.

Maas, A. L./ Daly, R. E./ Pham, P. T./ Huang, D./ Ng, A. Y./ Potts, C. (2011): Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, pp. 142–150.

Mairesse, F./ Walker, M. A./ Mehl, M. R./ Moore, R. K. (2007): Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of artificial intelligence research*, *30*, pp. 457–500.

Malmasi, S./ Cahill, A. (2015): Measuring Feature Diversity in Native Language Identification. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 49–55.

Manning, C./ Surdeanu, M./ Bauer, J./ Finkel, J./ Bethard, S./ McClosky, D. (2014): The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Mattheier, K. J. (1987): Alter, Generation. In: Ammon, U./ Dittmar, N./ Mattheier, K./ Trudgill, P. (Eds.): *Sociolinguistics. Soziolinguistik*. Berlin/New York, pp. 78–82.

Mc Laughlin, F. (2014): Senegalese digital repertoires in superdiversity: A case study from Seneweb. *Discourse, Context and Media*, *4–5*, pp. 29–37.

Mc Laughlin, G. H. (1969): SMOG grading: A new readability formula. *Journal of reading*, *12* (8), pp. 639–646.

McCarthy, P. M. (2005): *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD dissertation, University of Memphis.

McCarthy, P. M./ Jarvis, S. (2007): vocd: A theoretical and empirical evaluation. *Language Testing*, *24* (4), pp. 459–488.

McCarthy, P. M./ Jarvis, S. (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42* (2), pp. 381–392.

McCullagh, P. (2018): *Generalized linear models*. Routledge.

McCutchen, D. (1996): A capacity theory of writing: Working memory in composition. *Educational psychology review*, *8* (3), pp. 299–325.

McEnery, T./ Hardie, A. (2012): *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T./ Wilson, A. (2001): *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

McEnery, T./ Xiao, R./ Tono, Y. (2006): *Corpus-based language studies: An advanced resource book*. London/New York: Routledge.

McNamara, D. S./ Crossley, S. A./ McCarthy, P. M. (2010): Linguistic features of writing quality. *Written communication*, *27* (1), pp. 57–86.

McNamara, D. S./ Graesser, A. C. (2012): Coh-Metrix: An automated tool for theoretical and applied natural language processing. In: *Applied natural language processing: Identification, investigation and resolution*. IGI Global, pp. 188–205.

Mierswa, I./ Wurst, M./ Klinkenberg, R./ Scholz, M./ Euler, T. (2006): Yale: Rapid prototyping for complex data mining tasks. In: *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 935–940.

Mihalcea, R./ Tarau, P. (2004): Textrank: Bringing order into text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, pp. 404–411.

Mikolov, T./ Chen, K./ Corrado, G./ Dean, J. (2013): Efficient estimation of word representations in vector space. *International Conference on Learning representations: Workshops Track*.

Milin, P./ Divjak, D./ Dimitrijević, S./ Baayen, R. H. (2016): Towards cognitively plausible data science in language research. *Cognitive Linguistics*, *27* (4), pp. 507–526.

Miller, T. (2018): Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, pp. 1–38.

Ming, Y./ Cao, S./ Zhang, R./ Li, Z./ Chen, Y./ Song, Y./ Qu, H. (2018): Understanding Hidden Memories of Recurrent Neural Networks. In: *Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST 2017)*, pp. 13–24.

Mitchell, T. M. (1997): *Machine learning*. Boston, MA: McGraw-Hill.

Mohseni, S./ Zarei, N./ Ragan, E. D. (2018): A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. Retrieved from https://arxiv.org/abs/1811.11839.

Moisl, H. (2015): *Cluster analysis for corpus linguistics*. Berlin, München, Boston: Walter de Gruyter.

Molnar, C. (2018a): iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software*, *3* (27), p. article 786.

Molnar, C. (2018b): *Interpretable machine learning*. *A Guide for Making Black Box Models Explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/.

Molnar, C./ Casalicchio, G./ Bischl, B. (2019): Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition. Retrieved from http://arxiv.org/abs/1904.03867.

Montavon, G./ Samek, W./ Müller, K.-R. (2017): Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, pp. 1–15.

Mooney, R. J. (2003): Machine Learning. In: Mitkov, R. (Ed.): *The Oxford Handbook of Computational Linguistics2*. Oxford, UK: Oxford University Press, pp. 376–394.

Morel, É./ Pekarek-Doehler, S. (2013): Multilingual "texts": Code-switching as an affiliation resource in a globalized community. *Revue Francaise de Linguistique appliquee*, *18* (2), pp. 29–43.

Moretti, F. (2013): *Distant reading*. London/New York: Verso Books.

Murakami, A. (2016): Modeling Systematicity and Individuality in Nonlinear Second Language Development: The Case of English Grammatical Morphemes. *Language Learning*, *66* (4), pp. 834–871.

Nadeem, F./ Nguyen, H./ Liu, Y./ Ostendorf, M. (2019): Automated Essay Scoring with Discourse-Aware Neural Models. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, pp. 484–493.

Narayanan, M./ Chen, E./ He, J./ Kim, B./ Gershman, S./ Doshi-Velez, F. (2018): How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. Retrieved from http://arxiv.org/abs/1802.00682.

Naur, P. (1966): The science of datalogy. *Communications of the ACM*, *9* (7).

Naur, P. (1974): *Concise survey of computer methods*. Lund, Sweden: Studentlitteratur.

Nelder, J. A./ Wedderburn, R. W. M. (1972): Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135* (3), pp. 370–384.

Nesset, T. (2019): Big data in Russian linguistics? *Zeitschrift für Slawistik*, *64* (2), pp. 157–174.

Neuland, E. (1987): Spiegelungen und Gegenspiegelungen. Anregungen für eine zukünftige Jugendsprachenforschung./Reflections and counter-reflections. Suggestions for future youth language research. *Zeitschrift für germanistische Linguistik*, *15*, pp. 58–82.

Neumann, A. (2012): Advantages and disadvantages of different text coding procedures for research and practice in a school context. *Measuring Writing: Recent Insights into Theory, Methodology and Practices*, *27*, pp. 33–54.

Neumann, A. (2016): Zugänge zur Bestimmung von Textqualität. *Forschungshandbuch empirische Schreibdidaktik*, pp. 203–219.

Neumann, A./ Lehmann, R. H. (2008): Schreiben Deutsch. In: DESI-Konsortium (Ed.): *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie*. Weinheim et al.: Beltz, pp. 89–103.

Nguyen, D.-P. (2017): *Text as Social and Cultural Data: a Computational Perspective on Variation in Text*. PhD dissertation, University of Twente.

Nguyen, D./ Doğruöz, A. S./ Rosé, C. P./ de Jong, F. (2016): Computational sociolinguistics: A survey. *Computational Linguistics*, *42* (3), pp. 537–593.

Nguyen, D./ Gravel, R./ Trieschnigg, D./ Meder, T. (2013): "How old do you think I am?": A study of language and age in Twitter. In: *Proceedings of the seventh international AAAI conference on weblogs and social media, 8-11 July 2013, Cambridge, Massachusetts, USA*, pp. 439–448.

Nguyen, D./ Trieschnigg, D./ Doğruöz, A. S./ Gravel, R./ Theune, M./ Meder, T./ de Jong, F. (2014): Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In:

Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014, Dublin, Ireland. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1950–1961.

Nielsen, M. A. (2015): *Neural networks and deep learning*. San Francisco, CA, USA: Determination Press.

Nobata, C./ Tetreault, J./ Thomas, A./ Mehdad, Y./ Chang, Y. (2016): Abusive language detection in online user content. In: *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, pp. 145–153.

Nussbaumer, M. (1996): Lernerorientierte Textanalyse: Eine Hilfe zum Textverfassen. In: Feilke, H./ Portmann-Tselikas, P. (Eds.): *Schreiben im Umbruch. Schreibforschung und schulisches Schreiben*. Stuttgart: Klett, pp. 96–112.

Nussbaumer, M./ Sieber, P. (1994): Texte analysieren mit dem Zürcher Textanalyseraster. In: Sieber, P. (Ed.): *Sprachfähigkeiten–besser als ihr Ruf und nötiger denn je*. Sauerländer, pp. 141–186.

Oakes, M. (2009): Corpus linguistics and stylometry. In: Lüdeling, A./ Kytö, M. (Eds.): *Corpus linguistics: An international handbook* (Vol. 2). Mouton de Gruyter, pp. 1070–1090.

Oakes, M. P. (2005): *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Olinghouse, N. G./ Santangelo, T./ Wilson, J. (2012): Examining the validity of single-occasion, single-genre, holistically scored writing assessments. *Studies in Writing*, *27*, pp. 55–62.

Östling, R./ Smolentzov, A./ Tyrefors Hinnerich, B./ Höglin, E. (2013): Automated essay scoring for swedish. In: *The 8th Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA, USA, June 13, 2013*. Association for Computational Linguistics, pp. 42–47.

Özyirmidokuz, E. K. (2014): Mining Unstructured Turkish Economy News Articles. *Procedia Economics and Finance*, *16*, pp. 320–328.

Packard, G./ Berger, J. (2017): How language shapes word of mouth's impact. *Journal of marketing Research*, *54* (4), pp. 572–588.

Page, L./ Brin, S./ Motwani, R./ Winograd, T. (1999): *The PageRank citation ranking: Bringing order to the web. Technical Report*. Stanford InfoLab.

Pang, B./ Lee, L./ Vaithyanathan, S. (2002): Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. Association for Computational Linguistics, pp. 79–86.

Paquot, M./ Plonsky, L. (2017): Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3* (1), pp. 61–94.

Park, G./ Schwartz, A./ Eichstaedt, J. C./ Kern, M. L./ Kosinski, M./ Stillwell, D. J./ Ungar, L. H./ Seligman, M. (2014): Automatic Personality Assessment through Social Media Language. *Journal of personality and social psychology*, *108* (6), pp. 934–952.

Pastor, G. C. (2016): *Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives - Fraseología computacional y basada en corpus: perspectivas*. Geneva: Tradulex.

Pearl, J. (2009): Causal inference in statistics: An overview. *Statistics surveys*, *3*, pp. 96–146.

Peersman, C./ Daelemans, W./ Van Vaerenbergh, L. (2011): Predicting age and gender in online social networks. In: *Proceedings of the 3rd international workshop on search and mining user-*

generated contents. ACM, pp. 37–44.

Peersman, C./ Daelemans, W./ Vandekerckhove, R./ Vandekerckhove, B./ Van Vaerenbergh, L. (2016): The Effects of Age, Gender and Region on Non-standard Linguistic Variation in Online Social Networks. Retrieved from http://arxiv.org/abs/1601.02431.

Peldszus, A./ Stede, M. (2013): From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *7* (1), pp. 1–31.

Pennacchiotti, M./ Popescu, A.-M. (2011): A Machine Learning Approach to Twitter User Classification. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 281–288.

Pennebaker, J. W./ Francis, M. E./ Booth, R. J. (2001): Linguistic inquiry and word count: LIWC 2001. *Operator's Manual*. Mahway: Lawrence Erlbaum Associates.

Pennebaker, J. W./ Mehl, M. R./ Niederhoffer, K. G. (2002): Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54* (1), pp. 547–577.

Pennington, J./ Socher, R./ Manning, C. (2014): Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perelman, L. (2014): When "the state of the art" is counting words. *Assessing Writing*, *21*, pp. 104–111.

Periñán-pascual, C. (2017): Bridging the gap within text-data analytics: A computer environment for data analysis in linguistic research. *Revista de Lenguas para Fines Específicos*, *23* (2), pp. 111–132.

Persing, I./ Ng, V. (2015): Modeling Argument Strength in Student Essays. In: *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 543–552.

Phakiti, A./ Costa, P. De/ Plonsky, L./ Starfield, S. (2018): Applied Linguistics Research: Current Issues, Methods, and Trends. In: Phakiti, A./ Costa, P. De/ Plonsky, L./ Starfield, S. (Eds.): *The Palgrave Handbook of Applied Linguistics Research Methodology*. London: Palgrave Macmillan, pp. 5–29.

Phillips, L./ Dowling, C./ Shaffer, K./ Hodas, N./ Volkova, S./ Group, A./ Northwest, P. (2017): Using Social Media To Predict the Future: A Systematic Literature Review. Retrieved from https://arxiv.org/abs/1706.06134, pp. 1–55.

Pilán, I./ Vajjala, S./ Volodina, E. (2016): A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, *7* (1), pp. 143–159.

Pilán, I./ Volodina, E./ Zesch, T. (2016): Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2101–2111.

Pitler, E./ Nenkova, A. (2008): Revisiting readability: A unified framework for predicting text quality. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 186–195.

Plaisant, C./ Rose, J./ Yu, B./ Auvil, L./ Kirschenbaum, Matthew GSmith, M. N./ Clement, T./ Lord, G. (2006): Exploring erotics in Emily Dickinson's correspondence with text mining and visual

interfaces. In: *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06)*. Chapel Hill, NC, USA: ACM, pp. 141–150.

Polio, C./ Shea, M. C. (2014): An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, pp. 10–27.

Pölitz, C. (2016a): *Automatic methods to extract latent meanings in large text corpora*. PhD dissertation, Technische Universität Dortmund.

Pölitz, C. (2016b): Data Mining Software for Corpus Linguistics with Application in Diachronic Linguistics. *Journal for Language Technology and Computational Linguistics (JLCL)*, *31* (1), pp. 51–71.

Porhet, C./ Ochs, M./ Saubesty, J./ De Montcheuil, G./ Bertrand, R. (2017): Mining a multimodal corpus of doctor's training for virtual patient's feedbacks. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, pp. 473–478.

Prensky, M. (2001): Digital natives, digital immigrants part 1. *On the horizon*, *9* (5), pp. 1–6.

Prensky, M. (2009): H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: journal of online education*, *5* (3), p. article 1.

Preoţiuc-Pietro, D./ Devlin Marier, R. (2019): Analyzing Linguistic Differences between Owner and Staff Attributed Tweets. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2848–2853.

Present-Thomas, R. L./ Weltens, B./ de Jong, J. H. A. L. (2013): Defining proficiency: A comparative analysis of CEFR level classification methods in a written learner corpus. *Dutch Journal of Applied Linguistics*, *2* (1), pp. 57–76.

Pryzant, R./ Basu, S./ Sone, K. (2018): Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analysing and Interpreting Neural Networks*. Brussels, Belgium: Association for Computational Linguistics, pp. 125–135.

Pryzant, R./ Chung, Y./ Jurafsky, D. (2017): Predicting sales from the language of product descriptions. In: *Special Interest Group on Information Retrieval (SIGIR) eCommerce Workshop*. ACM.

Quinlan, J. R. (2014): *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann.

Rabinovich, E./ Nisioi, S./ Ordan, N./ Wintner, S. (2016): On the Similarities Between Native, Non-native and Translated Texts. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Germany: Association for Computational Linguistics, pp. 1870–1881.

Radinsky, K./ Horvitz, E. (2013): Mining the web to predict future events. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pp. 255–264.

Raghunadha Reddy, T./ Vishnu Vardhan, B./ Vijayapal Reddy, P. (2016): A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, *11* (5), pp. 3092–3102.

Rangel, F./ Rosso, P. (2016): On the impact of emotions on author profiling. *Information processing & management*, *52* (1), pp. 73–92.

Rangel, F./ Rosso, P./ Montes-y-Gómez, M./ Potthast, M./ Stein, B. (2018): Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter. In: *Working Notes Papers of the CLEF*.

Rangel, F./ Rosso, P./ Verhoeven, B./ Daelemans, W./ Potthast, M./ Stein, B. (2016): Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: Balog, K./ Cappellato, L./ Ferro, N./ Craig, M. (Eds.): *CEUR Workshop Proceedings. Working Notes of the CLEF 2016 Evaluation Labs*, pp. 750–784.

Rao, D./ Yarowsky, D./ Shreevats, A./ Gupta, M. (2010): Classifying latent user attributes in twitter. In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, pp. 37–44.

Rayson, P. (2002): *Matrix : A statistical method and software tool for linguistic analysis through corpus comparison*. PhD dissertation, Lancaster University.

Rayson, P. (2008a): Computational Tools and Methods for Corpus Compilation and Analysis. In: Biber, D./ Reppen, R. (Eds.): *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press, pp. 32–49.

Rayson, P. (2008b): From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13* (4), pp. 519–549.

Rayson, P./ Archer, D. (2008): Key domain analysis: Mining text in the humanities and social sciences. In: *Workshop on Text Mining Applications in the Social Sciences in conjunction with the 4th International Conference on e-Social Science*.

Rayson, P./ Mariani, J./ Anderson-Cooper, B./ Baron, A./ Gullick, D./ Moore, A./ Wattam, S. (2017): Towards interactive multidimensional visualisations for corpus linguistics. *Journal for language technology and computational linguistics*, *31* (1), pp. 27–49.

Ribeiro, M. T./ Guestrin, C. (2018): Anchors: High-Precision Model-Agnostic Explanations. In: *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1527–1535.

Ribeiro, M. T./ Singh, S./ Guestrin, C. (2016a): Model-Agnostic Interpretability of Machine Learning. In: *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, pp. 91–95.

Ribeiro, M. T./ Singh, S./ Guestrin, C. (2016b): "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1135–1144.

Riehl, C. M. (2007): Varietätengebrauch und Varietätenkontakt in Südtirol und Ostbelgien. *Linguistik online*, *32* (3), pp. 105–117.

Rinaldi, F./ Schneider, G./ Kaljurand, K./ Hess, M./ Andronis, C./ Konstandi, O./ Persidis, A. (2007): Mining of relations between proteins over biomedical scientific literature using a deep linguistic approach. *Artificial intelligence in medicine*, *39* (2), pp. 127–136.

Rocha, A./ Scheirer, W. J./ Forstall, C. W./ Cavalcante, T./ Theophilo, A./ Shen, B./ Carvalho, A. R. B./ Stamatatos, E. (2017): Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, *12* (1), pp. 5–33.

Rockwell, G./ Berendt, B. (2016): On big data and text mining in the humanities. In: ElAtia, S./ Ipperciel, D./ Zaiane, O. R. (Eds.): *Data Mining and Learning Analytics: Applications in Educational Research*. John Wiley & Sons, pp. 29–40.

Rosenthal, S./ McKeown, K. (2011): Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 763–772.

Rudner, L. M./ Liang, T. (2002): Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, *1* (2).

Rustagi, M./ Prasath, R. R./ Goswami, S./ Sarkar, S. (2009): Learning age and gender of blogger from stylistic variation. In: *International Conference on Pattern Recognition and Machine Intelligence*. Berlin: Springer, pp. 205–212.

Salazar Aguilar, M. A./ Moreno Rodríguez, G. J./ Cabrera-Ríos, M. (2006): Statistical characterization and optimization of artificial neural networks in time series forecasting: The one-period forecast case. *Computación y Sistemas*, *10* (1), pp. 69–81.

Salzberg, S. L. (1997): On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, *1* (3), pp. 317–328.

Sammons, M./ Christodoulopoulos, C./ Kordjamshidi, P./ Khashabi, D./ Srikumar, V./ Vijayakumar, P./ Bokhari, M./ Wu, X./ Roth, D. (2016): EDISON: Feature Extraction for NLP, Simplified. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4085–4092.

Santillán, E. (2009): *Digitale Jugendkommunikation in der Informationsgesellschaft*. PhD dissertation, Universität Wien.

Sap, M./ Park, G./ Eichstaedt, J. C./ Kern, M. L./ Stillwell, D./ Kosinski, M./ Ungar, L. H./ Schwartz, A. (2014): Developing Age and Gender Predictive Lexica over Social Media. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1146–1151.

Sarawagi, S. (2008): Information extraction. *Foundations and Trends in Databases*, *1* (3), pp. 261–377.

Scardamalia, M./ Bereiter, C. (1987): Knowledge telling and knowledge transforming in written composition. *Advances in applied psycholinguistics*, *2*, pp. 142–175.

Schäfer, R./ Barbaresi, A./ Bildhauer, F. (2013): The good, the bad, and the hazy: Design decisions in web corpus construction. In: *Proceedings of the 8th Web as Corpus Workshop*. Lancaster, UK: Association for Computational Linguistics, pp. 7–15.

Schler, J./ Koppel, M./ Argamon, S./ Pennebaker, J. W. (2006): Effects of Age and Gender on Blogging. In: *AAAI spring symposium: Computational approaches to analyzing weblogs* (Vol. 6), pp. 199–205.

Schober, O. (2007): Zur Sprachsituation in Südtirol. In: Heller, H. (Ed.): *Fremdheit im Prozess der Globalisierung*. Wien: LIT, pp. 133–149.

Scholbeck, C. A./ Molnar, C./ Heumann, C./ Bischl, B./ Casalicchio, G. (2019): Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model Agnostic Interpretations. Retrieved from http://arxiv.org/abs/1904.03959.

Schoonen, R. (2012): The validity and generalizability of writing scores: The effect of rater, task and language. In: Van Steendam, E./ Tillema, M./ Rijlaarsdam, G./ Van Den Bergh, H. (Eds.): *Measuring Writing: Recent Insights into Theory, Methodology and Practices*. Leiden, Boston, Netherlands: Brill, pp. 1–22.

Schreiber, B. R. (2015): "I am what I am": Multilingual identity and digital translanguaging. *Language Learning and Technology*, *19* (3), pp. 69–87.

Schriver, K. A. (1989): Evaluating Text Quality: The Continuum from Text-focused to Reader-focused Methods. *IEEE Transactions on Professional Communication*, *32* (4), pp. 238–255.

Schwartz, A./ Eichstaedt, J. C./ Kern, M. L./ Dziurzynski, L./ Ramones, S./ Agrawal, M./ Shah, A./ Kosinski, M./ Stillwell, D./ Seligman, M./ Ungar, L. H. (2013): Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8* (9), p. article e73791.

Sculley, D./ Holt, G./ Golovin, D./ Davydov, E./ Phillips, T./ Ebner, D./ Chaudhary, V./ Young, M./ Crespo, J.-F./ Dennison, D. (2015): Hidden technical debt in machine learning systems. In: Cortes, C./ Lawrence, N. D./ Lee, D. D./ Sugiyama, M./ Garnett, R. (Eds.): *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Curran Associates, pp. 2503–2511.

Shermis, M. D. (2014): State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, pp. 53–76.

Shermis, M. D./ Burstein, J. C. (2003): *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates.

Shmueli, G. (2010): To explain or to predict? *Statistical science*, *25* (3), pp. 289–310.

Siebenhaar, B. (2006): Gibt es eine jugendspezifische Varietätenwahl in Schweizer Chaträumen? In: Dürscheid, C./ Spitzmüller, J. (Eds.): *Perspektiven der Jugendsprachforschung. Trends and Developments in Youth Language Research*. Bern, Frankfurt a. M.: Peter Lang, pp. 227–239.

Siebenhaar, B. (2008): Quantitative Approaches to Linguistic Variation in IRC: Implications for Qualitative Research. *Language@Internet*, *5* (4), pp. 1–14.

Siever, T./ Schoblinski, P./ Runkehl, J. (Eds.) (2005): *Websprache.net: Sprache und Kommunikation im Internet*. Berlin/New York: De Gruyter.

Siirtola, H./ Säily, T./ Nevalainen, T./ Räihä, K.-J. (2014): Text Variation Explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics*, *19* (3), pp. 417–429.

Simaki, V./ Aravantinou, C./ Mporas, I./ Kondyli, M./ Megalooikonomou, V. (2017): Sociolinguistic Features for Author Gender Identification: From Qualitative Evidence to Quantitative Analysis. *Journal of Quantitative Linguistics*, *24* (1), pp. 65–84.

Simaki, V./ Mporas, I./ Megalooikonomou, V. (2016a): Age Identification of Twitter Users: Classification Methods and Sociolinguistic Analysis. In: *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pp. 385–395.

Simaki, V./ Mporas, I./ Megalooikonomou, V. (2016b): Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, *9* (4), pp. 868–876.

Simaki, V./ Simakis, P./ Paradis, C./ Kerren, A. (2017): Identifying the Authors' National Variety of English in Social Media Texts. In: *Proceedings of the 11th Biennial Conference on Recent Advances In Natural Language Processing (RANLP'17), 2-8 September 2017, Varna, Bulgaria*. Varna, Bulgaria, pp. 671–678.

Singh, J. P./ Dwivedi, Y. K./ Rana, N. P./ Kumar, A./ Kapoor, K. K. (2017): Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, pp. 1–21.

Singh, J. P./ Irani, S./ Rana, N. P./ Dwivedi, Y. K./ Saumya, S./ Kumar Roy, P. (2017): Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*, pp. 346–355.

Smith, G. D./ Ebrahim, S. (2002): Data dredging, bias, or confounding. *BMJ (Clinical research)*, *325* (7378), pp. 1437–1438.

Sokolova, M./ Lapalme, G. (2009): A systematic analysis of performance measures for classification

tasks. *Information Processing & Management*, *45* (4), pp. 427–437.

Speelman, D./ Heylen, K./ Geeraerts, D. (Eds.) (2018): *Mixed-effects regression models in linguistics*. Cham: Springer.

Spina, S. (2019a): Emoticons as Multifunctional and Pragmatic Resources: a Corpus-based Study on Twitter. In: Stemle, E./ Wigham, C. R. (Eds.): *Building computer-mediated communication corpora for sociolinguistic analysis*. Clermont-Ferrand: Presses universitaires Blaise Pascal, pp. 123–146.

Spina, S. (2019b): Role of Emoticons as Structural Markers in Twitter Interactions. *Discourse Processes*, *56* (4), pp. 345–362.

Stæhr, A. (2015): Reflexivity in Facebook interaction: Enregisterment across written and spoken language practices. *Discourse, Context and Media*, *8*, pp. 30–45.

Stamatatos, E. (2009): A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60* (3), pp. 538–556.

Stamatatos, E. (2016): Universality of Stylistic Traits in Texts. In: *Creativity and Universality in Language*. Springer, pp. 143–155.

Stamatatos, E./ Fakotakis, N./ Kokkinakis, G. (2000): Automatic text categorization in terms of genre and author. *Computational linguistics*, *26* (4), pp. 471–495.

Staniak, M./ Biecek, P. (2018): Explanations of model predictions with live and breakDown packages. *The R Journal*, *10* (2), pp. 395–409.

Stemle, E./ Onysko, A. (2015): Automated L1 identification in English learner essays and its implications for language transfer. In: Peukert, H. (Ed.): *Transfer Effects in Multilingual Language Development* (Vol. 4). Amsterdam/Philadelphia: John Benjamins, pp. 297–321.

Steyvers, M./ Griffiths, T. (2007): Probabilistic topic models. In: Landauer, T. K./ McNamara, D. S./ Dennis, S./ Kintsch, W. (Eds.): *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 424–440.

Stoop, W./ van den Bosch, A. P. J. (2014): Using idiolects and sociolects to improve word prediction. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 318–327.

Storrer, A. (2000): Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet. In: Voß, G. G./ Holly, W./ Boehnke, K. (Eds.): *Neue Medien im Alltag*. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 151–175.

Storrer, A. (2013): Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze–empirische Befunde. In: Plewnia, A./ Witt, A. (Eds.): *Sprachverfall? Dynamik–Wandel–Variation. Jahrbuch des Instituts für Deutsche Sprache*. Berlin/Boston: De Gruyter, pp. 171–196.

Štrumbelj, E./ Kononenko, I. (2014): Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41* (3), pp. 647–665.

Stulpe, A./ Lemke, M. (2016): Blended Reading: Theoretische und praktische Dimension der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung. In: Lemke, M./ Wiedemann, G. (Eds.): *Text Mining in den Sozialwissenschaften*. Wiesbaden: Springer VS, pp. 17–61.

Sun, T./ Gaut, A./ Tang, S./ Huang, Y./ ElSherief, M./ Zhao, J./ Mirza, D./ Belding, E./ Chang, K.-W./ Wang, W. Y. (2019): Mitigating Gender Bias in Natural Language Processing: Literature Review.

In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1630–1640.

Sushil, M./ Šuster, S./ Daelemans, W. (2018): Rule induction for global explanation of trained models. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 82–97.

Szatlóczki, G./ Pákáski, M./ Kálmán, J./ Hoffmann, I./ Tóth, L./ Bánréti, Z./ Gosztolya, G./ Vincze, V. (2016): Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 181–187.

Szmrecsanyi, B. (2019): Register in variationist linguistics. *Register Studies*, *1* (1), pp. 76–99.

Szmrecsanyi, B./ Grafmiller, J./ Bresnan, J./ Rosenbach, A./ Tagliamonte, S./ Todd, S. (2017): Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics*, *2* (1), p. article 86.

Taboada, M./ Brooke, J./ Tofiloski, M./ Voll, K./ Stede, M. (2011): Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37* (2), pp. 267–307.

Tagg, C./ Seargeant, P. (2014): Audience design and language choice in the construction and maintenance of translocal communities on social network sites. In: Seargeant, P./ Tagg, C. (Eds.): *The Language of Social Media*. London: Palgrave Macmillan, pp. 161–185.

Taghipour, K./ Ng, H. T. (2016): A neural approach to automated essay scoring. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891.

Tagliamonte, S. A./ Denis, D. (2008): Linguistic ruin? LOL! Instant messaging and teen language. *American speech*, *83* (1), pp. 3–34.

Taguchi, N./ Crawford, W./ Wetzel, D. Z. (2013): What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *Tesol Quarterly*, *47* (2), pp. 420–430.

Tan, C./ Lee, L./ Pang, B. (2014): The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 175–185.

Teich, E./ Degaetano-Ortlieb, S./ Fankhauser, P./ Kermes, H./ Lapshinova-Koltunski, E. (2015): The linguistic construal of disciplinarity: A data-mining approach using register features. *Journal of the Association for Information Science and Technology*, *67* (7), pp. 1668–1678.

Teich, E./ Degaetano-Ortlieb, S./ Kermes, H./ Lapshinova-Koltunski, E. (2013): Scientific registers and disciplinary diversification: A comparable corpus approach. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 59–68.

Teich, E./ Frankhauser, P. (2009): Exploring a corpus of scientific texts using data mining. *Language and Computers*, *71* (1), pp. 233–247.

Thelwall, M./ Wilkinson, D./ Uppal, S. (2009): Data Mining Emotion in Social Network Communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, *61* (1), pp. 190–199.

Therneau, T./ Atkinson, B./ Ripley, B./ Ripley, M. B. (2015): Package 'rpart.' R Foundation for Statistical

Computing.

Thiagarajan, J. J./ Kailkhura, B./ Sattigeri, P./ Ramamurthy, K. N. (2016): TreeView: Peeking into deep neural networks via feature-space partitioning. In: *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain: Curran Associates.

Thimm, C. (2002): Generationsspezifische Wortschätze. In: Cruse, D. A./ Hundsnurscher, F./ Job, M./ Lutzeier, P. R. (Eds.): *Lexikologie. Handbücher zur Sprach- und Kommunikationswissenschaft 21*. Berlin: De Gruyter, pp. 880–888.

Thornham, H./ McFarlane, A. (2011): Discourses of the Digital Native. *Information, Communication & Society*, *14* (2), pp. 258–279.

Tognini-Bonelli, E. (2001): *Corpus linguistics at work*. Amsterdam: John Benjamins.

Trudgill, P./ Trudgill, S. (1974): *The social differentiation of English in Norwich* (Vol. 13). Cambridge, UK: Cambridge University Press.

Tumasjan, A./ Sprenger, T. O./ Sandner, P. G./ Welpe, I. M. (2010): Predicting elections with twitter: What 140 characters reveal about political sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, pp. 178–185.

Turney, P. D. (2002): Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 417–424.

Vajjala, S. (2018): Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, *28* (1), pp. 79–105.

Vajjala, S./ Meurers, D./ Eitel, A./ Scheiter, K. (2016): Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, December 11-17 2016*. Osaka, Japan, pp. 38–48.

van Cranenburgh, A./ Bod, R. (2017): A Data-oriented Model of Literary Language. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1228–1238.

Van Den Bergh, H./ De Maeyer, S./ Van Weijen, D./ Tillema, M. (2012): Generalizability of text quality scores. In: Van Steendam, E./ Tillema, M./ Rijlaarsdam, G./ Van Den Bergh, H. (Eds.): *Measuring Writing: Recent Insights into Theory, Methodology and Practices*. Leiden/Boston: Brill, pp. 23–32.

Van den Bosch, A. (2009): Machine learning. In: Lüdeling, A./ Kytö, M. (Eds.): *Corpus Linguistics. An International Handbook* (Vol. 2). Berlin/ New York: Mouton de Gruyter, pp. 855–873.

van der Waa, J./ Robeer, M./ van Diggelen, J./ Brinkhuis, M./ Neerincx, M. (2018): Contrastive Explanations with Local Foil Trees. In: *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. Stockholm, Sweden, pp. 41–47.

Van Gompel, M./ Van Den Bosch, A. (2016): Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software*, *4* (1).

Van Halteren, H./ Baayen, H./ Tweedie, F./ Haverkort, M./ Neijt, A. (2005): New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, *12* (1), pp. 65–77.

Vardigan, M./ Heus, P./ Thomas, W. (2008): Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, *3* (1), pp. 107–113.

Varghese, A./ Varde, A./ Peng, J./ Fitzpatrick, E. (2013): The CollOrder System for Detecting and Correcting Odd Collocations in L2 Written English. In: *Demo paper in IEEE International Conference on Information and Communication Systems (ICICS)*, pp. 155–160.

Venturi, G./ Bellandi, T./ Dell'Orletta, F./ Montemagni, S. (2015): NLP-Based Readability Assessment of Health-Related Texts: A Case Study on Italian Informed Consent Forms. In: *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pp. 131–141.

Verheijen, L. (2017): WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication. In: Fišer, D./ Beißwenger, M. (Eds.): *Investigating Computer-mediated Communication. Corpus-based Approaches to Language in the Digital World*. Ljubljana, Slovenia: Ljubljana University Press, pp. 72–101.

Vikoler, M. (2016): *Der Einfluss des Italienischen auf die deutsche Jugendsprache in Südtirol*. Bachelor's thesis, University of Bologna.

Volansky, V./ Ordan, N./ Wintner, S. (2015): On the features of translationese. *Digital Scholarship in the Humanities*, *30* (1), pp. 98–118.

Wachsmuth, H./ Naderi, N./ Habernal, I./ Hou, Y./ Hirst, G./ Gurevych, I./ Stein, B. (2017): Argumentation quality assessment: Theory vs. practice. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2), pp. 250–255.

Wachter, S./ Mittelstadt, B./ Russell, C. (2018): Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31* (2), pp. 841–887.

Wang, B./ Zubiaga, A./ Liakata, M./ Procter, R. (2015): Making the most of tweet-inherent features for social spam detection on twitter. In: *Proceedings of the 5th Workshop on Making Sense of Microposts, Florence Italy, 18 May 2015*, pp. 10–16.

Wang, C./ Blei, D. M. (2011): Collaborative topic modeling for recommending scientific articles. In: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 448–456.

Wei, Z./ Liu, Y./ Li, Y. (2016): Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, pp. 195–200.

Weir, G. R. S./ Dos Santos, E./ Cartwright, B./ Frank, R. (2016): Positing the problem: enhancing classification of extremist web content through textual analysis. In: *Proceedings of the IEEE International Conference on Cybercrime and Computer Forensics (ICCCF)*. IEEE.

Weiß, Z. (2015): *More Linguistically Motivated Features of Language Complexity in Readability Classification of German Textbooks: Implementation and Evaluation*. Bachelor's thesis, Universität Tübingen.

Weiß, Z. (2017): *Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects*. Master's thesis, Universität Tübingen.

Weiss, Z./ Dittrich, S./ Meurers, D. (2018): A Linguistically-informed Search Engine to Identifiy Reading Material for Functional Illiteracy Classes. In: *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pp. 79–90.

Weiss, Z./ Meurers, D. (2019): Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 380–393.

Weiß, Z./ Meurers, D. (2018): Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 303–317.

Weiß, Z./ Riemenschneider, A./ Schröter, P./ Meurers, D. (2019): Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 30–45.

Weld, D. S./ Bansal, G. (2018): The Challenge of Crafting Intelligible Intelligence. *Communication of the ACM*, *62* (6), pp. 70–79.

Wiedemann, G. (2013): Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, *38* (4), pp. 332–357.

Wiegand, M./ Bocionek, C./ Ruppenhofer, J. (2016): Opinion Holder and Target Extraction on Opinion Compounds–A Linguistic Approach. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 800–810.

Wilkinson, M. D./ Dumontier, M./ Aalbersberg, Ij. J./ Appleton, G./ Axton, M./ Baak, A./ Blomberg, N./ Boiten, J.-W./ da Silva Santos, L. B./ Bourne, P. E. (2016): The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*.

Wilmsmeier, S./ Brinkhaus, M./ Hennecke, V. (2016): Ratingverfahren zur Messung von Schreibkompetenz in Schülertexten. *Bulletin VALS-ASLA*, *103*, pp. 101–117.

Wilson, T./ Wiebe, J./ Hoffmann, P. (2005): Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, Canada: Association for Computational Linguistics, pp. 347–354.

Witten, I. H./ Frank, E./ Hall, M./ Pal, C. J. (2016): *Data Mining: Practical machine learning tools and techniques*. Amsterdam et al.: Morgan Kaufmann.

Wolfe-Quintero, K./ Inagaki, S./ Kim, H.-Y. (1998): *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Second language teaching and curriculum center, University Hawaii Press.

Wolpert, D. H./ Macready, W. G. (1995): *No free lunch theorems for search*. Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wong, S.-M. J./ Dras, M. (2009): Contrastive analysis and native language identification. In: *Proceedings of the Australasian Language Technology Association Workshop 2009*. Sydney, Australia, pp. 53–61.

Wu, T./ Wen, S./ Xiang, Y./ Zhou, W. (2018): Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security*, *76*, pp. 265–284.

Yang, H. (2009): *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches*. PhD dissertation, University of Texas Austin.

Yannakoudakis, H./ Briscoe, T./ Medlock, B. (2011): A new dataset and method for automatically grading ESOL texts. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 180–189.

Yu, B. (2008): An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, *23* (3), pp. 327–343.

Zeroual, I./ Lakhouaja, A. (2018): Data science in light of natural language processing: An overview. *Procedia Computer Science*, *127*, pp. 82–91.

Zesch, T./ Wojatzki, M./ Scholten-Akoun, D. (2015): Task-Independent Features for Automated Essay Grading. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 224–232.

Zhang, W./ Lau, R. Y. K./ Li, C. (2014): Adaptive Big Data Analytics for Deceptive Review Detection in Online Social Media. In: *35th International Conference on Information Systems "Building a Better World Through Information Systems" (ICIS 2014)*. Aukland, New Zealand: Association for Information Systems, pp. 1–19.

Zheng, X./ Zhu, S./ Lin, Z. (2013): Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach. *Decision Support Systems*, *56* (1), pp. 211–222.

Zien, A./ Krämer, N./ Sonnenburg, S./ Rätsch, G. (2009): The feature importance ranking measure. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 694–709.

Zifonun, G./ Hoffmann, L./ Strecker, B. (1997): *Grammatik der deutschen Sprache* (Vol. 1). Berlin / New York: Walter de Gruyter.

Zupanc, K./ Bosnic, Z. (2015): Advances in the field of automated essay evaluation. *Informatica*, *39* (4), pp. 383–395.

Zuur, A. F./ Ieno, E. N./ Elphick, C. S. (2010): A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, *1* (1), pp. 3–14.

30 List of figures

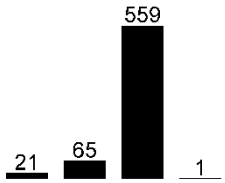
31 List of tables

32 List of model outputs

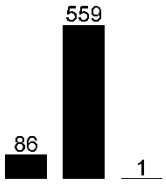
Appendix

A: Questionnaire items used in corpus study 1

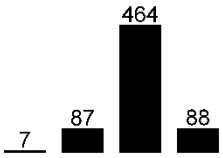
(A) formal completeness



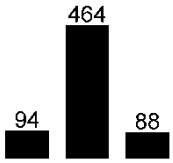
Q03_Text_beginnt_mit
The text begins with...
- the main part.
- an initial part, but with inadequate or no introduction.
- an introduction.
- Not determinable (NA)



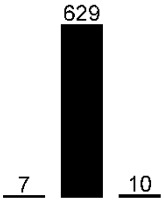
Q03_Text_beginnt_mit.rec
The text begins with...
- *no real introduction.*
- *an introduction.*
- *not determinable (NA)*



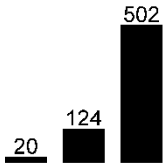
Q04_Thema_in_Einleitung
Is the topic of the text mentioned in the introduction (e.g. the topic of youth (in itself); the question whether youth is enviable; Hans Magnus Enzensberger and the interview, ...) clear?
- No.
- Yes, but it does not become clear what the announced topic is.
- Yes, it becomes clear what the announced topic is.
- Not determinable (NA)



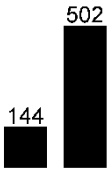
Q04_Thema_in_Einleitung.rec
Is the topic of the text mentioned in the introduction (e.g. the topic of youth (in itself); the question whether youth is enviable; Hans Magnus Enzensberger and the interview, ...) clear?
- *No, the topic is not clear.*
- *Yes, the topic is clear.*
- *Not determinable (NA)*



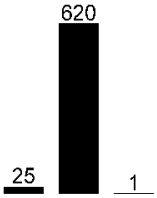
Q05_im_Hauptteil
In the main part of the text...
- the thematic core(s) is/are not elaborated.
- the thematic core is/the thematic cores are elaborated.
- not determinable (NA)



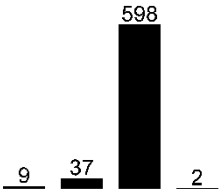
Q06_Text_endet
The text ends...
- abrupt in/after the main part.
- with a final part, but without a proper ending.
- with a proper ending.
- not determinable (NA)



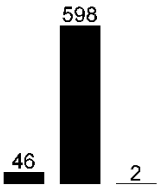
Q06_Text_endet.rec
The text ends...
- abrupt or without proper ending.
- with a proper ending.



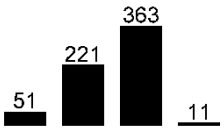
Q07_weitere_Textteile
Does the text contain further parts of the text (e.g. excursuses)?
- Yes
- No
- Not determinable (NA)



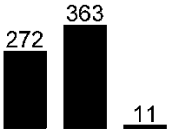
Q08_pers_Stellungnahme
Is there a personal statement in the text (relating to the whole text)?
- No, not present.
- Yes, as a positioning to one's own theme (if it deviates from the one in the input text).
- Yes, as positioning to statements of the input text.
- Not determinable (NA)



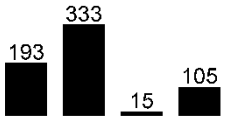
Q08_pers_Stellungnahme.rec
Is there a personal statement in the text (relating to the whole text)?
- A personal statement is not present or does not refer to the input text.
- Yes, as positioning to statements of the input text.
- Not determinable (NA)



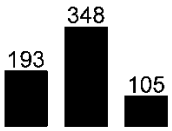
Q09_pers_Meinung_deutlich
Is the opinion of the writer expressed in the personal statement clear?
- No, the writer's opinion is not clear.
- Yes, the writer's opinion is partly, but not completely clear.
- Yes, the writer's opinion is clear.



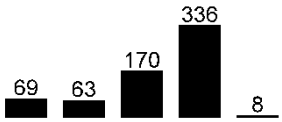
Q09_pers_Meinung_deutlich.rec
Is the opinion of the writer expressed in the personal statement clear?
- No, the writer's opinion is not clear.
- Yes, the writer's opinion is partly, but not completely clear.
- Yes, the writer's opinion is clear.



Q11_Stellungnahme_vs_Enzensberger
Which of the following possibilities is applicable the opinion statement?
- The statement contains both positions.
- The statement is (mostly) rejecting the statements of Hans Magnus Enzensberger.
- The statement is (mostly) in agreement with Hans Magnus Enzensberger's statements.
- Not determinable (NA)

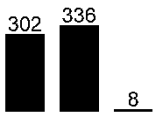


Q11_Stellungnahme_vs_Enzensberger.rec
Which of the following possibilities is applicable the opinion statement?
- The statement contains both positions.
- The statement is exclusively positive or negative
- Not determinable (NA)



Q12_Fazit_ableitbar

- No conclusion exists.
- A conclusion exists, but the conclusion cannot be derived from what has been written.
- A conclusion exists. The conclusion can be partially derived from what has been written.
- A conclusion exists. The conclusion refers to what has been written or can be derived from what has been written.
- Not determinable (NA)



Q12_Fazit_ableitbar.rec

Is there a conclusion in the text? If so, can the conclusion be derived from what has been written?

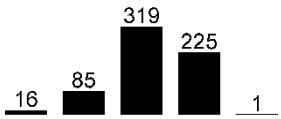
- *There is no conclusion that refers to the rest of the text.*
- *A conclusion exists and refers to what has been before.*
- *Not determinable (NA)*

(B) content

Q13_expliziter_Bezug_Input

Is there an explicit reference to the input text?

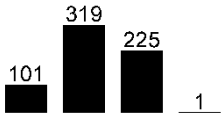
- No
- Yes



Q18_Thema_Gesamttext_vorhanden

Can you assign one topic for the whole text?

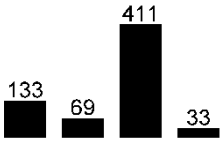
- No
- Rather no
- Rather yes
- Yes
- Not determinable (NA)



Q18_Thema_Gesamttext_vorhanden.rec

Can you assign one topic for the whole text?

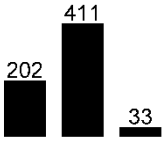
- *No/Rather no*
- *Rather yes*
- *Yes*
- *Not determinable (NA)*



Q21_Ankuendigung_Erfuellung_Thema

Is the topic announced in the first part of the text? If yes: does the text comply with the announced topic?

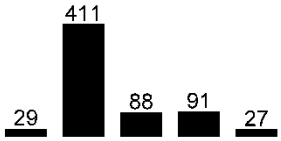
- No topic is announced in the introduction.
- A topic is announced but the author does not stick to it.
- A topic is announced; the text complies to the announced topic.
- Not determinable (NA)



Q21_Ankuendigung_Erfuellung_Thema.rec

Is the topic announced in the first part of the text? If yes: does the text comply with the announced topic?

- *No topic is announced in the introduction that the text complies with.*
- *A topic is announced; the text complies to the announced topic.*
- *Not determinable (NA)*

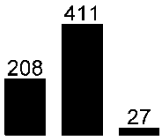


Q22_dominante_Themenentfaltung

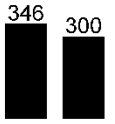
Which basic form of topic development is dominant in this text?

- Other forms of topic development not mentioned here
- Argumentative topic development
- Descriptive topic development

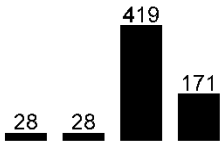
- Explicative topic development
- Not determinable (NA)



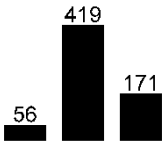
Q22_dominante_Themenentfaltung.rec
Which basic form of topic development is dominant in this text?
- Other forms of topic development
- Argumentative topic development
- Not determinable (NA)



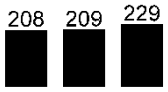
Q23_weitere_Form_Themenentfaltung_vorhanden
In addition to the form of topic development already mentioned, is there another form of topic development in this text or is it not clear which form of topic development is dominant?
- There is another form of topic development available/it is not clear which form is dominant.
- There is no other form of topic development.



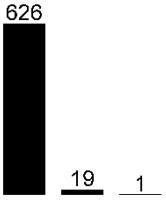
Q29_Argumentationsstrategien
Which argumentation strategy is primarily pursued in this text?
- Only counterarguments are mentioned.
- Only pro-arguments are mentioned
- Pro- and counterarguments are mentioned.
- Not determinable (NA)



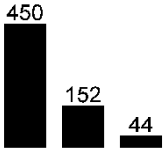
Q29_Argumentationsstrategien.rec
Which argumentation strategy is primarily pursued in this text?
- Only pro- or only counterarguments are mentioned.
- Pro- and counterarguments are mentioned.
- Not determinable (NA)



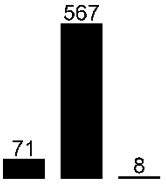
Q30_konzessive_Argumentation
Is the argumentation concessive?
- No
- Yes
- Not determinable (NA)



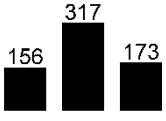
Q31_explizite_Adressatenorientierung
Is there an explicit addressing of a fictitious or actual recipient?
- No
- Yes
- Not determinable (NA)



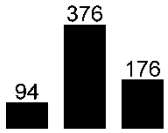
Q32_Darstellung_Sachverhalte_subjektiv
Presentation of the facts: Which characteristic applies to the text?
- The aspects mentioned by the writer are mainly linked to concrete or personal experiences.
- Information is considered which is linked to abstract or not directly accessible facts and has little to do with the writer's immediate environment.
- Not determinable



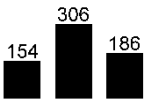
Q33_Emotionalität
Is the text emotional?
- No/rather not.
- Yes/mostly yes.
- Not determinable (NA)



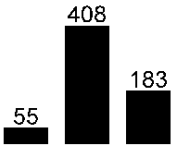
Q34_objektive_Argumentation
Subjectivity or objectivity of argumentation: Which characteristic applies to the text?
- The argumentation is (rather) subjective.
- The argumentation is (rather) objective.
- Not determinable (NA)



Q35_abwiegen_und_enschraenken
Argumentation strategy: Which characteristic applies to the text?
- The author assesses and adjusts.
- The author does not assess and adjust.
- Not determinable (NA)

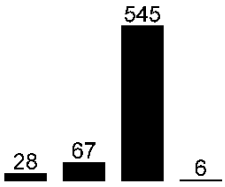


Q36_Argumentationsgang_deutlich_logisch_nachvollziehbar
Course of argumentation: Which characteristic applies to the text?
- There is no clearly recognizable and logically comprehensible course of argumentation that leads to a certain result or conclusion.
- There is a clearly recognizable and logically comprehensible line of argumentation that leads to a certain result or conclusion.
- Not determinable (NA)

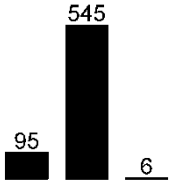


Q37_Argumente_begruendet_Position_bezogen
Justifications of arguments: Which characteristic applies to the text?
- There is no clear reasoning/ no clear position is taken.
- There is a clear reasoning and a clear position is taken.
- Not determinable (NA)

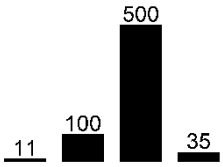
(C) formal and linguistic means of text arrangement



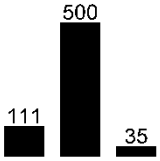
Q38_Umbrueche_Textteile_getrennt
Are there line breaks between introduction, main part and ending?
- No
- Not all
- Yes
- Not determinable (NA)



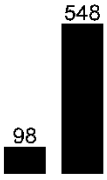
Q38_Umbrueche_Textteile_getrennt.rec
Are there line breaks between introduction, main part and ending?
- *No*
- *Yes*
- *Not determinable (NA)*



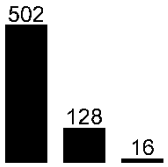
Q39_Umbrueche_gelungen
Are the line breaks between introduction, main part and ending appropriate?
- No
- Not all
- Yes
- Not determinable (NA)



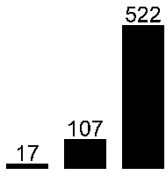
Q39_Umbrueche_gelungen.rec
Are the line breaks between introduction, main part and ending appropriate?
- *No*
- *Yes*
- *Not determinable (NA)*



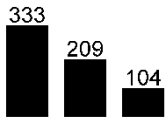
Q40_weitere_Umbrueche
Are there other line breaks despite the ones separating introduction, main part and ending?
- No
- Yes



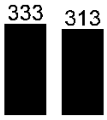
Q41_Anfang_Textfunktion_angekuendigt
Are the text functions announced in the introduction?
- No
- Yes
- Not determinable (NA)



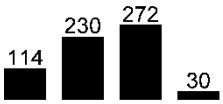
Q42_Angekuend_Textfunktion_eingehalten
Does the text (mostly) adhere to the announced text functions?
- No
- Yes
- Not determinable (NA)



Q43_Textgliederung_inhaltlich_sprachlich
Is the text structured through textual means?
- No, no or almost no textual structuring is present.
- Yes, textual structuring is partly present.
- Yes, textual structuring is mostly present.

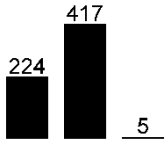


Q43_Textgliederung_inhaltlich_sprachlich.rec
Is the text structured through textual means?
- *No.*
- *Yes, at least partly.*

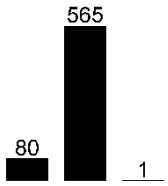


Q44_Textgliederung_inhaltlich_graphisch_unterstuetzt
Does the formal and textual structure support the reception of the text?
- (Mainly) no.
- Partly.
- (Mainly) yes.
- Not determinable (NA)

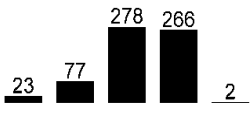
(D) overall impression



Q45_Aufgabenstellung_erfuellt
Does the text fulfil the task (the text is an argumentative essay; in the text there is a discussion of the input text; in the text is a personal statement made)?
- No
- Yes
- Not determinable (NA)

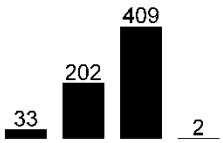


Q47_Qualitaet_konsistent
Is the text quality fluctuating or constant?
- Fluctuating
- Constant
- Not determinable (NA)

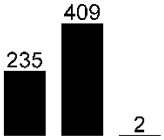


Q48_koherent
Does the text appear coherent overall?
- No.
- Rather no.
- Rather yes.
- Yes.
- Not determinable (NA)

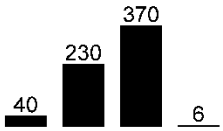
Q48_koherent.rec
Does the text appear coherent overall?
- *No / rather no.*
- *Yes / rather yes.*
- *Not determinable (NA)*



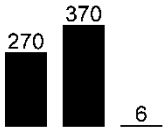
Q49_Gesamtthema_nachvollziehbar
Comprehensibility of the overall theme: Which characteristic applies to the text?
- The overall theme of the text is not comprehensible.
- The overall theme of the text is partially comprehensible.
- The overall theme of the text is comprehensible.
- Not determinable (NA)



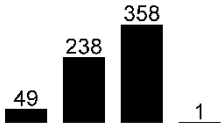
Q49_Gesamtthema_nachvollziehbar.rec
Comprehensibility of the overall theme: Which characteristic applies to the text?
- *The overall theme of the text is not entirely comprehensible.*
- *The overall theme of the text is comprehensible.*
- *Not determinable (NA)*



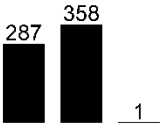
Q50_konzeptionell_zusammenh
Conception of the text: Which characteristic applies to the text?
- The text is conceptually incoherent.
- The text is partially conceptually coherent.
- The text is conceptually coherent.
- Not determinable (NA)



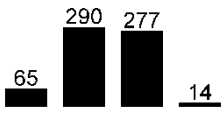
Q50_konzeptionell_zusammenh.rec
Conception of the text: Which characteristic applies to the text?
- *The text is conceptually incoherent.*
- *The text is conceptually coherent.*
- *Not determinable (NA)*



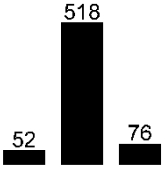
Q51_inhaltlich_klarer_Aufbau
Content structure of the text: Which characteristic applies to the text?
- The text does not have a clear structure.
- The text has a partially clear structure.
- The text has a clear structure.
- Not determinable (NA)



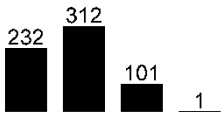
Q51_inhaltlich_klarer_Aufbau.rec
Content structure of the text: Which characteristic applies to the text?
- *The text does not have a clear structure.*
- *The text has a clear structure.*
- *Not determinable (NA)*



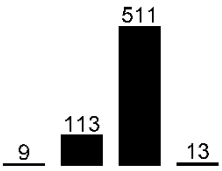
Q52_Absatzstruktur_nachvollz
Paragraph structure: Which characteristic applies to the text?
- The text does not have a comprehensible paragraph structure.
- The text has a partially comprehensible paragraph structure.
- The text has a comprehensible paragraph structure.
- Not determinable (NA)



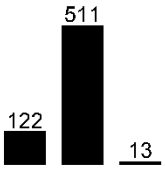
Q53_Fazit_zu_Input
Conclusion: Which characteristic applies to the text?
- The conclusion does not refer to the text.
- The conclusion refers to the text.
- Not determinable (NA)



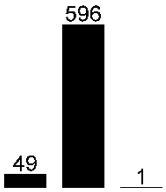
Q54_Anz_inhaltlich_sprachlich_Gliederungsmerk
Textual structuring signals: Which characteristic applies to the text?
- In the text no textual structuring signals are used.
- Some textual structuring signals are used in the text.
- Many textual structuring signals are used in the text.
- Not determinable (NA)



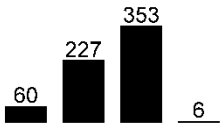
Q55_eindeutige_Bezuege
References: Which characteristic applies to the text?
- The references in the text are not clear.
- The references in the text are partially clear.
- The references in the text are clear.
- Not determinable (NA)



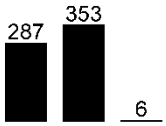
Q55_eindeutige_Bezuege.rec
References: Which characteristic applies to the text?
- *The references in the text are not or rather not clear.*
- *The references in the text are clear.*
- *Not determinable (NA)*



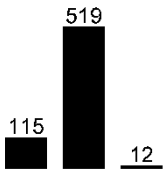
Q56_Konnektorenprobleme
Use of connectives: Which characteristic applies to the text?
- The use of connectives causes some problems.
- The use of con connectives does not cause any problems.
- Not determinable (NA)



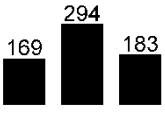
Q57_inhaltliche_Spruenge
Jumps in content: Which characteristic applies to the text?
- There are many content jumps in the text.
- There are some content jumps in the text.
- There are no content jumps in the text.
- Not determinable (NA)



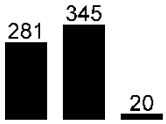
Q57_inhaltliche_Spruenge.rec
Jumps in content: Which characteristic applies to the text?
- *There are content jumps in the text.*
- *There are no content jumps in the text.*
- *Not determinable (NA)*



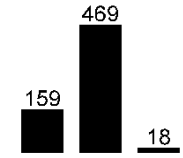
Q58_Roter_Faden
Common theme: Which characteristic applies to the text?
- There is no common theme in the text.
- There is a common theme in the text.
- Not determinable (NA)



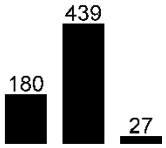
Q60_ueberzeugende Argumentation
The following aspects apply to the text:
- The argumentation is not convincing.
- The argumentation is convincing.
- Not determinable (NA)



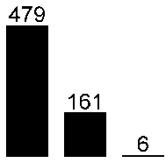
Q61_hat_gefallen
The following aspects apply to the text:
- The text was appealing.
- The text was not appealing.
- Not determinable (NA)



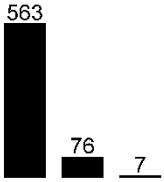
Q62_inhaltlich_nachvollziehbar_klar
The following aspects apply to the text:
- The text is not comprehensible/clear, but rather confusing.
- The text is comprehensible/clear in terms of content.
- Not determinable (NA)



Q63_interessant_langweilig
The following aspects apply to the text: The text...
- is boring.
- is interesting.
- not determinable (NA)



Q64_unterhaltsam
The following aspects apply to the text: The text...
- is not entertaining.
- is entertaining.
- not determinable (NA)



Q65_humorvoll_ironisch
The following aspects apply to the text: The text...
- is not humorous, ironic.
- is humorous, ironic.
- not determinable (NA)

B: Linguistic complexity measures used in corpus study 1

F11	PPsPerSentence	Syntax	Form	Phrase complexity
F12	PPsPerTUnit	Syntax	Form	Phrase complexity
F13	PPsPerClause	Syntax	Form	Phrase complexity
F14	PPsPerFiniteClause	Syntax	Form	Phrase complexity
F15	VPsPerSentence	Syntax	Form	Phrase complexity
F16	VPsPerTUnit	Syntax	Form	Phrase complexity
F17	VPsPerClause	Syntax	Form	Phrase complexity
F18	VPsPerFiniteClause	Syntax	Form	Phrase complexity
F19	VZsPerSentence	Syntax	Form	Phrase complexity
F20	VZsPerTUnit	Syntax	Form	Phrase complexity
F21	VZsPerClause	Syntax	Form	Phrase complexity
F22	VZsPerFiniteClause	Syntax	Form	Phrase complexity
F23	NPsPerSentence	Syntax	Form	Phrase complexity
F24	NPsPerTUnit	Syntax	Form	Phrase complexity
F25	NPsPerClause	Syntax	Form	Phrase complexity
F26	NPsPerFiniteClause	Syntax	Form	Phrase complexity
F27	complexNominalsPerSentence	Syntax	Form	Phrase complexity
F28	complexNominalsPerTUnit	Syntax	Form	Phrase complexity
F29	complexNominalsPerClause	Syntax	Form	Phrase complexity
F30	complexNominalsPerFiniteClause	Syntax	Form	Phrase complexity
F31	wordsInNPsPerNP	Syntax	Form	Text length
F32	wordsInVPsPerVP	Syntax	Form	Text length
F33	wordsInPPsPerPP	Syntax	Form	Text length
F34	sumLongestDependenciesPerSentence	Syntax	Form	Dependencies
F35	sumLongestDependenciesPerTUnit	Syntax	Form	Dependencies
F36	sumLongestDependenciesPerClause	Syntax	Form	Dependencies
F37	sumLongestDependenciesPerFiniteClause	Syntax	Form	Dependencies
F38	depClausesPerSentence	Syntax	Form	Dependencies
F39	depClausesPerTUnit	Syntax	Form	Dependencies

F40	depClausesPerClause	Syntax	Form	Dependencies
F41	depClausesPerFiniteClause	Syntax	Form	Dependencies
F42	sumNonTerminalNodesPerSentence	Syntax	Form	Syntactic complexity: Parse tree features
F43	sumNonterminalsPerWord	Syntax	Form	Syntactic complexity: Parse tree features
F44	sumNonTerminalNodesPerTUnit	Syntax	Form	Syntactic complexity: Parse tree features
F45	sumNonTerminalNodesPerClause	Syntax	Form	Syntactic complexity: Parse tree features
F46	sumNonTerminalNodesPerFiniteClause	Syntax	Form	Syntactic complexity: Parse tree features
F47	sumParseTreeHeightsPerSentence	Syntax	Form	Syntactic complexity: Parse tree features
F48	sumParseTreeHeightsPerTUnit	Syntax	Form	Syntactic complexity: Parse tree features
F49	sumParseTreeHeightsPerClause	Syntax	Form	Syntactic complexity: Parse tree features
F50	sumParseTreeHeightsPerFiniteClause	Syntax	Form	Syntactic complexity: Parse tree features
F51	WordsPerSentence	Syntax	Form	Sentence length
F52	WordsPerTUnit	Syntax	Form	Sentence length
F53	WordsPerClause	Syntax	Form	Sentence length
F54	WordsPerFiniteClause	Syntax	Form	Sentence length
F55	coordinatedPhrasesPerSentence	Syntax	Form	Phrase coordination
F56	coordinatedPhrasesPerTUnit	Syntax	Form	Phrase coordination
F57	coordinatedPhrasesPerClause	Syntax	Form	Phrase coordination
F58	coordinatedPhrasesPerFiniteClause	Syntax	Form	Phrase coordination
F59	coordinatePhrasesPerSentence	Syntax	Form	Phrase coordination
F60	coordinatePhrasesPerTUnit	Syntax	Form	Phrase coordination
F61	coordinatePhrasesPerClause	Syntax	Form	Phrase coordination
F62	coordinatePhrasesPerFiniteClause	Syntax	Form	Phrase coordination
F63	complexTUnitsPerSentence	Syntax	Form	Syntactic complexity: T-units
F64	complexTUnitsPerTUnit	Syntax	Form	Syntactic complexity: T-units
F65	tUnitsPerSentence	Syntax	Form	Syntactic complexity: T-units
F66	clausesPerSentence	Syntax	Form	Clause types and sentence relations
F67	clausesPerTUnit	Syntax	Form	Clause types and sentence relations
F68	dependentClausesPerSentence	Syntax	Form	Clause types and sentence relations
F69	dependentClausesPerTUnit	Syntax	Form	Clause types and sentence relations
F70	dependentClausesPerClause	Syntax	Form	Clause types and sentence relations
F71	dependentClausesPerFiniteClause	Syntax	Form	Clause types and sentence relations

F72	conjunctionalClausesPerSentence	Syntax	Form	Clause types and sentence relations
F73	conjunctionalClausesPerTUnit	Syntax	Form	Clause types and sentence relations
F74	conjunctionalClausesPerClause	Syntax	Form	Clause types and sentence relations
F75	conjunctionalClausesPerDependentClauseWithConjunction	Syntax	Form	Clause types and sentence relations
F76	conjunctionalClausesPerFiniteClause	Syntax	Form	Clause types and sentence relations
F77	dependentClausesWithConjunctionPerSentence	Syntax	Form	Clause types and sentence relations
F78	dependentClausesWithConjunctionPerTUnit	Syntax	Form	Clause types and sentence relations
F79	dependentClausesWithConjunctionPerClause	Syntax	Form	Clause types and sentence relations
F80	dependentClausesWithConjunctionPerDependantClause	Syntax	Form	Clause types and sentence relations
F81	dependentClausesWithConjunctionPerFiniteClause	Syntax	Form	Clause types and sentence relations
F82	dependentClausesWithoutConjunctionPerSentence	Syntax	Form	Clause types and sentence relations
F83	dependentClausesWithoutConjunctionPerTUnit	Syntax	Form	Clause types and sentence relations
F84	dependentClausesWithoutConjunctionPerClause	Syntax	Form	Clause types and sentence relations
F85	dependentClausesWithoutConjunctionPerDependentClause	Syntax	Form	Clause types and sentence relations
F86	dependentClausesWithoutConjunctionPerFiniteClause	Syntax	Form	Clause types and sentence relations
F87	interrogativeClausesPerSentence	Syntax	Form	Clause types and sentence relations
F88	interrogativeClausesPerTUnit	Syntax	Form	Clause types and sentence relations
F89	interrogativeClausesPerClause	Syntax	Form	Clause types and sentence relations
F90	interrogativeClausesPerDependentClauseWithConjunction	Syntax	Form	Clause types and sentence relations
F91	interrogativeClausesPerFiniteClause	Syntax	Form	Clause types and sentence relations
F92	relativeClausePerSentence	Syntax	Form	Clause types and sentence relations
F93	relativeClausePerTUnit	Syntax	Form	Clause types and sentence relations
F94	relativeClausePerClause	Syntax	Form	Clause types and sentence relations
F95	relativeClausePerDependentClauseWithConjunction	Syntax	Form	Clause types and sentence relations
F96	relativeClausePerFiniteClause	Syntax	Form	Clause types and sentence relations
F97	eventivePassivePerSentence	Syntax	Form	Academic language: use of passive voice
F98	eventivePassivePerTUnit	Syntax	Form	Academic language: use of passive voice
F99	eventivePassivePerClause	Syntax	Form	Academic language: use of passive voice
F100	eventivePassivePerFiniteClause	Syntax	Form	Academic language: use of passive voice
F101	sentenceInfinitivesPerSentence	Syntax	Form	Sentence infinitives
F102	sentenceInfinitivesPerTUnit	Syntax	Form	Sentence infinitives
F103	sentenceInfinitivesPerClause	Syntax	Form	Sentence infinitives

F104	sentenceInfinitivesPerDependentClause	Syntax	Form	Sentence infinitives
F105	sentenceInfinitivesPerFiniteClause	Syntax	Form	Sentence infinitives
F106	NpDependentsPerNounWithDependents	Syntax	Form	Dependencies
F107	VpDependentsPerVerbWithDependents	Syntax	Form	Dependencies
F108	VpExclModalsDependentsPerVerbWithDependents	Syntax	Form	Dependencies
F109	NpModifiersPerNP	Syntax	Form	Dependencies
F110	VpModifiersPerVP	Syntax	Form	Dependencies
F111	longestDependency	Syntax	Form	Dependencies
F113	syllablesPerToken	Lexis	Form	Text length
F114	charactersPerToken	Lexis	Form	Text length
F120	uberIndex	Lexis	Form	Lexical diversity
F121	yulesK	Lexis	Form	Lexical diversity
F122	HDD	Lexis	Form	Lexical diversity
F123	MTLD	Lexis	Form	Lexical diversity
F124	lexicalTypesPerLexicalToken	Lexis	Form	Variation
F125	lexicalTypesPerToken	Lexis	Form	Variation
F126	nonAuxVerbTypesPerNonAuxToken	Lexis	Form	Variation
F127	nonAuxVerbTypesPerNonAuxVerbToken	Lexis	Form	Variation
F128	squaredNonAuxVerbTypesPerNonAuxVerb	Lexis	Form	Variation
F129	correctedNonAuxVerbTypesPerNonAuxVerb	Lexis	Form	Variation
F130	nonAuxVerbsPerToken	Lexis	Form	Variation
F131	seinInstancesPerVerb	Lexis	Form	Variation
F132	habenInstancesPerVerb	Lexis	Form	Variation
F133	nounsPerLexicalToken	Lexis	Form	Variation
F134	nounsPerToken	Lexis	Form	Variation
F135	verbsPerNoun	Lexis	Form	Variation
F136	adjectivesPerLexicalToken	Lexis	Form	Variation
F137	adverbsPerLexicalToken	Lexis	Form	Variation
F138	adjectivesAndAdverbsPerLexicalToken	Lexis	Form	Variation
F139	annotatedTypeFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency
F140	typeFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency
F141	lemmaFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency

F142	logAnnotatedTypeFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency
F143	logTypeFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency
F144	logLemmaFreqsPerTypeFormundInDlex	Lexis	Form	Lexical frequency
F145	logAnnotatedTypeFreqBand1PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F146	logAnnotatedTypeFreqBand2PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F147	logAnnotatedTypeFreqBand3PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F148	logAnnotatedTypeFreqBand4PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F149	logAnnotatedTypeFreqBand5PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F150	logAnnotatedTypeFreqBand6PerTypeFormundInDlex	Lexis	Form	Lexical frequency
F151	typesNotFormundInDlexPerLexicalType	Lexis	Form	Lexical frequency
F152	typesFormundInDlexPerLexicalType	Lexis	Form	Lexical frequency
F153	typeFreqsPerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F154	logTypeFreqsPerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F155	logAnnotatedTypeFreqBand1PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F156	logAnnotatedTypeFreqBand2PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F157	logAnnotatedTypeFreqBand3PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F158	logAnnotatedTypeFreqBand4PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F159	logAnnotatedTypeFreqBand5PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F160	logAnnotatedTypeFreqBand6PerTypeFormundInSubtlex	Lexis	Form	Lexical frequency
F161	typesNotFormundInSubtlexPerLexicalType	Lexis	Form	Lexical frequency
F162	typesFormundInSubtlexPerLexicalType	Lexis	Form	Lexical frequency
F163	typeFreqsPerTypeFormundInGoogle00	Lexis	Form	Lexical frequency
F164	logTypeFreqsPerTypeFormundInGoogle00	Lexis	Form	Lexical frequency
F174	typeFreqsPerTypeFormundInKCT	Lexis	Form	Lexical frequency
F175	lemmaFreqsPerTypeFormundInKCT	Lexis	Form	Lexical frequency
F178	logAnnotatedTypeFreqBand1PerTypeFormundInKCT	Lexis	Form	Lexical frequency
F179	logAnnotatedTypeFreqBand2PerTypeFormundInKCT	Lexis	Form	Lexical frequency
F180	logAnnotatedTypeFreqBand3PerTypeFormundInKCT	Lexis	Form	Lexical frequency
F181	logAnnotatedTypeFreqBand4PerTypeFormundInKCT	Lexis	Form	Lexical frequency
F182	logAnnotatedTypeFreqBand5PerTypeFormundInKCT	Lexis	Form	Lexical frequency
F183	lexTypesNotFormundInKCTPerLexicalType	Lexis	Form	Lexical frequency
F184	typesFormundInKCTPerLexicalType	Lexis	Form	Lexical frequency

F185	lexLemmasNotFormundInKCTPerLexicalLemma	Lexis	Form	Lexical frequency
F186	lemmasFormundInKCTPerLexicalLemma	Lexis	Form	Lexical frequency
F187	AoA_sumTypesAoAPerTypeFormundInKCT	Lexis	Form	Lexical frequency
F188	AoA_sumTypesMinAoAPerTypeFormundInKCT	Lexis	Form	Lexical frequency
F189	AoA_sumLemmaAoAPerLemmaFormundInKCT	Lexis	Form	Lexical frequency
F190	AoA_ sumLemmaMinAoAPerLemmaFormundInKCT	Lexis	Form	Lexical frequency
F193	hypernymPerTypeFormundInGnet	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F194	hyponymPerTypeFormundInGnet	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F195	synsetPerTypeFormundInGnet	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F196	lexUnitPerSynset	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F197	relationsPerSynset	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F198	framesPerVerbTypeFormundInGnet	Lexis	Meaning	Semantic relatedness, concrete vs. abstract language
F199	VerbsPerSentence	Syntax	Form	Academic language: verb use
F200	VerbsPerTUnit	Syntax	Form	Academic language: verb use
F201	VerbsPerClause	Syntax	Form	Academic language: verb use
F202	VerbsPerFiniteClause	Syntax	Form	Academic language: verb use
F203	istTPerToken	Lexis	Form	Academic language: Derivation
F204	eiTPerToken	Lexis	Form	Academic language: Derivation
F205	lingTPerToken	Lexis	Form	Academic language: Derivation
F206	keitTPerToken	Lexis	Form	Academic language: Derivation
F207	atTPerToken	Lexis	Form	Academic language: Derivation
F208	werkTPerToken	Lexis	Form	Academic language: Derivation
F209	schaftTPerToken	Lexis	Form	Academic language: Derivation
F210	enzTPerToken	Lexis	Form	Academic language: Derivation
F211	tumTPerToken	Lexis	Form	Academic language: Derivation
F212	astTPerToken	Lexis	Form	Academic language: Derivation
F215	urTPerToken	Lexis	Form	Academic language: Derivation
F216	heitTPerToken	Lexis	Form	Academic language: Derivation
F217	nisTPerToken	Lexis	Form	Academic language: Derivation
F218	wesenTPerToken	Lexis	Form	Academic language: Derivation
F219	atorTPerToken	Lexis	Form	Academic language: Derivation
F220	ismusTPerToken	Lexis	Form	Academic language: Derivation

F221	aturTPerToken	Lexis	Form	Academic language: Derivation
F222	entTPerToken	Lexis	Form	Academic language: Derivation
F223	antTPerToken	Lexis	Form	Academic language: Derivation
F225	ungTPerToken	Lexis	Form	Academic language: Derivation
F226	ionTPerToken	Lexis	Form	Academic language: Derivation
F227	compoundDepthsPerCompoundNoun	Lexis	Form	Academic language: Compounding
F228	compundNounsPerNP	Lexis	Form	Academic language: Compounding
F229	derivedNounsPerNoun	Lexis	Form	Academic language: Derivation
F230	NominativesPerNoun	Lexis	Form	Inflectional features
F231	AccusativesPerNoun	Lexis	Form	Inflectional features
F232	GenitivesPerNoun	Lexis	Form	Inflectional features
F233	DativesPerNoun	Lexis	Form	Inflectional features
F234	finiteVerbsPerVerb	Lexis	Form	Inflectional features
F235	infiniteVerbsPerVerb	Lexis	Form	Inflectional features
F236	participleVerbsPerVerb	Lexis	Form	Inflectional features
F237	imperativeVerbsPerFiniteVerb	Lexis	Form	Inflectional features
F238	subjunctiveMarkingsPerFiniteVerb	Lexis	Form	Inflectional features
F239	indicativeMarkingsPerFiniteVerb	Lexis	Form	Inflectional features
F241	firstPersonMarkingsPerFiniteVerb	Lexis	Form	Inflectional features
F242	secondPersonMarkingsPerFiniteVerb	Lexis	Form	Inflectional features
F243	thirdPersonMarkingsPerFiniteVerb	Lexis	Form	Inflectional features
F244	modalVerbsPerVerb	Lexis	Form	Academic language: verb use
F245	auxiliaryVerbsPerVerb	Lexis	Form	Academic language: verb use
F246	localNounOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F247	globalNounOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F248	localArgOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F249	globalArgOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F250	localContentOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F251	globalContentOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F252	localStemOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F253	globalStemOverlapsPerSentence	Text	Meaning	Cohesion: global and local argument/noun/content overlap
F257	probSubSubsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles

F258	probSubObjsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F259	probSubOthsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F260	probSubNotsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F261	probObjSubsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F262	probObjObjsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F263	probObjOthsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F264	probObjNotsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F265	probOthSubsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F266	probOthObjsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F267	probOthOthsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F268	probOthNotsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F269	probNotSubsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F270	probNotObjsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F271	probNotOthsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F272	probNotNotsPerTransition	Text	Meaning	Cohesion: Transition of grammatical roles
F273	pronounsPerToken	Lexis	Form	Variation
F274	pronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F275	pronounsPerNoun	Lexis	Form	Variation
F276	persPronounsPerToken	Lexis	Form	Variation
F277	persPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F278	persPronounsPerNoun	Lexis	Form	Variation
F279	possPronounsPerToken	Lexis	Form	Variation
F280	possPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F281	possPronounsPerNoun	Lexis	Form	Variation
F282	3PPersPronounsPerToken	Lexis	Form	Variation
F283	3PPersPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F284	3PPersPronounsPerNoun	Lexis	Form	Variation
F285	3PPossPronounsPerToken	Lexis	Form	Variation
F286	3PPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F287	3PPossPronounsPerNoun	Lexis	Form	Variation
F288	3PPersAndPossPronounsPerToken	Lexis	Form	Variation
F289	3PPersAndPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation

F290	3PPersAndPossPronounsPerNoun	Lexis	Form	Variation
F291	1PPersPronounsPerToken	Lexis	Form	Variation
F292	1PPersPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F293	1PPersPronounsPerNoun	Lexis	Form	Variation
F294	1PPossPronounsPerToken	Lexis	Form	Variation
F295	1PPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F296	1PPossPronounsPerNouns	Lexis	Form	Variation
F297	1PPersAndPossPronounsPerToken	Lexis	Form	Variation
F298	1PPersAndPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F299	1PPersAndPossPronounsPerNoun	Lexis	Form	Variation
F300	2PPersPronounsPerToken	Lexis	Form	Variation
F301	2PPersPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F302	2PPersPronounsPerNoun	Lexis	Form	Variation
F303	2PPossPronounsPerToken	Lexis	Form	Variation
F304	2PPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F305	2PPossPronounsPerNouns	Lexis	Form	Variation
F306	2PPersAndPossPronounsPerToken	Lexis	Form	Variation
F307	2PPersAndPossPronounsPerTokenInSentencePerSentence	Lexis	Form	Variation
F308	2PPersAndPossPronounsPerNoun	Lexis	Form	Variation
F309	definiteArticlesPerArticle	Lexis	Form	Variation
F310	definiteArticlesPerTokenInSentencePerSentence	Lexis	Form	Variation
F311	indefiniteArticlesPerArticle	Lexis	Form	Variation
F312	indefiniteArticlesPerTokenInSentencePerSentence	Lexis	Form	Variation
F313	properNamesPerNoun	Lexis	Form	Variation
F314	properNamesPerToken	Lexis	Form	Variation
F315	properNamesPerSentence	Lexis	Form	Variation
F319	AdversativeConcessiveConnectivePerSentence	Text	Meaning	Cohesion: Connectives (list of Eisenberg 2016)
F321	all multi ConnectivePerSentence	Text	Meaning	Cohesion: Connectives (list of Breindl et al. 2014)
F323	all multi ConnectivePerSentence	Text	Meaning	Cohesion: Connectives (list of Eisenberg 2016)
F325	adversativeConnectivePerSentence	Text	Meaning	Cohesion: Connectives (list of Eisenberg 2016)
F327	concessiveConnectivePerSentence	Text	Meaning	Cohesion: Connectives (list of Breindl et al. 2014)