DOTTORATO DI RICERCA IN

TRADUZIONE, INTERPRETAZIONE E INTERCULTURALITA'

Ciclo 36

Settore Concorsuale: 10/L1 - LINGUE, LETTERATURE E CULTURE INGLESE E ANGLO - AMERICANA

Settore Scientifico Disciplinare: L-LIN/12 - LINGUA E TRADUZIONE - LINGUA INGLESE

BRIDGING THE GAP EXPLORING THE COGNITIVE IMPACT OF INTERPRETBANK ON CHINESE INTERPRETING TRAINEES

Presentata da: Zhiqiang Du

Coordinatore Dottorato	**Supervisore**
Chiara Elefante	Ricardo Munoz Martin

Esame finale anno 2024

We are all in the gutter,
but some of us
are looking at the stars

Oscar Wilde

E quindi uscimmo a riveder le stelle

Dante Alighieri

To my wife, CHEN Wanlu

献给我的妻子陈芄漉

acknowledgments

My interest in computer-assisted interpreting tools departed from my master's degree study, as I continually pondered how I could improve my interpreting quality through such tools. As the field of computer-assisted interpreting tools continues evolving alongside rapid advances in artificial intelligence, my academic interest in enhancing human-computer interaction has deepened. I feel fortunate to have had the opportunity to pursue a PhD project in the lovely town of Forli. The journey through a PhD program is rarely smooth, but I feel fortunate to have reached this milestone with supportive and helpful guidance. The completion of this dissertation would have been impossible without the invaluable support and guidance of several individuals and groups.

First and foremost, I would like to express my deepest gratitude to my supervisors: Professors Ricardo Muñoz Martín at the University of Bologna and Lei Victoria Lai Cheng at the University of Macau. Their knowledgeable input helped shape the research design, goals, informant recruitment process, and writing of this dissertation. Their insightful feedback and unwavering support were instrumental in refining this work and encouraging me to approach seemingly impossible tasks with a positive attitude.

Conducting a project is one thing, but telling its story is another. I am also indebted to the external reviewers, Dr. Bianca Prandi and Dr. Ho Chen-En. Their valuable time, invested effort, and comments to reviewing this dissertation inspired me tremendously and encouraged me to think more critically about my work from different perspectives.

I extend my sincere gratitude to the interpreting practitioners, trainers, and researchers who contributed to this work to varying degrees: Deng Yi, Francesca Frittella, Han Chao, He Sui, Liu Yu, Christopher D. Mellinger, Christian Olalla Soler, Richard D. Samson, Tian Ye, and Wang Weiwei. They were involved in experiment text creation, informant recruitment, and research design. Special thanks also go to Jennifer Monroe and Marissa López Paredes for their speech recordings.

I would like to acknowledge and thank all the informants who participated in the pilot study and the main study. Their devoted time, willingness to accompany me throughout the entire data collection period, and promptness in reporting any technical issues were invaluable. I am also grateful to my five PhD raters for providing their expertise and valuable evaluations.

Finally, I take this opportunity to thank the service and staff of the Forli library, especially Maria Letizia Montanari, for providing access to external literature and resources, which proved immensely helpful in broadening the horizons of my research and incorporating insights from other related disciplines.

abstract

This is an exploratory research project to investigate how to develop a cognitive situated approach to studying aspects of simultaneous interpreting in quantitative, confirmatory research approaches. On the surface, this project studied the potential benefits of using InterpretBank in twenty-two Chinese L1 and English L2 interpreting trainees. This dissertation also aims to open up new possibilities while it strives to connect with our rich tradition—in particular, it combines emergent tendencies and trends that seem to lead in the right direction. For these reasons, sometimes it may feel like a hybrid.

Enrolled in Chinese MA interpreting programs, the informants were mostly 2nd-year, females, with an average age of 24.7. The source materials for these tasks were (topic-, register- and speaker) matching pairs of the transcripts of popular science podcasts. Informants worked on one transcript from each pair to compile their glossaries. The other text in each pair was edited for SI and enhanced with 33 potential problem triggers (unigrams, bigrams, trigrams). Three terms out of those 33 were repeated twice to study rehearsal and recall. Added terms were selected using BootCaT and AntConc. The scripts were validated by an English L1 interpreter and interpreting trainer, recorded by three English L1 speakers, and streamed via Microsoft Stream.

This exploratory study adopted a pretest and posttest design, collecting data remotely through screen recording (TechSmith Capture), keylogging (Pynput), and surveys (Microsoft Forms and PsyToolkit) in three cycles, each corresponding to one text pair. The independent variable was the use of Excel or InterpretBank. After Cycle I (pre-test and baseline), the sample was split into control (Excel) and experimental (InterpretBank) groups. All informants received a treatment. Excel informants, on searching multimodal information; experimental informants, on the use of relevant InterpretBank features. Tool choice was compulsory in Cycle II (post-test 1), but not in Cycle III (post-test 2). Individual glossaries were replaced by master glossaries that informants revised before each RSI task. We adopted a cluster of fluency and accuracy indicators plus ear-key span and eye-voice span.

All informants complied their own glossaries on the text A in each pair, which were merged and adapted by the researcher into master glossaries that were then returned to informants for them to review, tweak, and use such master glossaries in the booth tasks. InterpretBank informants spent less time on glossary compilation, generated more terms, and took less *time per term* than Excel informants did when compiling their glossaries. However, InterpretBank glossaries compiled with automatic term extraction were less diverse and longer. No significant differences between groups were noted in *fluency* indicators across the cycles, except for an increase in *bumps* (production flow gaps between 200 and 600ms) in Cycle II for the InterpretBank group. InterpretBank informants produced more correct renditions in Cycles II and III, but no statistically significant difference was observed per cycle between groups in other accuracy indicators (i.e., correct, adequate, wrong, and skipped terms).

Five interpreting PhD volunteers holistically assessed the quality of informants' renderings. Their assessments were checked for inter-rater reliability and also cross-referenced with the clusters of fluency and accuracy quantitative performance indicators. InterpretBank informants consistently outperformed Excel informants in quality ratings throughout the cycles, suggesting a positive impact of InterpretBank on RSI rendering quality. Some InterpretBank implementations may raise concerns regarding cognitive ergonomics, particularly in the context of logographic languages like Chinese, potentially diminishing its utility. In a nutshell, from the perspective of determining the potential benefits of using InterpretBank for Chinese interpreting trainees, results were mixed. From the perspective of developing methods for the cognitive situated study of interpreting, I humbly think the project holds promise.

riassunto

Questo progetto è incentrato sullo studio dei potenziali benefici dell'uso di InterpretBank in ventidue interpreti (L1 cinese e L2 inglese). Tutti i partecipanti hanno svolto tre cicli di attività, ciascuno composto da un task di compilazione di glossari e un task in cabina RSI. Il primo ciclo è stato utilizzato come punto di riferimento comportamentale. In seguito, gli informatori sono stati divisi in due gruppi e trattati in maniera differente, ovvero svolgendo i cicli II e III utilizzando strumenti diversi (Excel o InterpretBank). La scelta dello strumento era obbligata nel Ciclo II e libera nel Ciclo III. I dati sono stati raccolti da remoto.

I partecipanti, iscritti a un master di interpretariato in una università cinese di alto livello, erano per lo più studenti del secondo anno, di sesso femminile, con un'età media di 24,7 anni. I materiali usati come fonte per i task erano coppie di script (con dei temi e i rispettivi relatori) di podcast di divulgazione scientifica. I partecipanti hanno lavorato con un testo della coppia per compilare i loro glossari, mentre l'altro testo è stato modificato per l'interpretazione simultanea e arricchito con 33 potenziali trigger di problemi (unigrammi, bigrammi, trigrammi). Nei testi di interpretazione simultanea, tre di quei 33 termini sono stati ripetuti due volte per studiare gli effetti della ripetizione e del richiamo. I termini aggiunti sono stati selezionati utilizzando BootCaT e AntConc. Gli script sono stati convalidati da un interprete madrelingua inglese e docente di interpretariato, registrati da tre parlanti nativi inglesi e trasmessi tramite Microsoft Stream.

Questo studio esplorativo ha adottato un design di pre-test e post-test, raccogliendo dati da remoto attraverso la registrazione dello schermo (TechSmith Capture), il keylogging (Pynput) e sondaggi (Microsoft Forms e PsyToolkit) in tre cicli, ciascuno corrispondente a una coppia di testi. La variabile indipendente era l'uso di Excel o InterpretBank. Dopo il Ciclo I (pre-test e punto di riferimento), il campione è stato diviso in gruppi, di controllo (Excel) e sperimentale (InterpretBank). I partecipanti che hanno lavorato con Excel si sono concentrati sulla ricerca di informazioni multimodali, mentre i partecipanti sperimentali si sono concentrati sull'uso di funzionalità rilevanti di InterpretBank. La scelta dello strumento era obbligata nel Ciclo II (post-test 1), ma non nel Ciclo III (post-test 2). I glossari individuali sono stati sostituiti da glossari generali che gli informatori hanno revisionato prima di ogni task RSI. Nello studio è stato adottato un gruppo di indicatori di fluenza e di accuratezza, oltre all'ear-key span e all'eye-voice span.

Tutti i partecipanti hanno compilato i propri glossari sul testo A di ogni coppia. I glossari individuali sono stati uniti e adottati come glossari generali dai partecipanti nei task in cabina.

I partecipanti che hanno utilizzato InterpretBank hanno impiegato meno tempo nella compilazione dei glossari, hanno generato più termini e hanno impiegato meno tempo per ogni termine rispetto agli utenti di Excel. Tuttavia, i glossari InterpretBank compilati con l'estrazione automatica dei termini sono risultati meno diversificati e più lunghi. Non sono state rilevate differenze significative tra i gruppi negli indicatori di fluidità nei vari cicli, a eccezione di un aumento dei "bump" (vuoti di produzione tra 200 e 600 ms) nel II ciclo per il gruppo InterpretBank. Gli informatori di InterpretBank hanno prodotto un maggior numero di rese corrette nei cicli II e III, ma non è stata osservata alcuna differenza statisticamente significativa tra i gruppi per quanto riguarda gli altri indicatori di accuratezza (ovvero, termini adeguati, errati e omessi).

Cinque volontari tra gli studenti di dottorato in interpretazione hanno valutato olisticamente la qualità della resa degli informatori. Le loro valutazioni sono state controllate per verificare l'affidabilità inter-rater e sono state incrociate con i cluster di indicatori di performance quantitativi di fluidità e accuratezza.

Gli informatori di InterpretBank hanno costantemente superato gli informatori di Excel nelle valutazioni della qualità durante tutti i cicli, suggerendo un impatto positivo di InterpretBank sulla qualità dei rendering RSI. Tuttavia, alcune implementazioni di InterpretBank possono sollevare problemi di ergonomia cognitiva, in particolare nel contesto di lingue logografiche come il cinese, riducendone potenzialmente l'utilità.

glossary

ASR	Automatic Speech Recognition
CAI	Computer Assisted Interpreting
CI	Consecutive Interpreting
cognitive demand	Environmental or task-related factors which tend to induce a certain level or degree of cognitive efforts
cognitive effort	Level or degree of mental resources or energy applied to meet self-generated, environmental or task demands
cognitive load	a broad term that may refer to either the demand or the effort, or both
CSV	comma-separated values
CTIS	Cognitive Translation and Interpreting Studies
E2K	ear-key span
I2V	eye-voice span
EVS	ear-voice span
EVS1	Chunk-initial EVS
EVS2	Chunk-final EVS
IKI	Inter keystroke Interval (time span between two keypresses)
multimodal ethograms	Grid of synchronized, aligned behavioral data including the source speech soundtrack, multimodal movie with voice-over, SI rendering/source speech soundtrack and keylogging data.
RSI	Remote Simultaneous Interpreting
RT	reaction time or response time
SI	Simultaneous Interpreting
WM	working memory

table of contents

list of figures xv

list of tables xviii

Chapter 1 introduction 1

1.1 Cognitive framework 3

1.2 Situating the human mind 5

1.3 Language and communication 7

1.4 Interpreting 11

1.5 Computer-assisted interpreting tools 13

1.5.1 Human-computer interaction 14

1.5.2 Research on CAI tools 14

1.5.3 InterpretBank 18

1.6 Quality evaluation of CAI-supported interpreting 20

1.6.1 Terminology accuracy 21

1.6.2 Evaluating the performance of CAI tool users 22

1.7 Research question and hypotheses 23

1.8 Summary 24

Chapter 2 materials and methods 27

2.1 Informants 28

2.2 Input 30

2.2.1 Source text preparation 30

2.2.2 Potential problem triggers 32

2.2.3 Source speech recordings 34

2.3 Application variables 35

2.3.1 Excel 36

2.3.2 InterpretBank 36

2.4 Data collection applications 40

2.4.1 Microsoft Stream 41

2.4.2 Keylogger 41

2.4.3 Screen recording 44

2.5 Study design 46

2.5.1 Glossary tasks 47

2.5.2 Booth tasks 48

2.5.3 InterpretBank training 49

2.5.4 Survey 49
2.5.5 Holistic assessment 50
2.5.5.1 *Raters* 50
2.5.5.2 *Audio preparation* 51
2.5.5.3 *Rating procedures* 53
2.6 Data collection 56
2.7 Data analysis 57
2.7.1 Data cleaning procedures 58
2.7.1.1 *Procedures for glossary task* 58
2.7.1.2 *Procedures for booth tasks* 59
2.7.1.3 *Punctual events and span constructs* 65
2.7.1.4 *Coding behaviors* 67
2.7.1.5 *Ear-key span and eye-voice span* 70
2.7.2 Indicators 72
2.7.2.1 *Indicators in glossary tasks* 72
2.7.2.2 *Indicators in booth tasks* 78
2.8 Summary 80

Chapter 3 results 83

3.1 Glossaries 83
3.1.1 Individual glossary 83
3.1.1.1 *Compilation* 84
3.1.1.2 *Sources of consultation* 90
3.1.1.3 *Glossary contents* 96
3.1.2 Glossary review (*Memo* mode) 98
3.2 Fluency analysis 99
3.2.1 False starts 105
3.2.2 Self-corrections 106
3.2.3 Fillers 107
3.2.4 Repetitions 109
3.2.5 Bumps 110
3.2.6 Respites 111
3.2.7 Chunk-initial EVS (*EVS1*) 113
3.2.8 Chunk-final EVS (*EVS2*) 114
3.2.9 Duration of source speech chunks and EVS 115
3.3 Term accuracy analysis 120
3.3.1 Correct renditions 123
3.3.2 Adequate renditions 125
3.3.3 Wrong renditions 126
3.3.4 Skipped terms 127
3.3.5 Repeated potential problem triggers 129

3.4 Search behavior of the InterpretBank group in Cycles II and III 131
3.4.1 Search workflows with InterpretBank 132
3.4.2 Ear-key span and eye-voice span 137
3.4.3 Problem triggers with InterpretBank search 139
3.4.4 Search duration and dropped chunks 151
3.5 Survey analysis 154
3.5.1 Surveys of the InterpretBank informants 154
3.5.1.1 *Overall opinion* 154
3.5.1.2 *Glossary task* 156
3.5.1.3 *Booth task* 158
3.5.2 Survey for the Excel group 160
3.5.2.1 *Overall opinion* 160
3.5.2.2 *Glossary task* 161
3.5.2.3 *Booth task* 162
3.6 Holistic quality assessment of informants' audios 163
3.6.1 Individual rater analysis 163
3.7 Summary 171

Chapter 4 discussion 173

4.1 H1: InterpretBank improves efficiency at glossary compilation 173
4.2 H2: InterpretBank improves the quality of RSI rendering 177
4.3 H3: InterpretBank improves efficiency when producing the RSI rendering 179
4.3.1 Fluency: Content cluster 179
4.3.2 Fluency: Time cluster 182
4.3.3 Term accuracy: potential problem triggers 185
4.3.4 Term accuracy: repeated terms 189
4.3.5 Summary 191
4.4 H4: Improved documentation performance will yield better RSI rendering quality(H3) .. 194
4.5 H5: Improvements using InterpretBank but also attitudes, will lead to keep using it 197
4.6 Compilation of glossaries 200
4.6.1 Sources 205
4.6.2 Contents 207
4.6.3 Flashcard use (InterpretBank group) 207
4.7 Findings on search behavior from the InterpretBank group 207
4.7.1 Searching workflow 208
4.7.2 Ear-key span and eye-voice span 211
4.7.3 Potential problem triggers with InterpretBank search 213
4.7.4 Search duration and dropped chunks 215
4.8 Duration of source speech chunks and EVS 216
4.9 Holistic assessment by raters 217
4.10 Summary 218

Chapter 5 conclusion, limitations, and further research ... 223

5.1 General conclusions and implications ... 225

5.2 Limitations ... 228

5.3 Future work ... 230

references ... 233

appendices ... 251

Appendix A Profiling survey for informants ... 253

Appendix B Informants' profiles ... 255

Appendix C Potential problem triggers ... 256

Appendix D Survey for Excel group in Cycle III ... 259

Appendix E Surveys for InterpretBank group in Cycles II and III ... 260

Appendix F Follow-up survey ... 261

Appendix G Information sheet for Phd raters ... 262

Appendix H Chunks and sentences of source speech texts ... 264

Appendix I Other applications / services in glossary tasks in Cycles II &III ... 276

Appendix J Individual rating scores per text ... 278

list of figures

1. A representation of cascade processing. 9
2. Source speech arrangement. 31
3. Procedures to identify 33 potential problem triggers. 32
4. Components of source speech soundtrack. 35
5. InterpretBank welcome view. 37
6. InterpretBank’s edit mode. 37
7. Automatically terminology extraction. 38
8. Manual extraction option in InterpretBank. 39
9. Screenshot of memo mode. 39
10. Booth mode in InterpretBank. 40
11. Sample logfile from Pynput. 43
12. Voice-over screen recording. 45
13. A simplified randomized audio procedure. 52
14. Individual survey structure. 54
15. Assessment of one audio in the PsyToolkit survey. 55
16. Project website. 56
17. Conversion table from Unix time to chrono time. 58
18. Individual multimodal data synchronization. 59
19. The first alignment workflow. 60
20. Importing Video and Audio into Adobe Premiere Pro. 61
21. The second alignment workflow. 62
22. Partial alignment list showing timestamps using MS Excel. 63
23. The third alignment workflow. 64
24. Computation for universal time. 65
25. Representation of punctual events and span construct. 66
26. Coded behavior in Adobe Premiere Pro for glossary task. 68
27. Workspace in Adobe Audition 2024. 68
28. Layers for coding indicators in Audacity®. 69
29. Representation of ear-key span and eye-voice span. 70
30. Layered approach to task models. 73
31. Labels and structures of subtasks and corresponding strategies. 74
32. Behavior of informants in the glossary task of Cycle I. 87
33. Behavior of informants in the glossary task of Cycle II. 88
34. Behavior of informants in the glossary task of Cycle III. 89
35. Application/service percentage in Cycle I. 92

36. Group-level application/service percentage in Cycle I. 93
37. Application/service percentage in Cycles II and III....94
38. Term counts in individual glossaries, master glossaries, and diversity rates. 97
39. Distribution of false starts across cycles by informants....105
40. Within-group comparison of *false starts* across cycles. 105
41. Comparison of false starts across cycles between groups....105
42. Distribution of *self-corrections* across cycles by informants. 106
43. Within-group comparison of *self-corrections* across cycles. 106
44. Comparison of *self-corrections* across cycles between groups. 107
45. Distribution of fillers across cycles by informant....107
46. Within-group comparison of *fillers* across cycles....108
47. Comparison of *fillers* across cycles between groups. 108
48. Distribution of *repetitions* across cycles by informants. 109
49. Within-group comparison of *repetitions* across cycles. 109
50. Comparison of *repetitions* across cycles between groups. 109
51. Distribution of *bumps* across cycles by informants. 110
52. Within-group comparison of *bumps* across cycles....110
53. Comparison of *bumps* across cycles between groups....111
54. Distribution of *respites* across cycles by informants....111
55. Within-group comparison of *respites* between groups. 112
56. Comparison of *respites* across cycles between groups. 112
57. Distribution of *Chunk-initial EVS* across cycles by informants....113
58. Within-group comparison of *Chunk-initial EVS* across cycles. 113
59. Comparison of *Chunk-initial EVS* across cycles between groups....113
60. Distribution of *Chunk-final EVS* across cycles by informants....114
61. Within-group comparison of *Chunk-final EVS* across cycles. 114
62. Comparison of *Chunk-final EVS* across cycles between groups. 115
63. *EVS1*, *EVS2*, and chunks duration in Cycle I (Excel Group). 117
64. *EVS1*, *EVS2*, and chunks duration in Cycle I (InterpretBank Group)....117
65. *EVS1*, *EVS2*, and chunks duration in Cycle II (Excel Group). 118
66. *EVS1*, *EVS2*, and chunks duration in Cycle II (InterpretBank Group). 118
67. *EVS1*, *EVS2*, and chunks duration in Cycle III (Excel Group). 119
68. *EVS1*, *EVS2*, and chunks duration in Cycle III (InterpretBank Group). 119
69. Percentage of accuracy indicators across cycles by two groups. 122
70. Distribution of *correct* renditions across cycles by informants. 123
71. Within-group comparison of *correct* renditions across cycles....124
72. Comparison of *correct* renditions across cycles between groups. 124
73. Distribution of *adequate* renditions across cycles by informants....125
74. Within-group comparison of *adequate* renditions across cycles. 125
75. Comparison of *adequate* renditions across cycles between groups. 125
76. Distribution of *wrong* renditions across cycles by informants. 126

77. Within-group comparison of *wrong* renditions across cycles. 126
78. Comparison of *wrong* renditions across cycles between groups. 127
79. Distribution of *skipped terms* across cycles by informants. 127
80. Within-group comparison of *skipped terms* across cycles. 128
81. Comparison of *skipped terms* across cycles between groups. 128
82. Stacked percentage bar plot of first-time terms by cycles and groups. 129
83. Stacked percentage bar plot of *repeated* terms by cycles and groups. 130
84. Workflow of search with InterpretBank in Cycle II. 133
85. Workflow of search with InterpretBank in Cycle III. 134
86. Scatter plots of E2K and I2V across Cycles II and III. 138
87. Accuracy of problem triggers and InterpretBank usage for Cycle II. 140
88. Accuracy of problem triggers and InterpretBank usage for Cycle III. 142
89. Search for *repeated terms* in problem triggers in Cycle II. 149
90. Search for *repeated terms* in problem triggers in Cycle III. 150
91. Event counts and durations with overlaps for Cycles II and III. 153
92. Self-assessment of interpreting performance. 154
93. Intent to continue using InterpretBank for glossary preparation. 155
94. Intent to continue using InterpretBank for booth tasks. 155
95. Attitudes toward automatic extraction of technical terms. 156
96. Views on the need for manual extraction. 157
97. InterpretBank convenience for compiling glossaries. 157
98. Frequency of using booth mode for technical terms. 158
99. Confidence in booth mode's term retrieval capabilities. 159
100. Impact of Booth mode on stress levels when handling technical terms. 159
101. Tau-b values for Félix in Cycles I, II, and III. 167
102. Tau-b values for Jules in Cycles I, II, and III. 167
103. Tau-b values for Luc Cycles I, II, and III. 168
104. Tau-b values for Maxime in Cycles I, II, and III. 168
105. Tau-b values for Quentin in Cycles I, II, and III. 169
106. Average percentage of quality ratings for two groups. 178

list of tables

1. Linguistic features of texts in glossary tasks and booth tasks. 31
2. Source speech script features for booth tasks. 34
3. Sample of behaviors, activities, and subtasks in glossary tasks. 77
4. Ethogram subtasks and activities in glossary tasks. 84
5. Automatic term extraction in Cycles III. 85
6. Group-level application/service in Cycles II and III. 95
7. Memo mode usage in Cycle II and Cycle III. 98
8. Fluency indicators across cycles for two groups. 100
9. Overall fluency performance by each informant across Cycles. 104
10. Correlation coefficients and p-values among EVS vs. chunks. 120
11. Accuracy indicators for two groups (Shapiro-Wilk test included). 121
12. Count and percentage of correct searches and correct renditions. 137
13. Statistics and Shapiro-Wilk test for E2K and I2V in Cycles II and III. 138
14. N-gram analysis of problem triggers using InterpretBank in Cycle II. 145
15. N-gram analysis of problem triggers using InterpretBank in Cycle III. 147
16. N-gram counts by accuracy and InterpretBank usage. 148
17. Distances between *repeated terms and1st2Reps* . 148
18. Overall opinion questions for Excel informants. 160
19. Responses to statements related to glossary tasks. 161
20. Responses to statements related to booth tasks. 162
21. Average evaluation scores by raters and groups. 164
22. Correlation between fluency and accuracy indicators among five raters. 166
23. Raters' Taub variation (differences between Max and Min values). 166
24. Glossary compilation indicators for two groups in Cycles II and III. 174
25. Average quality ratings for the recordings. 177
26. Average numbers for accurate and inaccurate renditions. 188
27. P-values in statistical tests for fluency and accuracy indicators 192
28. Glossary task indicators and interpreting quality for two groups. 194
29. Documentation behavior indicators and interpreting quality. 195
30. Individual subtasks and activities percentage in Cycle I. 202
31. Individual subtasks and activities percentage in Cycle II. 203
32. Individual subtasks and activities percentage in Cycle III. 204
33. Correct renditions of *repeated* terms and *1st2Rep* in Cycles II and III 214
34. Potential problem triggers in Cycle I. 256
35. Potential problem triggers in Cycle II. 257
36. Potential problem triggers in Cycle III. 258
37. Informants' rating scores per text. 278

Chapter 1

introduction

Computer-assisted interpreting (CAI) tools have been used for a decade, and a new generation now adds voice recognition that may be incorporated in consecutive interpreting (S. Chen & Kruger, 2023) and simultaneous interpreting. The COVID pandemic led to an explosion in the remote interpreting market and a few studies have found that CAI tools may be beneficial for interpreters when at task (e.g., Defrancq & Fantinuoli, 2021; Fantinuoli *et al.*, 2022; Prandi, 2023).

The goal of this doctoral research project was to develop and formulate a situated approach to the assessment of CAI tool use that will take on board the cognitive dynamics of full tasks of remote SI, mainly to capture the impact of glossary compilation on the booth task. It further aimed to test the usefulness of InterpretBank as an example of third-generation CAI tools for Chinese-L1 interpreting trainees and its potential effects on the quality of renderings.

To elucidate this subject matter, we shall first explicate the terminology employed, subsequently focusing on the area of remote interpreting and its correlation with documentation behavior and conclude with an examination of computer-assisted interpreting tools. *Documentation behavior*, in the present study, refers to the full cycle of finding, processing, storing, and retrieving information. Finding information involves activities like performing web searches (e.g., using keywords in search engines) and consulting online reference materials (print sources are included but rare). Processing information entails analyzing, summarizing, or reorganizing the found information (e.g., translation, pronunciation) to make it more understandable and applicable. Storing information covers actions such as creating new entries in a glossary, knowledgebase, or database, as well as taking notes or writing documentation. Retrieving information refers to searching for and accessing information stored in glossaries, knowledgebases, personal notes, or documentation when needed.

Some parameters in behavioral research defy comprehensive quantification, but quantitative research in this exploratory study is not solely about computing numbers, but about reaching intersubjectively valid understandings. This should make this research project more reliable, but reliability should not be confused with exhaustivity. We caution readers against expecting definitive conclusions about the quality and utility of InterpretBank; rather, our objective is modest—we aim to glean insights through selected quantitative indicators, but our focus is on the methods. In particular, we aim to test several approaches, constructs and indicators that might substantiate a cognitive translatological (situated) approach (among others) to studying interpreting behavior and task results. The present study is thus mainly exploratory, so rather than looking at a target and see a bullet hit it, it may sometimes feel more like watching fireworks in the dark. Yet precisely because it explores new venues for empirical, quantitative, confirmatory research, it sometimes seems to be close to it.

This dissertation adopted the IMRaD (introduction, methods, results, and discussion) structure to organize the content. The first chapter lays the foundation for the research by providing a cognitive framework for language, communication, and interpreting. It discusses human-computer interaction and reviews existing research on CAI tools, specifically InterpretBank. The chapter also covers the quality evaluation of interpreting renditions (see **§ 1.6**), including terminology accuracy and evaluating the performance of CAI tool users. It outlines the research question and hypotheses in **§ 1.7**. This chapter culminates with the formulation of a research question and hypotheses concerning the utilization of CAI-supported interpreting activities, followed by a brief conclusion. Bridging the gap between theoretical exposition and CAI tool, the ensuing chapter transitions from a general discussion on the principles of CTIS to a focused exploration of the specific impacts and implications of InterpretBank in the realm of CAI study.

The second chapter describes the research methods employed in the study. It details the profiling of 22 informants, the input materials (source text preparation, potential problem triggers, and source speech recording), and the application variables (Excel and InterpretBank). The data collection applications, including Microsoft Stream, Pynput (keylogger), and TechSmith Capture (screen recording), are also included. The study design is composed of glossary tasks, booth tasks, InterpretBank training, surveys, and holistic assessment (see **§ 2.5**). Finally, the chapter outlines the data collection and analysis procedures, including data cleaning, coding behaviors, and calculating indicators.

The third chapter presents the results of the study. It analyzes behavioral data collected from individual glossary compilation. Quantitative indicators in fluency analysis have been examined regarding *false starts*, *self-corrections*, *fillers*, *repetitions*, *bumps*, *respites*, *chunk-initial* and *chunk-final ear-voice spans (EVS)*, and the duration of source speech chunks and EVS (see **§ 3.2**). Term accuracy analysis was also performed by evaluating the rendering of potential problem triggers both as *first-time terms* and *repeated terms* (see **§ 3.3**). The search behavior of the InterpretBank group in Cycles II and III was analyzed, covering search workflows, ear-key span, and eye-voice span, problem triggers with InterpretBank search, and *search duration* and *dropped chunks*. The chapter also presents survey results for both the InterpretBank and Excel groups, covering overall opinions, glossary tasks, and booth tasks. Finally, it covers also the results of holistic assessment by raters, whose performance was analyzed through inter-rater reliability analysis but also attending to individual differences in the potential impact of typical quantitative quality parameters, to critically assess the validity of the approach to interpreting quality evaluation.

The fourth chapter opens with the discussion of the research hypotheses. It further discusses the compilation of glossaries and search behavior findings from the InterpretBank group, the duration of source speech chunks and EVS, and holistic assessment by raters. Chapter five addresses the general conclusions and implications of the research, highlighting the main findings and their implications for relevant stakeholders. It also acknowledges the limitations of the study and provides suggestions for future work in the field of CAI and RSI within a cognitive-situated perspective.

Throughout this dissertation, the pronoun *we* will be used to refer to myself. The use of *we* is not merely a way to show modesty; rather, it implies that the work has been promoted and carried out not just by the author, but also by those who have contributed to the research in various ways. This includes but is not limited to, those who helped create the source speech, those who assisted in spreading the word for recruiting informants, and those who contributed to the research design. However, I am the sole author of this dissertation, in the most legal, restrictive sense. This work is thus mine; I assert it is original, and I am the only one to blame.

1.1 Cognitive framework

The cognitive study of simultaneous interpreting has traditionally focused on memory, and the use of glossaries in interpreting may also be approached, in terms of *extended cognition* (Clark & Chalmers, 1998; Sprevak, 2019), as resorting to an external memory to support other cognitive functions. Cognitive psychology and psycholinguistics entertain over 20 types of memories, understood from distinct perspectives. Systematic research on memory truly began in the 1960s, with Atkinson & Shiffrin's (1968) model, which included a sensory register and two stores, one for short-term remembering, and the other one for long-term memories. Since then, the *short memory* or *working memory* (WM) store sparked debates on its nature and functions and stimulated further models.

Baddeley & Hitch (1974) proposed a multicomponent model of WM, incorporating elements for processing visual and auditory information. Baddeley (2000) enlarged this model to include an episodic buffer. This new component aimed to provide a more comprehensive, modular explanation of WM processes. Baddeley & Hitch's model has undergone further significant changes, evolving to be more situated and multimodal (Muñoz & Tiselius, in press). The central executive has been replaced by Norman & Shallice's (1986) supervisory attention system. The episodic buffer has become a slave store linked to long-term memory. The visuospatial sketchpad now processes haptic feedback, and the phonological loop now also deals with sign and lip reading, as well as environmental sounds (Yao, 2021).

In contrast to this multi-component framework, Cowan (1999) proposed a simpler model of WM. In his view, memory is a single repository for storing and manipulating information. In order to be used, information needs to be activated, and this can happen to various degrees. The information activated in your (long-term) memory is your "working memory". Here, WM is like a *pan* where various types of information can be placed and manipulated. Cowan's model does not negate the contributions of previous models (Cowan, 2017), but rather offers a more consolidated explanation, suggesting a singular, flexible memory system capable of various manipulations. Cowan's WM model is not only noteworthy for its simple, *Occamian* approach, but also for its emphasis on the role of attention, a factor that previous models may not have sufficiently addressed. This aspect is crucial because, as defined by Merriam-Webster dictionary, *attention* is 'the act or state of applying the mind to something'. This is intimately linked with our understanding of memory within cognitive translatology (see Muñoz, 2010, 2023; Muñoz & González, 2021), a situated approach that we justify below.

In Cowan's framework, memory is seen as supporting action through steering attention. For instance, when trying to remember the password of a website, we may allocate more memory resources to this focus of interest by focusing our attention on it. Rehearsing the symbol string in memory becomes a task by itself, which shows the practical and realistic workings of this model. Cowan's approach is particularly relevant in our study of interpreters' behaviors when using CAI tools (see also Mizuno, 2005). Cowan's emphasis on the interplay between memory and attention provides a robust framework for explaining the cognitive processes our informants engage in during RSI tasks with the support of CAI tools. In adopting this model, we move away from the traditional focus in interpreting research on the stable capacities of interpreters and even on the purported innate faculties of their cognitive systems and into the changing patterns of cognitive resource management and allocation, the way interpreters steer their attention while multitasking and regulate their mental efforts while interacting with the environment.

In other words, multicomponent WM models are not considered so important in the present study, even though we acknowledge that WM is as crucial as portrayed in other research approaches. We adopt, in brief, a situated approach that is distinctively human-oriented. Departing from this point, we discuss the cognitive demand (input) and cognitive effort (output) in the interaction between human, environment, and stimuli. Cognitive demand refers to the tendency of environmental stimuli and tasks to prompt different degrees of cognitive effort that can be linked to goal-oriented, adaptive behavior. Demands may be measured by focusing on relevant independent variables, such as sentence length, vocabulary frequency, term density, and speech delivery rate. Cognitive effort refers to mentally investing higher amounts of metaphorical mental energy to carry out a task or handle several tasks simultaneously. Cognitive effort is an adaptive response that is affected by demands; it is usually measured through physiological indicators, such as pupil dilation and heart rate variability, and behavioral indicators, such as the frequency and length of fillers and pauses. Cognitive demands should reflect the general tendency, whereas cognitive effort is the actual individual and changing response involving the allocation and management of attentional resources to meet those demands (Muñoz, personal communication).

1.2 Situating the human mind

Comprising approximately 100 billion neurons (Herculano-Houzel, 2009), the human brain exhibits a rich interconnectivity. Neurons are dynamically activated, and continually altered by experiences and environmental factors. Traveling waves of electrical activity are ubiquitous in neural networks, regardless of whether one is awake or asleep (Erazo-Toscano & Osan, 2023). Reading this text or listening to speech activates neurons and sparks interactions and connections of relevant information in your mind that draw from prior experience and build meaning for the incoming stimuli. Neuron activation spreads like a fire, wave-like activity (Foster & Scheinost, 2024), and activating one piece of information can activate the next, and the next, and perhaps an area that we may consider a self-contained whole (e.g., a term, a concept, a memory). So re-reading these words will strengthen the neural connections that make those patterns and increase the depth of your learning (Sousa, 2022). Artificial neural networks (ANNs) only vaguely resemble the real ones. There is a small (but growing) number of ANNs but more than 200 biological natural networks. ANNs are organized in layers, usually up to 20, whereas biological neural networks may have thousands of them, since it can rather be seen as a biological, changing, multidimensional mesh. Connections in biological networks are more *across* than *within* neural networks, so that the whole brain can be described as a single neural network. Furthermore, biological networks have a refractory period when they cannot *fire* (immediately after having done so) that does not exist in ANNs that, in turn, need back-propagation to learn, which is unnecessary in neurons. The brain is plastic, and neurons grow and die, and change their connections, whereas ANNs can adjust but remain basically the same.

There are other differences between human cognition and digital processing. However, we cannot cover them all. For instance, in computers, transmission delays are often due to the response time of transistors, which switch within 5 nanoseconds. In contrast, human neuron responses typically take 1–5 milliseconds (ms) due to biochemical processes and synaptic mechanisms (Groh & Gazzaniga, 2003). Compared to machines, brain reactions are slower. Yet human brains predict the next stimulus to come and get ready for it by adjusting the activated information in memory, as priming effects have consistently shown (Cowan, 1988; Altarriba & Basnight-Brown, 2007; Chmiel, 2018). Computers do not do that. To find the appropriate term, they often need to look for it through the whole stored information. Thus, despite computers performing faster than humans at first glance, the fundamental nature of linear information processes in computers remains unchanged.

Humans engage with their environment through sensory receptors: sight, hearing, smell, touch, and taste. The human brain activates a widespread system of neurons that appear to integrate sensory information with states of sensory inputs and behavioral responses (Lovallo, 2016). Each sense plays a distinct role in our perception and interpretation of stimuli. For instance, the eyes process visual inputs, such as observing slides in a conference or reading text on a screen. Ears and kinesthetic awareness detect auditory cues and the navigation of spaces, crucial for tasks like using digital devices in conference interpreting. Computers also lack the ability to process and integrate sensory information holistically (Doherty, 2020).

Multisensory integration enables us to coordinate sensory inputs, tailoring our responses to the specific demands of our attention, especially in multimodal information processing (Muñoz, personal communication). For instance, interpreters may simultaneously watch a presenter's slides and listen to their spoken words through headphones. In such scenarios, interpreters might not notice a colleague passing a note in the booth, not because they cannot do it, but due to their attention being focused on a different source (or, rather, sources) of information (e.g., text in the slides, and auditory signal of speeches). Our brain dynamically filters certain sensory inputs over others, depending on situational needs in a specific scenario. This phenomenon is a key aspect of situated cognition.

Situated cognition approaches cognitive processes with a special focus on interacting with the environment in realistic social situations (Robbins & Aydede, 2009), emphasizing the dynamic brain activity also during problem-solving (Anderson, 2007). It aims to "...understand the development and constitution of cognitive systems in changing environments or real-world situations" (Krickel, 2023, p. 4). Situated cognition asserts that cognitive states are far too ephemeral and complex to be conceived of as stable entities. They are rather processes, and these mental processes are not brain-internal, self-contained activities isolated from the rest of the body and the environment. Cognitive processes unfold as we interact with the physical or social environment. In other words, cognition can be external to the brain, in that it results from actions involving tools, digital devices, and social interactions (Krickel, 2023).

By extending cognition beyond the brain boundary and considering interactions with the environment, new possibilities for thinking and action emerge. A fundamental characteristic of situated cognition is that both agents and environments shape human interaction, there is no divide between internal and external factors. For instance, when we consult a term in a glossary, we depart from an "internal" need generated by an "external" input (the source speech). We may "internally" seek in our memory and then "externally" in a glossary, but we do so with the support of "internal" processes such as reading and assessing the information. Once we "internally" choose a certain rendering from the "external" palette of options, we process it into the "internal" action plan and then we utter it, turning it into an "external" input for other parties to process, whose reaction we often assess "internally" as feedback hints. This is the dynamic interaction with the environment that situated approaches such as cognitive translatology (see Muñoz, 2010, 2023; Muñoz & González, 2021) are mainly interested in.

Heersmink (2015) proposes three types of information flow in situated cognitive systems: (1) One-way information flow from artifact to agent, (e.g., websites and dictionaries), where the agent typically does not influence the informational content. (2) Two-way information flow, from agent to artifact and back. This is common when humans offload information onto their environment to ease memory burdens, creating cognitive artifacts like notetaking for interpreters. (3) Reciprocal information flow, where cognitive artifacts are integral to ongoing information-processing tasks. For instance, during SI tasks, term retrieval using CAI tools and checking screen information involves multiple subtasks. Typing letters into the CAI tool to look for a glossary entry, and the results from the CAI tool may influence the interpreter's subsequent rendering. This process, whereby interpreters may tend to offload part of their cognitive demands onto the CAI tool, may shape later steps in the interpreting process.

In situated cognitive systems, multiple agents or artifacts can interact within the environment (Heersmink, 2015). For instance, interpreters listen to speakers, speak out the rendering via a remote interpreting platform, and check translations on a CAI tool in another device. In remote interpreting, interpreters may collaborate with a boothmate on a single task, taking turns and utilizing different CAI tools, exemplifying distributed cognition where a collective of agents performs a task using cognitive artifacts. For instance, in RSI, interpreters engage in complex cognitive processes. They listen to speakers and produce rendering in a target language, often through a remote interpreting platform. Additionally, they may consult various CAI tools on separate devices for translation search. This setup exemplifies interpreters being embedded in a context that involves interacting with multiple artifacts. Taking it a step further, remote interpreting scenarios often involve collaboration with a boothmate via a remote interpreting platform. Interpreters work together on a single interpreting task, engaging in turn-taking and potentially employing different CAI tools. This situation illustrates another instance of distributed cognition, where a collective of agents collaborates to solve a specific problem or perform a cognitive task. The use of diverse cognitive artifacts in this process reflects the dynamic and distributed nature of cognition in RSI.

1.3 Language and communication

Language has a double nature. On the one hand, it structures and mediates but does not necessarily constrain the workings of our minds. On the other hand, it is part of a complex, interactive, imperfect system of symbolic communication (Brice, 2021). Language is imperfect because there are as many versions of language as there are speakers, organized in language families, dialects, sociolects, registers, uses, and the like. That is why interpreters have always been the companions of progress and civilization. In ancient China, 舌人 'tongue man' referred to official interpreters during the Zhou dynasty (1046–771 BC). They were expected to be well-versed in foreign languages and familiar with different places (Lung, 2005). Their early role highlights the longstanding human effort to overcome language barriers.

Language is dynamic; it evolves in interaction, leading to coining new words and forgetting others into the annals of history due to lack of use. This fluid nature of languages emerges through interaction. Language is not confined to combinations of symbols from a closed set. The material forms these symbols adopt when used entail speech, intonation, pace, body language, gestures, facial expressions, references to visual elements in the environment, etc. Tools like smartphones, the Internet, and large language models (LLMs) are revolutionizing human interaction. This diversification has broadened the palette of our interactions, as digitalization impacts nearly all aspects of everyday life. For interpreters, "augmented reality" applications exemplify this trend (Gieshoff & Schuler, 2022). We can also use natural language to interact with machines, such as for prompting image generation in tools like Midjourney and Stable Diffusion.[1] This showcases the evolving landscape of language interaction and its potential future developments.

Meaning is the business of language. Muñoz & Rojo (2018) argue that it results from a cognitive constructive process, which is not directly transferrable, and is rooted within the human mind. Objects like books, tables, and printed words do not inherently possess meaning; instead, they are interpreted thanks to accumulated life experience. For instance, consider a three-year-old child encountering a ball for the first time. The child initially lacks knowledge about it. However, if the parents consistently refer to it as 球 (*qiú,* 'ball') in Chinese, the child gradually forms a mental representation of 球 through repeated exposure to the word, alongside visual and auditory stimuli. Eventually, the child responds to the auditory stimulus of 球 by gazing at or reaching for a ball, illustrating the interplay between language and accumulated experiences in information processing. Meaning is thus not merely individual, but it is also a social construct that we interiorize. Our mental experience is richer than what we codify. That is one of the main reasons that make communication possible at all (Muñoz & Rojo, 2018).

The reading process prompts an interaction between newly acquired knowledge and existing knowledge, influenced by experience. That is, reading involves cognitive processes to recognize and comprehend stimuli by aligning our perceptions with our stored knowledge (Levering & Kurtz, 2019). During this process, readers tend to develop a model that we can simplify as consisting of three levels of mental elaboration in a (narrative) text (Dijk & Kintsch, 1983; Kintsch, 1998): the *text surface,* the *textbase,* and the *situation model.*

Apprehending the text surface enables conceptualization beyond immediate sensory experience. In this model, readers develop the text surface, which symbolically reproduces verbatim text information. It is a superficial engagement with the text, focusing on word-to-word information without deep comprehension (Wannagat *et al.*, 2022). It may be likened to entering the stimulus into the short-term memory. The second level, the textbase, involves readers grasping the ideas described in the text as a progressively aggregated whole, rather than a string of specific words or concepts (Zwaan & Rapp, 2006). It is a construction process that still does not adopt any particular perspective and can be likened to the immediate product resulting from processing a text structure in working memory. The highest level of representation, the situation model, involves constructing a mental

[1] Midjourney: https://www.midjourney.com/
Stable Diffusion: https://stability.ai/

representation of the situation described in the text, activating knowledge that extends beyond the explicitly stated content (Kintsch, 1998). Readers often rely on their prior knowledge to fill gaps in the text. This is when we *make sense* of what we read. That is, we establish links with the rest of the information we have in our long-term memory.

The situation model is central to comprehension (Zwaan & Rapp, 2006) and is thought to govern both production and discourse comprehension (Kintsch & van Dijk, 1978; Morales *et al.*, 2022). Generally, when people comprehend, they construct mental models by integrating the incoming information with their stored knowledge (Morales *et al.*, 2022). Kintsch's model involves more than just constructing the mental representation of the text itself. It also involves comprehending and constructing the mental representation of what the text is about, integrating the information derived from the text in a step-by-step bottom-up manner. During the process of constructing the mental representation, the model also heavily relies on inference generation in language comprehension (Davoudi & Moghadam, 2015). Inference making is an essential process for language comprehension, as it allows the generation of information that is not explicitly stated in the text (Cain & Oakhill, 1999). Readers, therefore, are not passive receivers of information; instead, they play an active role in making inferences, in predicting what comes next.

Figure 1. A representation of cascade processing.

The Cascade Model of discourse comprehension, as proposed by McClelland (1979), illustrates the scheme of discourse information processing at a macro-level. The concept of a cascade can be likened to that of a waterfall, as depicted in **Figure 1** after Oakes & Rakiso's (2019, p. 103) representation.

In this analogy, information flow is likened to water cascading down a waterfall, starting at the top and traveling downwards, taking various paths influenced by intervening elements. Ultimately, all streams converge in a pool at the bottom, symbolizing the culmination of the information processing journey (Harley, 2014; Oakes & Rakison, 2019). So, discourse information processing is complex in that it entails parallel threads, and not a linear, successive, straightforward process. The top level of processing influences tasks that require a high degree of information integration (e.g., inference and prediction, formerly labeled *anticipation* as well), like taking different paths among stones in the waterfall.

More than a simple hierarchical structure, the Cascade Model highlights the dynamic nature of language use (both with successive and overlapping processes). In this model, information flows continuously, allowing each processing stage to activate the subsequent stage without necessarily waiting for the completion of the previous one. That is, there are overlaps in the processing stages (McClelland, 1979; Harley, 2014). For instance, cascaded models suggest that phonological word forms can activate before lexical selection is finalized (McClelland, 1979). This means that as a speaker is preparing to say a word, the sounds of that word (phonology) begin to get ready even before the speaker has fully selected the exact word she intends to use (see Morsella & Miozzo, 2002; Navarrete & Costa, 2005; Goldrick & Blumstein, 2006; and also Q. Zhang *et al.*, 2018; Bao *et al.*, 2023 on Chinese spoken word production).

In psychology and human behavior analysis, the Cascade Model is broadly applied to emphasize parallel, looping mental processes, spanning from comprehension to production. Díaz (2011, 2020) examined SI preparation as related to situation models. The cascade analogy can be a useful way to conceptualize information processing. In the picture of a cascade, you can imagine each flat or level represents a different processing stage, with information (like water) flowing through and undergoing transformations simultaneously. However, the cascade analogy only captures certain aspects of processing. For example, bottom-up processes, where lower-level features are integrated into higher-level representations, are not well represented in this analogy. Readers of this dissertation should be aware that the cascade analogy is limited in its ability to depict all facets of information processing. It primarily illustrates how processing branches out and can occur at different levels concurrently but does not fully encapsulate other mechanisms like bottom-up integration. The analogy should be treated as a partial metaphor rather than a comprehensive model.

1.4 Interpreting

Interpreters read many documents, a time-consuming task that entails building knowledge structures in their minds (Moss & Schunn, 2015). Faced with time-sensitive preparation tasks, interpreters often experience stress and anxiety due to scarce domain knowledge, compared to that of the speakers. Interpreters may immerse themselves in materials related to the coming speech they will interpret, from domain knowledge to conference specifics. This reading process, often time-consuming, does not merely cram background information into the interpreter's mind. Instead, it activates and integrates relevant knowledge, enabling interpreters to align their understanding closely with the speech content.

Preparation materials may include, e.g., previous conference proceedings, colleagues' glossaries, and minutes from earlier meetings. Additionally, interpreters often receive supplementary materials such as PowerPoint slides, speech drafts, or other documents prior to the conference (Díaz, 2011; Jiang, 2013). However, most non-analogical resources (e.g., manuscripts, images, and audio recordings) are often beyond the reach of CAI tools' digital search capabilities. During the preparation phase/process, interpreters can construct their understanding of the source texts and activate stored information including that from experience. Moreover, the process also widens and deepens the interpreters' familiarity with the topic, and this, in turn, impacts interpreting quality, information completeness, delivery fluency, etc (Chen *et al.*, 2021).

Despite the benefits of the preparation phase, interpreters often face challenges in this process. The comprehension process may be influenced by factors such as task requirements (Díaz & Torres, 2019) and individual differences, including time constraints, the volume of documents, unfamiliar terms, and pertinent specialized terms. To face these challenges, a personalized glossary can prove to be invaluable. This glossary aids not only in understanding the documents but also in preparing for the potential linguistic structures and terminology in the forthcoming speech, especially with challenging specialized terms.

The connection between experience and the mental representation of source texts is relevant for both novice and experienced interpreters (Ho 2021) because this kind of information has usually undergone multisensory (multimodal) integration. Discourse comprehension is of course not limited to the stage of preparing for interpreting, but it expands into the interpreting process, where a high level of multitasking is required. We humans do not necessarily excel at, or get totally used to multitasking, but interpreting, particularly SI, is widely recognized as a complex cognitive task where multitasking is a paramount skill.

Mellinger & Hanson (2019) show that interpreters are better than non-interpreters at tasks requiring them to remember and process what one hears and sees. Interpreters tend to be skilled at handling both auditory (hearing) and visual (seeing) information (Ghiselli, 2022). This is important because interpreters sometimes have to work with both spoken language and written texts. Response times are claimed to be a significant factor in interpreting performance in these tasks. Furthermore, the response duration in WM tasks has been shown to correlate with the prediction of academic skills and achievements. For instance, in SI, interpreters listen to the source language (L2) and rapidly switch tasks to produce speech in the target language (L1) while selectively inhibiting specific elements of the L2 content mapping with their renditions. This selective inhibition is a sophisticated cognitive feature of task switching. Moving from one task to the next, people shift their focus of attention and adjust their actions by bringing up relevant information and changing their behavior accordingly (Dong & Li 2020; Muñoz & Apfelthaler 2021).

Compared to monolingual comprehension, simultaneous interpreting is an extreme case of language processing (Obler, 2012). It requires detailed control, monitoring, and access to both receptive (listening) and productive (speaking) processes in two languages simultaneously. It is a non-natural cognitive activity that requires extensive training and improvement through experience.

Cowan's WM model (1999, 2000) emphasizes the role of attention in relation to WM capacity, and Cowan *et al.* (2003) highlight the critical role of attention-switching in reading and listening span tasks. The focus of attention refers to "… subjective or phenomenological idea of the information of which one is aware at a given moment" (2005, p. 7). Interpreters are required to rapidly shift their focus between listening and speaking, a cognitive process facilitated by the ability to concentrate on one task at a time. They can do this because the phonological information stays in their memory for a short time, allowing them to recall and interpret it even after they have heard it (Cowan, 2000).

Undoubtedly, memory, attention, and multitasking control are inextricably linked in interpreters' minds, which hold various types of information simultaneously in *chunks*. That is, in "[...] stored units formed from integrating smaller pieces of information" (Eysenck & Keane, 2020, p. 804). Chunking actually refers to two complementary, simultaneous, interacting but opposed processes: bottom-up and top-down chunking. Bottom-up chunking involves combining smaller bits of information as represented into larger ones. Top-down chunking involves dividing the input flow into smaller units or chunks. Chunking allows us to break down complex information into manageable pieces and also to reconstruct it into coherent wholes. When applied to SI, chunking refers to the process by which interpreters "[…] segment the input into smaller fragments that can be encoded without having to wait for the entire sentence to unfold" (Seeber, 2011, p. 194). Cowan argued that separate items or chunks can be combined into a single, larger chunk only if they can be present in the focus of attention at the same time (Cowan, 2000, 2001). Chunks may thus include sentence-based information, names, entities, numbers, or unfamiliar terms.

As for *remote interpreting*, it involves interpreting a speaker from a location different from that of the interpreter, facilitated by information and communications technology, as described by ISO 18841:2018 (en) *Interpreting services—General requirements and recommendations.* Remote interpreting is closely associated with processing information with the support of technology. The technologies employed in remote interpreting may include devices such as videophones, web cameras, and computer screens (Corpas, 2016). Virtual meetings are on the rise, promising substantial savings for interpreters in time, and for clients in cost (Annalisa, 2015).

Recent developments have directed research toward remote simultaneous interpreting (RSI), focusing on aspects such as interpreters' physiological stress (Bower 2015; Li *et al.* 2022; Olalla *et al.* 2023), technology literacy (Drechsel, 2019), automatic speech recognition (Fantinuoli *et al.*, 2022), CAI tool design (Corpas, 2022), user perspectives (Gilbert *et al.*, 2022; Frittella, 2023), note-taking effort (Kuang & Zheng, 2023), and live captioning (Yuan & Wang, 2023). These areas of RSI research are all new compared to on-site SI research, but the focus increasingly gravitates toward the capabilities of the technology used (Corpas, 2016), and human-computer interaction (Salaets & Brône, 2020). Such shift may "[…] alter the interpreting task and paradigm" (Mellinger & Hanson, 2018, p. 369), and it probably entails changes in the cognitive demands placed on interpreters.

The use of appropriate terminology is of utmost importance in facilitating the delivery of high-quality interpretation services (Xu, 2018; Prandi, 2023). In preparing for interpreting tasks, interpreters may opt for using technology to compile glossaries. Tools for interpreters to prepare, store, and share their glossaries have proliferated (Riccardi *et al.* 2020). Some CAI tools or online services argue that interpreters can rely on glossaries automatically compiled through their built-in specific algorithms. As discussed, preparatory reading is both taxing in terms of time and cognitively demanding. However, glossaries built by reading and extracting terms personally entail reading comprehension, which is also beneficial in constructing a text-based understanding and furthering domain knowledge. Building a *read-first* glossary (see **§ 1.5.3**) not only provides close domain knowledge about the upcoming tasks but also activates prior knowledge from experience.

During RSI, interpreters need to navigate a complex work environment with numerous variables that may influence their behavior. For instance, as speech signals unfold rapidly, interpreters must allocate cognitive resources effectively and maintain selective attention to process multimodal information, such as auditory signals and graphic text from monitors, and integrate them to produce a coherent target output. The cognitive demands in such a situated environment are dynamic and constantly evolving within the interpreting tasks. However, research is scarce that explores how CAI tools influence the cognitive processes and demands within RSI.

1.5 Computer-assisted interpreting tools

Human-computer interaction (HCI) investigates how technology supports cognitive processes, which is relevant for understanding how computer-assisted interpreting (CAI) tools can enhance interpreters' performance. CAI tools, such as InterpretBank, offer features like term extraction, automatic speech recognition, and term retrieval. Research has arguably shown that these tools can improve term accuracy and reduce errors and omissions in interpreting (see **§ 1.5.3**). Even though InterpretBank is less well-known among Chinese interpreters and trainees.

Limited studies conducted in China have reported potential benefits of using InterpretBank in improving term accuracy (e.g., Ge, 2023).

1.5.1 Human-computer interaction

Human-computer Interaction (HCI) covers scenarios where a variety of technologies and human actions coexist and can even be thought of as a unified system. In order to effectively establish cognitive ecosystems with out-of-skull tools, we need a profound comprehension of the intricate cognitive interactions that underlie multilectal mediated processes of oral communication. HCI and social cognition are fundamental concepts in comprehending how technology enhances human cognition. Grinschgl & Neubauer (2022) emphasize that individuals often rely on technology to amplify their cognitive abilities.

HCI primarily focuses on the user experience. *Distributed cognition* (Hutchins, 1995; Hollan *et al.*, 2000), posits that cognitive processes are not confined to an individual, but rather distributed across individuals, artifacts, and their environment (but see Muñoz, 2023, pp. 131–140). This notion is especially pertinent to address HCI in interpreting, as it explores how technologies, such as computers and software applications, support cognitive processes. HCI research has expanded to consider how humans interact with computers in task execution, with an emphasis on the role of distributed cognition and also *extended* cognition. One popular aspect of extended cognition is cognitive *offloading,* whereby individuals or groups transfer cognitive tasks to their environment, often through the use of technology. In this process, people employ physical actions to modify the information processing demands of a task, thereby reducing cognitive effort (Morrison & Richmond, 2020; Grinschgl *et al.*, 2021).

In Cognitive Translatology, Muñoz (2010, 2023; Muñoz & González, 2021) suggests that the implications of distributed cognition become apparent when examining the work of translators and translator trainees, particularly in group settings. When employing computer-assisted translation, the cognitive processes involved in translation are facilitated and distributed among individuals as they collaborate, either implicitly or explicitly, with computers. To conclude, the interplay between HCI and social, extended, and distributed cognition underscores the role of technology in enhancing both individual and group-level cognitive processes.

1.5.2 Research on CAI tools

Interpreters have always used tools to support their performance, such as pen and paper. Technological advancements have made remote interpreting possible in ways that telephone interpreting could not have imagined. With the widespread accessibility of computers, it was only natural that interpreting tools would also be computerized. Computer-assisted interpreting tools (CAI tools) are all digital tools that support and aid interpreters in many ways. CAI tools have replaced traditional information resources, such as paper dictionaries with electronic versions, replaced communication channels with instant messaging and other digital sources of information, and even enabled remote consultation for information or interpreter collaboration. In time, applications specifically designed for interpreters have emerged. Hence, a second, narrower, or more specialized take on the notion of CAI tool restricts it to dedicated software packages designed for interpreting tasks. Hence, a second, narrower, or more specialized take on the notion of CAI tool restricts it to dedicated software packages designed for the tasks.

Some common features of CAI tools, in the narrow sense, are terminology management, speech recognition, note-taking, and audio/video recording, as well as virtual booths for collaborative interpreting. According to Fantinuoli (2023), the first generation of CAI tools simply replaced traditional tools for compiling or using glossaries. The second generation of tools takes advantage of features that only computers make possible, such as automatic glossary extraction and glossary memorization aids. The third generation of tools integrates artificial intelligence features (e.g., automatic speech recognition). Since documentation behavior is the main topic of our project, unless otherwise specified, the label *CAI tool* here will mainly refer to (terminology management features in) a CAI tool package.

When evaluating the impact of a specific tool on interpreters' performance, researchers tend to focus on results, e.g., SI delivery. Xu's study evaluated the effectiveness of corpus-based tools in improving SI term accuracy for 22 Chinese interpreting trainees from a UK university, who were divided into a test group using corpus tools (i.e., Syllabs Tools and Sketch Engine) and a control group employing traditional methods by manually extracting terms (Xu, 2018). Xu found that the test group exhibited significantly higher terminological accuracy and fewer omissions, compared to the control group. Participants using corpus tools also reported better term recall. However, Xu's observation that corpus-based tools required less preparation time needs further exploration. It only considered limited aspects (e.g., terminological accuracy), and there was no discussion on term extraction results. Furthermore, an interpreter's experience in preparing a glossary might impact its retrieval during interpretation tasks. Xu's study did not explore the potential relationship between glossary compilation and interpreting quality.

Pérez's (2018) also examined the impact of using a corpus management program on vocabulary preparation for interpreting assignments. The study, conducted with 27 final-year Translation and Interpreting students at the University of Málaga, aimed to determine whether using a corpus management program would positively influence interpreting outcomes. Participants were divided into two groups: Group 1 prepared vocabulary without a corpus management program, while Group 2 used a corpus management tool (either AntConc or WordSmith). Both groups used tools to extract relevant vocabulary and terminology from input documents. Students in Group 2 were found to consistently perform better across various topics, as evidenced by a higher percentage of matches (terms correctly interpreted) in their renderings.

CAI tools' advanced functions—such as automated term extraction and automatic speech recognition paired with automated retrieval of pre-set glossary entries—have been shown to benefit users in digital working environments (Defrancq & Fantinuoli, 2021; Frittella & Rodríguez, 2022; Tammasrisawat &

Rangponsumrit, 2023). In these studies, term accuracy is widely discussed for evaluating interpreters' performance when working with different CAI tools, but few studies have explored aspects of documentation for RSI preparation from a cognitive perspective. In fact, documentation uses multimodal information resources, which require interpreters' demanding cognitive activities: integrating target or rich content knowledge from different information resources and maintaining working memory contents to remember some terms to comprehend the domain knowledge.

Both integrating contents and keeping them active in memory contribute to cognitive difficulties, particularly for inexperienced interpreting trainees. They have to complete the RSI preparation and proceed with RSI tasks with limited domain-specific knowledge from glossary preparation. Their rendering quality may not remain constant but evolves and changes as they progress through the task. Meanwhile, the choice of digital assistance tools could possibly draw their attention to the tools' operation and usage, rather than the essential elements of novel knowledge. Under these circumstances, if users approve of the CAI's function model, they could more effectively select novel terms, and organize, and integrate term retrieval with the CAI tools into the RSI workflow. Based on this notion, we raised the fifth hypothesis, which merges the first four, as will be seen in **§ 1.7**: considering the aforementioned scenarios, those who perform efficiently in documentation with InterpretBank, have a good quality of RSI rendering, and show individual attitudes of liking InterpretBank, would continue using InterpretBank.

The above considerations regarding a human-computer interaction approach to preparing for an RSI task materialize in aspects such as that the testing scenario is remote interpreting and that informants had to split their attention toward different input and output information resources like typing letters on the keyboard, keeping eyes on the screen movements which may locate correct terms in the master glossary, acoustic signal from the source speech recording, and monitoring their own delivery. The integration of CAI tools into the RSI process is already taking place in many real-world scenarios, with PCs or tablets and other emerging technologies of remote interpreting blurring the line between the booth task and information-seeking platforms (see, e.g., Corpas, 2021). Human-computer interaction may scaffold the exploratory discussion on the cognitive benefits that CAI tools provide for interpreting trainees.

Prandi's (2023) study compared the process and performance of simultaneous interpreting using traditional digital glossaries, CAI tools, and CAI tools with integrated automatic speech recognition (ASR), involving nine advanced interpreting students. The experiment collected data through performance measures, behavioral measures, and a questionnaire, using speeches with inserted terminology and an ASR-CAI mock-up simulating perfect ASR performance. Results showed that terminological accuracy was highest with the ASR-CAI tool, followed by the CAI tool and the PDF glossary, with fewer severe errors and omissions occurring when using the ASR-CAI tool. Behavioral measures, such as ear-voice span and fixation durations, also improved with the ASR-CAI tool, which participants rated as most useful and least distracting, despite its inability to look up terms not automatically recognized.

Fantinuoli *et al.* (2022) evaluated KUDO Interpreter Assist Tool (actually, the performance of informants' use of the tool) using datasets of glossaries and speech audio files. They assessed term extraction quality as rated by professional interpreters and real-time suggestion performance through precision/recall against expected results. While target-term accuracy was rated well, term relevance judgments varied. The tool performed strongly for specialized terms and numerals after ASR fine-tuning but recall of general speeches was not so full of promise. To the best of my knowledge, this might be the first time that recall has been raised in relation to the CAI tool performance assessment.

Previous research failed to check whether correctly or wrongly rendered terms had actually been searched. Human memory may store information according to salience, and also according to its relevance for the person who remembers. Hence, terms that have been entered in the glossary because they seem relevant or particularly salient may later be remembered without the interpreters resorting to the glossary. This might distort or raise doubts regarding the CAI tool's real support. Furthermore, prior CAI tool assessments did not consider the powerful impact of repetition. The interpreters' likelihood of consulting the terms in the glossary very probably becomes less and less pressing with more and more repetitions of a given term. Any task evolves over time, and performance varies from the onset to later stages. In brief, those assessments did not envision glossary use as situated activities building on prior experience and co-occurrence dynamics.

To address these gaps, the present study focused both on glossary compilation and glossary consultation, and their interrelationships. We adopted some strategies and decisions from Prandi's (2017) study of InterpretBank. She had informants interpret three short speeches of similar length and lexical density, using InterpretBank, Excel, and Word glossaries. Prandi's source speeches were designed with a fixed internal structure, containing 36 terms each (18 at the end of sentences, 18 in the middle; 12 unigrams, 12 bigrams, 12 trigrams). Half of the terms were expected to require a glossary search. We also drew inspiration from Frittella (2022) evaluation of the SmarTerp, another CAI tool.[2]

Building on these studies, the present research project set out to study the benefits of using a terminology management tool for interpreting, from complementary or alternative perspectives. Atabekova *et al.* (2018) surveyed the tools the interpreters use to manage their multilingual glossaries and found that 63% of the respondents mentioned using InterpretBank for their professional activities. The tool is popular due to its versatility and the various modes it provides for managing glossaries and using them during interpreting sessions. Hence, our choice fell on InterpretBank.

[2] SmarTerp: https://www.eitdigital.eu/fileadmin/2021/innovation-factory/new/digital-tech/EIT-Digital-Factsheet-Smarterp.pdf

1.5.3 InterpretBank

InterpretBank is a 3rd-generation CAI tool developed by Dr Claudio Fantinuoli.[3] It offers two types of term extraction: *automated* and *manual*. In the automated term extraction mode, InterpretBank autonomously identifies and extracts terms from the source text. Manual term extraction involves reading the source text, selecting terms, and entering them into InterpretBank's records. We will call this procedure *read-first glossary compilation* since *manual* is a misleading label, however popular. Read-first term extraction with InterpretBank is quite similar to doing it with a text processor and using an Excel spreadsheet to compile the glossary.

We refer to manual extraction as *read-first*, to highlight that it entails reading the text to choose the entries for the glossary. This may happen in different ways. Compilers may read a bit as a first step and then stop to select terms, or they may read the whole text or a good section thereof, and then go back to the text they read to select terms. The actual procedure may vary, but reading the text is always a part of it. Such *read-first* glossary compilation might have a differential impact on the memory of the compiler, compared to automatic extraction, which will very likely not provide the same level of understanding. In this context, *read-first* glossary compilation facilitates more contextual understanding and it can help compilers better comprehend texts and build not just the *textbase* (not just passively receiving information at the language level), but also the *situation model* (activating and connecting the novel information with the compilers' prior knowledge in their memories) for the glossary compiler (cf. Kintsch, 1998, 2013).

Within *read-first* term extraction, there are at least two approaches. One approach is to identify an unfamiliar term, stop reading, search for its translation, enter it, and then return to reading. Let us call this the read-first *looping* approach. The other approach is to identify unfamiliar terms, transfer them into a spreadsheet, and continue reading. Let us call this one the read-first *straight-through* approach. The read-first looping and straight-through approaches represent two activities for dealing with novel information. If read-first extraction is conducted without checking definitions while reading and simply selecting terms, these terms become isolated words, disconnected from each other, and unlikely to be memorized as easily as with the looping approach.

If users work on one term at a time, stop reading the source text to check definitions, and then return to it, they may benefit in two ways. One benefit involves checking translations, which requires maintaining items active in mind for searching, typing, or validating translations. This process probably fosters the activation of prior knowledge. The specific way of encoding memory traces is unknown, but search activities are known to enhance the memorization of novel terms (Spink *et al.*, 2002; Rosman *et al.*, 2016). Another benefit is that, with prior exposure to definitions, readers allocate less attention to new words during the reading process, as noted by Elgort *et al.* (2023). This approach thus aids in the reading of the source text, particularly in reducing the mental demands placed by lengthy texts.

[3] InterpretBank: https://www.interpretbank.com

Recently, InterpretBank has garnered interest due to its purported impact on interpreters' performance from various perspectives. Guo *et al.* (2022) reviewed 27 empirical studies on CAI tools and found that six utilized InterpretBank, making it the most studied CAI tool in interpreting research. InterpretBank's booth support has been explored on European users. Despite its popularity in Europe, where InterpretBank is "the tool students are most often introduced to, followed by Interplex and Interpreter's Help" (Prandi, 2020, p. 4), it is less known among Chinese interpreters and trainers. Wan & Yuan (2022) surveyed interpreting trainees in China and found that InterpretBank was not even part of their toolkit, whose most frequently used CAI tool was online dictionaries. Liu (2022)'s discussion on InterpretBank's limitations, such as financial implications, platform dependency, and learning curve, echoes these findings. The financial burden of a commercial license for InterpretBank is also highlighted by Costa *et al.* (2017).

Defrancq & Fantinuoli (2021) compared first-time users' English to Dutch performance with and without InterpretBank's support for numbers (through automatic speech recognition) and explored their interaction with the application. They found that InterpretBank was very precise (96%), and its latency was low enough to offer interpreters a transcript before they had to deliver the number. The application generally met ergonomic requirements, and the participants were moderately satisfied with the system's usability. The participants' interactions with ASR support were varied, and they consulted the renditions in just over half of the cases. ASR use improved performance, in that the share of complete renditions increased in most cases and for almost all types of numbers. Defrancq & Fantinuoli (2021) concluded that, despite limitations and challenges, their informants appreciated the ASR support system, especially as a safety net. The system displayed numbers with an average latency below the interpreters' average ear-voice span (EVS), so it did not interfere with the interpreting process. However, the individual analysis yielded a significant beneficial effect only in two of the six informants.

Fantinuoli & Montecchio (2023) investigated the maximum system latency that is cognitively acceptable for interpreters performing SI. That is, the maximum EVS that interpreters can cope with while using InterpretBank without negatively impacting their performance. They found that interpreters can cope with a system latency of three seconds without any major impact on the rendition of the original text, both in terms of accuracy and fluency. These results may be influenced by the controlled testing environment, and the limited number of studies in this area suggests a need for further research to comprehensively understand user response times and reactions in real-world settings. Moreover, in their study, the conditions were not randomized: the participants interpreted with a latency of 1 second (s), then 2 s, and finally 3 s. Without prior practice with ASR support, there may have been a learning effect, albeit moderate.

Three empirical studies have been conducted in China, in master's theses—(Zhou, 2019; J. Zhang, 2021; Ge, 2023)—investigating InterpretBank's impact on trainee interpreters' performance. They all conclude that InterpretBank brings benefits to term accuracy. For instance, Ge (2023) used a pretest-posttest design and reported a 23.1% improvement in term efficiency and a 16.7% higher share of correct term renderings in eight MA interpreting trainees, as judged by four professional interpreters. Although these results should be interpreted with caution, they highlight the potential benefits of using InterpretBank. These findings raise the question of whether it should be used for and by Chinese interpreting trainees.

Tammasrisawat & Rangponsumrit (2023) studied four trainee interpreters who performed SI from English to Thai with and without InterpretBank's ASR support. With a glossary prepared by the researchers in advance, using InterpretBank led to a significant reduction in the rates of errors and omissions, improving the quality of term rendition. They also found that data visualization could significantly impact the interpreting process and interpreters' rendition quality. In cases when InterpretBank failed to provide any suggestions or when multiple suggestions were displayed at once on the interface, the interpreting process could potentially break down, and it caused interpreters to make mistakes in their renditions.

InterpretBank may be used without ASR. In this case, users need to enter searches themselves. Two options are available: regular, and fuzzy search. The regular search works like standard dictionary searches: alphabetically, from the beginning of any word in a term. That is, it may be found by typing a string of the second or third word in a complex, plurilexical term. Fuzzy search, in contrast, allows users to enter any correlative string of characters in a term to retrieve it, whether initial or not. Fuzzy search allows users to enter any correlative string of characters in a term to retrieve it, whether initial or not. Of course, keys pressed in either kind of search may be in a wrong order, so that the resulting characters in the search bar of InterpretBank do not match the correct character string stored in the glossary. Typos may happen anytime with or without corrections and users may or may not correct their input before they enter the last keypress. For instance, an informant in the present study typed *dopinemne* without any correction, and another informant typed *dopamine* without any correction. The final character strings shown in the search bar are *dopamine* and *dopinemne*. The retrieval results from InterpretBank are different in these cases because only *dopamine* is in the master glossary. *Dopinemne* cannot match any terms in the InterpretBank.

1.6 Quality evaluation of CAI-supported interpreting

The rapid development of CAI tools has witnessed the development of several performance evaluation approaches for various SI tasks. A noteworthy challenge associated with these advancements is the need to take in all factors that may impinge on results. That is, isolated, micro-analyses focused on accuracy measures tend to ignore the environment and important variables. Often, they may not scrutinize the implicit choices in their methods and even their notion of interpreting quality. Do CAI tool quality studies measure the performance of the tool or of the interpreter? Evaluating interpreters' performance tends to fall into two main traditions, variously portrayed as quantitative vs qualitative, or rubric-based vs holistic. The evolution of assessment methodologies in language testing has been significantly influenced by research on teacher assessment strategies. Although the assessment by trainers has been a long-standing concern in language testing, research on teacher assessment has shed light on two fundamental approaches, holistic and analytical, and their interaction (Sadler, 2009; Crisp, 2017; Phung & Michell, 2022).

Both have problems and advantages that cannot be ignored. Our approach was to perform a quantitative analysis of typical SI quality indicators (**§ 1.6.1**) but only for descriptive purposes. The quality of the informants' renderings was holistically assessed by raters other than the researcher (**§ 1.6.2**). Then both results were cross-referenced. In brief, **§ 1.6.1** and **§ 1.6.2** will dive into the literature on the evolving evaluation approaches and the key indicators of interest in this study.

1.6.1 Terminology accuracy

This section only refers to interpreters' rendering performance in the booth tasks. Terminological work requires domain-specific terminology and specialized phraseology (Corpas, 2022) but there is no gold standard to assess efficiency in terminology management in the booths. In the present study, *efficiency* is understood as a function of (assumed) invested cognitive effort and quality of the rendering, instead of more restricted interpretations of the notion that limit it to aspects such as time spent or delivery speed. In other words, we believe that the less effort interpreters exert, the better quality of their renderings—and, thus, the more efficient their behavior.

Prandi (2023) studied CAI tool users' performanceperformance under three conditions: with automatic speech recognition (sim-ASR), CAI (no-ASR), and traditional glossary consultation (with PDF) in simultaneous interpreting. She found that terminological accuracy was highest in the sim-ASR condition (96.3%), followed by no-ASR (86.26%), and the PDF condition (78%). Statistically significant differences were found between sim-ASR and PDF, and between sim-ASR and no-ASR, indicating that automatic term recognition in sim-ASR facilitated more accurate renditions compared to manual lookups. Fewer severe term errors and omissions were observed in the sim-ASR condition compared to the PDF glossary, with the CAI tool falling in between. Significantly fewer sentences were omitted in the sim-ASR condition compared to the PDF condition, suggesting that sim-ASR support reduced errors and omissions related to term searches more effectively than the PDF condition, with the CAI tool providing some benefits but not as many as sim-ASR. The actual consultations were not reported upon.

Frittella (2023) reported on the performance of another CAI tool, SmarTerp, and found the tool to be effective for simple, isolated items like acronyms and named entities. Interpreters achieved 90-100% accuracy, but accuracy dropped to 45-79% on more complex items with multiple co-occurring problems. Frittella focused on the cognitive ergonomics of the interface and did not address the possible costs of higher cognitive efforts when performing RSI with SmarTerp. Specific issues included pronunciation errors for rare named entities despite transcriptions, misunderstandings of numerical magnitudes like *trillion*, and omissions of contextual information not displayed by the tool. Hence, factors such as the size of chunks being interpreted, familiarity with the terms, and background knowledge may have played a role, and multitasking may have been more demanding.

Both InterpretBank and SmartTerp have been argued to support high accuracy in terminology rendering in SI tasks, but the studies might have limitations that were out of scope or that have been overlooked, including a focus on isolated informants' renditions, absence of process analysis of users' response to problem triggers, and potential challenges in effectively managing intricate terminology in real-time interpreting scenarios. For instance, Tarasenko & Amelina (2020) argue that the effectiveness of mastering terminology resources in CAI systems largely depends on prior knowledge. In brief, these quantitative indicators are relevant and informative, but they do not exhaust the available options for a comprehensive CAI tool evaluation regarding documentation. Furthermore, no matter how accurate SI rendering is, this is one aspect of quality, which is usually evaluated by addressing it in an implicit, intuitive, and holistic manner.

1.6.2 Evaluating the performance of CAI tool users

Quality assessment in interpreting may not raise as many doubts and eyebrows as it does in Machine Translation, but concerns about how close expert assessment is to the impressions end users have been legitimate. Blind and undirected evaluation refers to a holistic assessment approach where the raters evaluate the overall interpreting quality based on their intuition and general impression, rather than focusing on specific aspects or errors of interpreting delivery. This, in principle, avoids the risk of the assessment to include artifacts imposed by the researcher. This is why we chose this approach in this present study and, consequently, no explicit criteria were provided to the raters, and they were not required or encouraged to follow any explicit criteria were provided to the raters, and they were not required or encouraged to follow any particular rating guidelines. The only exception was in a rating trial conducted with the raters prior to the actual assessment, which involved some technical guidance on assessment procedures and answering raters' doubts and questions. This was carried out in the most neutral fashion possible to the best of our knowledge.

Chen *et al.* (2022, p. 16) found that the use of a holistic approach resulted in comparatively more reliable scores, irrespective of the interpreting instructions. Opinions are not unanimous, though. Han (2022b) discussed a scoring approach called *multi-methods scoring,* used by several testing agencies in the USA to assess interpreting performance. This method integrates two assessment techniques: itemized/atomistic analysis, where raters determine the accuracy of the test takers' rendition of specific words or phrases from the source language, and rating scale-based assessment, which employs a Likert-type scale to gauge the overall quality of the interpreting performance. Nevertheless, the absence of a well-established framework for integrating the scores from these two methods may pose challenges when trying to make accurate and reliable decisions based on the results. So far, there is no standard or golden rule for interpreting quality assessment. In this present study, we explored holistic assessment as well. By investigating this combination of assessment methods, we aimed to contribute to the development of more reliable and accurate evaluation methods for interpreting performance. After all, quality assessment plays a role, even if implicit, in many interpreting research projects adopting cognitive views. In our case, measuring the impact and support of a CAI tool implies a notion of quality and likely involves an approach to operationalizing that notion into an assessment method.

Comprehensive assessments play a pivotal role within a situated approach, i.e., in the present study. The raters' perception is dynamic and continuously evolving. They can perceive elements in the environment that are influenced by emotions, intonation, and accent of the speakers, among other factors. These explicit and implicit influencing factors may remain unchanged with alterations in the assessment criteria (Muñoz & Conde, 2007; Conde, 2009, 2012a, 2012b; Du & Muñoz, in preparation). Biases in human assessment may be due to differences between the characteristics of the raters and the testing environment, particularly when departing from idiosyncratic initial impressions. To address and minimize such biases, novel approaches to rater training can be implemented to enhance language testing skills and regulate factors that contribute to the variability of assessments (Lumley & McNamara, 1995). In this study, our strategy was to perform a holistic assessment, but to contrast it with an analytic approach (see **§ 2.5.5** and **§ 2.7.2**).

1.7 Research question and hypotheses

Readers of this dissertation will be familiar with research on information search and management in translation (e.g., Enríquez, 2013). In simultaneous interpreting, focused empirical studies on information search and management are more scarce but see, for instance, Will (2008) and Fantinuoli (2017). Discussions about practices largely remain in the general realm of interpreting preparation, and just a few empirical studies have tested the effectiveness of tool use by evaluating informants' SI rendering quality. These studies provide a point of departure and may sharpen our understanding of HCI processes. However, the relationship between glossary compilation with digital tools in SI tasks is relatively under-researched, leaving much room for further investigation into interpreting trainees' information-seeking behaviors.

The main goal of this project was to explore how InterpretBank impacts informants' performance from a cognitive perspective. The overall research question was:

is InterpretBank useful for Chinese-speaking
interpreting trainees before and during SI booth tasks?

The research question was unpacked into five hypotheses to scaffold the design of this exploratory study. Our hypotheses are necessarily wider than in confirmatory research. The discussion is organized around the stated hypotheses, which were the following:

H1: InterpretBank improves efficiency at glossary compilation
H2: InterpretBank improves the quality of RSI rendering
H3: InterpretBank improves efficiency when producing the RSI rendering
H4: Improved documentation performance will yield better RSI rendering quality (H3)
H5: Improvements using InterpretBank but also attitudes, will lead to keeping using it

These hypotheses form a sequence. This progressive exploration is like peeling an onion, helping us to gradually narrow down the research core by focusing on the influences of InterpretBank in each indicator for Chinese interpreting trainees. For a precise understanding of constructs such as documentation behavior, please refer to their operationalization (see **§ 3.1**).

1.8 Summary

This chapter delves into the evolving role and implications of CAI tools, with a special emphasis on InterpretBank, within the contexts of SI and RSI. It initiates with an overview of the advancement of CAI tools, particularly highlighting their expansion and increased relevance during and following the COVID-19 pandemic. This discussion points out the potential benefits these tools offer in both consecutive and simultaneous interpreting scenarios.

Central to this introduction is the cognitive framework (cognitive translatology) applied to study interpreting activities. We review the literature on various models of working memory (WM), including Baddeley & Hitch's multicomponent model and Cowan's single repository model, emphasizing their relevance in understanding the cognitive processes of interpreters during SI tasks. Other aspects include but are not limited to, memory in general and the interactive processes of the formation and adjustment of memory and knowledge structures. These models are pivotal in explaining how interpreters manage cognitive load, especially in complex tasks like SI and RSI. Therefore, situated cognition emerges as a key concept, suggesting that cognitive processes in interpreting are not isolated brain phenomena but the changing product of interactions with the physical and social environment. This perspective underscores the dynamic nature of cognition in RSI, where interpreters interact with various technological tools and their environment. Language and communication are explored in the context of interpreting, emphasizing the dynamic and interactive nature of language. The chapter also highlights the challenges and complexities inherent in language as a system of symbolic communication. The interpreting process, particularly in SI, is presented as demanding high levels of multitasking. The importance of memory, attention, and the ability to process and integrate multimodal information is underscored, illustrating the cognitive complexity of interpreters' tasks.

This chapter also explores various aspects of RSI, including the effects of stress, technology literacy, and the use of CAI tools like InterpretBank, are briefly discussed, indicating a growing research interest in this area. It then specifically focuses on CAI tools, especially InterpretBank. The challenges of using InterpretBank in various interpreting contexts are also reviewed, highlighting its influence on interpreters' efficiency and rendering quality.

Finally, the chapter addresses the challenges in evaluating the quality of CAI-supported interpreting, discussing methodologies and key indicators such as terminology efficiency and accuracy. It concludes with research questions and hypotheses centered on the effectiveness of InterpretBank for Chinese interpreting trainees, aiming to investigate its impact on improving the efficiency and quality of RSI rendering.

Chapter 2

materials and methods

This chapter provides an overview of the Chinese-speaking informants and the methodology used in the present study. Methods in this research project were approved by the ethical committee of the University of Bologna. We will first profile informants and explain how we split them into the InterpretBank group and Excel group (see **§ 2.1**). Then we will explain how the source texts were manipulated, involving editing transcripts of six online podcast episodes (**§ 2.2**).

Two applications, Excel and InterpretBank, were employed in the study, along with non-intrusive data collection tools including a Webplayer (MS Stream), a customized, python-based keylogger (Pynput), and a screen recorder (TechSmith Capture). Informants used their laptops or PCs to undertake glossary tasks and booth tasks (see below), simulating a remote interpreting scenario. All computer activities (screen, mouse movements, keystrokes, audio) were recorded by those data collection tools while informants autonomously played the source speech soundtracks in MS Stream.

The study was divided into four sessions:

(a) Cycle I: baseline, pretest; full sample

(b) online InterpretBank training: treatment; (different per group)

(c) Cycle II: control + experimental groups, compulsory tools, post-test 1

(d) Cycle III (control + experimental groups, no condition, post-test 2.

Each data collection session (Cycles I, II, and III), but not the online training, comprised an SI glossary task and an SI booth task. During the 2½ hour glossary task in each cycle, informants extracted terms and compiled individual glossaries from a source script (see **§ 2.2**). InterpretBank's automatic extraction feature would make glossaries compiled with it identical. Variation would thus be potentially concentrated on Excel informants. Furthermore, source scripts for the glossary compilation task and its corresponding booth task were reasonably similar but not identical. Particularly important for our goals was to make sure that the quality of the actual glossary did not undermine booth performance. Thus, after each glossary task, we collected all individual glossaries and compiled them into a master glossary, which included (1) all entries chosen by at least two informants, plus (2) all entries chosen by InterpretBank, plus (3) the missing entries from the list of 33 specifically chosen or added terms by the researcher as potential problem triggers in the booth source speech.

The informants received the master glossary back 30 minutes before the booth task. They had been instructed to review it and were also allowed to introduce as many changes as they deemed appropriate. Hence, the informants were able find what they expected but also what they might have missed, making thus their booth delivery independent of their skills at glossary compilation. We assumed that, with this strategy, memory effects derived from different glossary compilation approaches would remain more or less intact, but a shared master glossary would prevent differences to be related to poor glossary compilation skills.

In the booth tasks, informants performed SI from English into Chinese, with source-speech recordings averaging 13 minutes in each cycle. The informants' renditions were analyzed in terms of both typical and novel quantitative indicators, and a holistic assessment was performed to evaluate the interpreting quality of the informants across the three cycles. Then quantitative and qualitative approaches to quality were cross-referenced. The study also investigated the informants' perceptions of using CAI tools through surveys. To sum up, chapter 2 covers the data collection procedures, followed by an explanation of the data analysis procedures. This includes the process of data cleaning and the indicators involved in the analysis.

2.1 Informants

Twenty-two informants took part in this project as interpreters. All of them were Chinese L1 speakers with English as their L2 (Female: 11, Male: 11, age range: 22–34, average age: 24.7 ± 2.9). All informants were trainees in competitive programs in conference interpreting at Chinese universities and had completed at least two semesters of SI training before enrolling in this study. The informant profiles leave room for substantial individual variations. However, we can only hope to have captured the most relevant, defining factors. It seems that this fuzziness can only be fixed through impossibly large samples of a population ([trainee] simultaneous interpreters in the world) whose clear and extract delimitation seems hardly feasible. Prior to their participation, informants signed an informed consent form approved and provided by the University of Bologna. The informants were compensated for their participation with electronic gift cards worth 600 RMB, equivalent to approximately 77 euros. We used convenience sampling. Although informants were personally recruited by the researcher, there was no prior acquaintance between anyone involved in the project.

Drawing from Direnga *et al.*'s (2016) criteria for *Self-generated identification codes*, the profiling survey's final question in **Appendix A** asked informants to create a unique code consisting of the first two letters of their mother's last name, the number of their siblings, and their birth month, for instance, XI0009. This is a simple approach, yet it aligns with the requirements for participant anonymity and consistent data linkage over time. *Self-generated identification codes* enhance privacy but, as noted by Calatrava *et al.* (2022), they do not guarantee complete anonymity and could potentially reveal group features (Excel group or InterpretBank group).

Since our study focused on introducing a new digital tool in an otherwise customary environment, to prevent confounders, informants needed to use their own computers. Six of them used macOS systems, while the others operated Windows. The use of different operating systems led to the need to install keylogging software that would work with these operating systems. To the best of our knowledge, there is no single keylogging solution that is compatible with both Windows and macOS systems, so we had to consider a customized keylogger. Besides, six of 22 informants reported experience with InterpretBank (see **Appendix A**). Further profiling details on the informants can be found in **Appendix B**.

At the beginning of the study—in Cycle I (pre-test)—the informants were treated as a single group. Afterward, they were split into two subgroups, an InterpretBank group and an Excel group with 12 and 10 informants, respectively. This grouping was based on a preliminary evaluation of Cycle I performance, while also considering the informants' prior experience with InterpretBank. The informants were quite evenly matched as to their performance, but those with that experience were assigned to the InterpretBank group to ensure that the differences between the InterpretBank and Excel groups were not due to this potential confounder. Although an equal split of 11 informants per group was possible, we intentionally assigned 12 informants to the InterpretBank group and 10 to the Excel group for two reasons: (1) to collect more data from the InterpretBank group, as that is the primary focus of the research. Having additional informants in this group helps ensure that sufficient data is captured and (2) to have additional informants in the InterpretBank group provides a buffer against potential attrition, maximizing data collection from a limited number of informants.

Later on, we divided the informants into two groups. InterpretBank informants received training on how to use InterpretBank for glossary compilation before SI tasks and term retrieval during the SI tasks (see **§ 2.5.3**). For the Excel group, we provided a mini-lecture on multimodal information search methods, such as Google search syntax, images, video, e-books, and the like.

Each group used either InterpretBank or MS Excel for glossary and booth tasks in Cycle II (post-test). Since we adopted a situated approach to study the informants' interaction with digital tools, the differential impacts for the corresponding tools were not assumed to be one-shot. We extended the analyses to cover more conditions (e.g., the impact of the strategies employed at glossary compilation on later glossary use) and factors (e.g., ultimate implicit judgment and motivation) in the data collection. Half of the InterpretBank informants came with prior experience, and we recommended them to practice before data collection sessions. Experience could thus be offset but in ways we could not control: no confirmation of practice was collected, so as not to be invasive. We thus simply assumed that the InterpretBank informants would be uneven, but that is not necessarily too problematic when the difference between them and the Excel informants is in focus: novelty is disruptive as well. We had Cycle I and Cycle II because we had experiences from Cycle I that informed Cycle II. We told the InterpretBank informants to practice before the second data collection session (Cycle II). We knew that they might have different levels of experience, so we recommended practice, but we did not formally test it or require reporting, so it would not be invasive.

For these reasons, a third cycle was added. In Cycle III (post-test) the informants were allowed to use the tool of their choice. This resulted in one member in each group using a tool other than the one assigned to their group. To preserve the integrity of the data, these two informants were omitted from the analysis of Cycle III, so as to ensure that the results accurately represented each group's consistent and, at least in Cycle III, willful use of either Excel or InterpretBank throughout the study.

2.2 Input

This study examined the incorporation of multiple sensory stimuli, encompassing both modified and prepared textual material for glossary tasks and booth tasks. A sequence of potential problem triggers was developed to establish a benchmark to assess terminological accuracy.

2.2.1 Source text preparation

We first chose an adequate, naturalistic source. *Dr. Huberman Lab* Podcast series, hosted by Andrew Huberman, a Stanford neurobiology associate professor, explores various health topics for the wide audience of this reputed popular science series.[4] Each episode in the *Dr. Huberman Lab* Podcast series is labeled with specific topic tags. We selected three of these tags and identified all corresponding episodes: *time perception and dopamine* (1); *immune system* (2); and *emotions* (3). For each tag, we selected two episodes from the same speaker, Dr. Huberman, to ensure consistency in the speech and style features. The speeches were then assigned to Cycles I, II, and III respectively (see **Figure 2**).

The podcast transcripts served as a point of departure, authentic source material from a domain expert, providing terminology usage in context and introducing the podcast topics. The six episodes were downloaded as .mp3 files and transcribed into English automatically with the Microsoft Stream service. This approach was chosen because acquiring terminology and subject knowledge prior to interpreting usually requires the transcription of spoken speeches (Gaber *et al.*, 2020). Stream service provides an interactive editing feature to correct the machine-generated script. Thus, we can check contents by listening to all audio stored in MS Stream and manually eliminate errors.

[4] Dr. Huberman Lab: https://www.hubermanlab.com/podcast

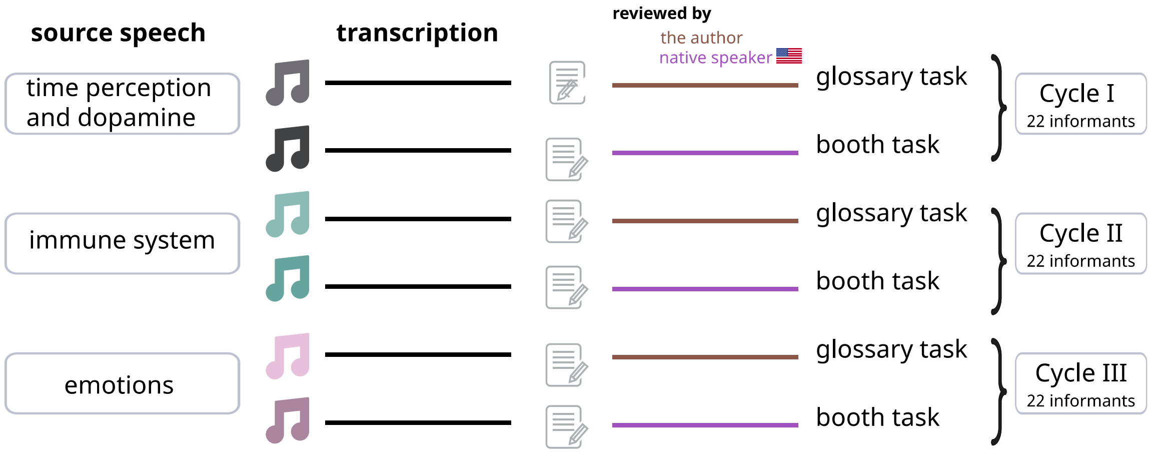


Figure 2. Source speech arrangement.

For each episode pair sharing the same topic, one was randomly chosen for the glossary task, using it as the source text for compiling the glossary. The remaining episode in each pair became the source text for speech recording preparation for the respective booth task. Editing followed two steps. First, irrelevant elements from the opening and closing sections, such as sponsors and advertising, were removed to leave a structured text, closer to SI typical sources. Then, terms were added to increase the density of relevant vocabulary (see below). After that, an English L1 interpreter and interpreter trainer, Prof Richard Samson from the University of Vic, Spain, reviewed them for naturalness and to ensure naturalness and an appropriate total word count balanced for a simultaneous interpretation task (see linguistic features of these six texts in **Table 1**).

task	glossary			booth		
Cycle	I	II	III	I	II	III
word count	8432	8247	8228	1686	1673	1752
complex word count*	1303	1044	1068	228	239	181
complex word share	15.45	12.66	12.98	13.31	14.28	10.33
avg. word frequency for content words	2.37	2.29	2.4	2.28	2.29	2.38
avg. word frequency for all words	3.10	3.06	3.13	3.07	2.99	3.13
full words % (lexical density)	50.66	49.76	47.31	51.43	52.6	49.29
nouns %	26.22	24.58	22.44	29.36	27.62	28.01
adjectives %	7.15	7.98	6.71	7.12	8.49	6.90
verbs %	10.40	10.53	11.06	10.68	11.48	9.30
sentence count	481	398	405	86	84	84
passive sentences count	59	66	85	25	19	17
passive sentences %	12.27	16.58	20.99	29.06	22.62	20.48
sentence length, number of words, mean	11.89	13.83	16.1	8.9	8.78	9.58
number of long sentences**	110	128	108	19	18	20

* Complex words are those those with three or more syllables.
** As a rule of thumb, sentences with more than 25 words are considered long.

Table 1. Linguistic features of texts in glossary tasks and booth tasks.

2.2.2 Potential problem triggers

To facilitate comparisons between InterpretBank and MS Excel use among informants, potential problem triggers were selected from the raw source speech script for booth tasks, always specialized terms to be delivered to a non-specialized audience. The terms were chosen to be not too frequent, but the selection should not be based on our preferences.

First, we will explain the process of selecting problem triggers and the rationale behind the choice of terms (see **Figure 3**). Then, we will describe the conditions of potential problem triggers. In order to select appropriate potential problem triggers, we first extracted keywords from the raw speech transcript for the booth task using AntConc (Anthony, 2022), a typical corpus analysis program, freely available in Win/Mac/Linux.

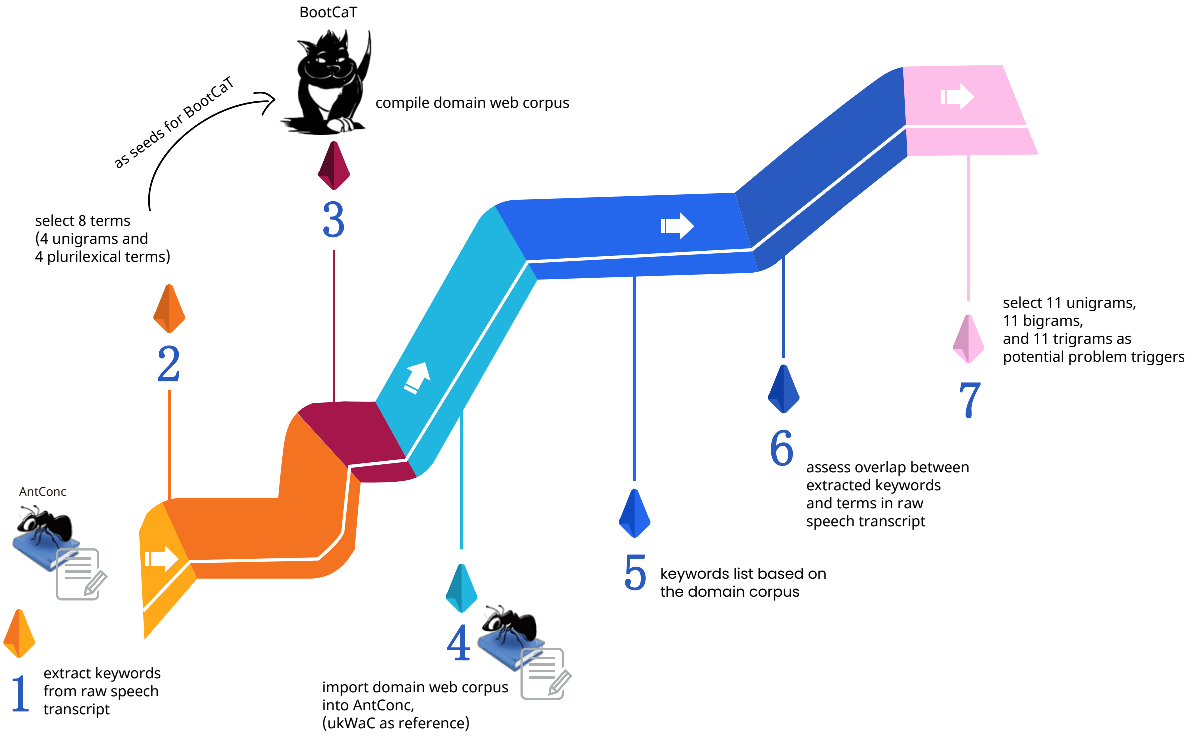


Figure 3. Procedures to identify 33 potential problem triggers.

Second, we selected eight terms from the specialized terms from AntConc based on the frequency—four unigrams and four plurilexical expressions—as *seeds* for BootCaT queries (Baroni & Bernardini, 2004). BootCaT is a Web crawler to compile full-text web-based corpora by extracting information from online resources. It employs keywords as *seeds* for web queries, similar to Google search. However, unlike Google search, BootCaT can conduct multiple Google API web queries concurrently with various keyword combinations (referred to as *tuples* in BootCaT). Users can configure the *tuple* length (i.e., the number of *seeds* included). In our study, we used four unigrams and four multi-word expressions as *seeds*, setting up *tuples* with 2 *seeds* for combination queries. Consequently, BootCaT automatically executed Google searches using these *tuple* combinations, efficiently harvesting relevant full-text documents from the web. This methodology led to the creation of a customized, domain-specific monolingual corpus.

Third, we used AntConc again to further refine our keyword selection. AntConc allowed us to extract keywords from a domain corpus by BootCaT, which were then compared to the terms in the speech script for the booth task. We imported the domain corpus compiled by BootCaT into AntConc, using the ukWaC corpus (Baroni *et al.*, 2008) as a reference. The ukWaC corpus, with 2 billion words, was created from web content within the .uk domain, employing medium-frequency content words from the British National Corpus as the starting points for its web crawl (Baroni *et al.*, 2008). This corpus, representing a broad spectrum of language use and a reference for medium-frequency word usage, helped establish that the domain terms were actually used in the domain context. The ukWaC corpus acted as a benchmark to validate that these domain terms have real usage in domain-specific texts, scaffolding their adequacy. So, our approach involved importing the domain corpus compiled by BootCaT into AntConc, using the ukWaC corpus as a benchmark. By doing this, we got a keyword list based on the domain corpus. In brief, these terms are frequently used within domain-specific information resources. This is because the selection of terms was based on their keyness value within AntConc's keyword list, prioritizing those terms most related to the topic of the corresponding speech.

Then we conducted an overlap assessment between the keywords extracted from AntConc with the terms present in the raw speech transcript. Terms have different lengths (e.g., plurilexical), and we wanted to account for the different conditions of terms. We decided to have them evenly distributed in case they had different impacts. Building on this overlap, we selected a set of 33 *first-time terms* as potential problem triggers, consisting of the 11 most frequent unigrams, 11 most frequent bigrams, and 11 most frequent trigrams.

This approach ensured a principled selection and validation process, to support that our compilation of potential problem triggers authentically reflected the regular and specialized language usage within the domain. We replicated these steps across all booth tasks in three cycles. To further enrich our analysis, we included two repetitions of one term from each category (i.e., *unigrams*, *bigrams*, and *trigrams*), totaling six repeated terms within each raw speech script. There were some other repetitions in the text, but this was a part of a side study. A study on other repetitions would require a second exploratory study.

These repetitions were intended to work as stimuli to observe and analyze the informants' ability to remember and accurately interpret terms that were not new under the potential cognitive difficulties faced while at the booth task. In other words, this complementary approach focused on the informants' WM maintenance, or failure to maintain and recall, in SI tasks. This side project within the dissertation aims to test the importance of cognitive dynamics in task performance. If the results prove interesting, further studies should consider incorporating a higher number of stimuli in different conditions to expand on these findings. We included a minimal number of items to determine whether it is worth further exploring it because it is actually a whole different dissertation.

Consequently, the total count of strategically placed potential problem trigger terms in each speech script can be seen as 39, consisting of 33 *first-time terms* and six *repeated terms*, or else as 33, plus six repetitions that study something else. In any case, as a result, the potential problem triggers belong to two categories: *first-time* terms and *repeated* terms. We will analyze the six repetitions separately when considering recall, but no distinction will be made when analyzing potential problem triggers because there is no way to discern whether informants actually remembered them, or whether this poses a different kind of problem, as to its nature. Full lists of potential problem triggers in each text are in **Appendix C**.

The raw source speech scripts with selected potential problem triggers were then reviewed, modified, and turned into recording scripts by Prof Samson. This reviewer adjusted the recording scripts to ensure the total word count was appropriate for SI tasks to take less than 15 minutes each. Additionally, potential problem triggers were strategically inserted in the recording scripts following the strategies laid out by Prandi (2017) and Frittella (2022). Two types of sentences were manipulated: 23 sentences, which had one problem trigger each, and 5 sentences, which had two problem triggers each. This resulted in a total of 28 target sentences with embedded terms, spread throughout the source speech (for detailed linguistic features of each script, please see **Table 2** in **§ 2.2.3**). These target sentences were not always immediately consecutive and could be followed by ordinary sentences without no problem triggers. This approach aimed to balance potential problem triggers so as to foster a fair evaluation of informants' performance in term accuracy in the booth task.

The set of potential problem triggers is not solely a benchmark for assessing informants' SI term accuracy and recall capacity assessment. It also serves as an indicator of how informants using InterpretBank respond to these potential problems with InterpretBank. Additionally, it partially contributes to the observation of ear-key span and eye-voice span (see more in **§ 2.7.1.5**).

2.2.3 Source speech recordings

We invited three native American speakers to read the edited texts and to record them as source speech soundtracks. The soundtracks were saved in .mp3 format.

Cycle	topic	word count	sentence counts	chunk counts	nsyll	dur(s)*	speech rate (nsyll/dur)
I	time perception	1686	86	86	2558	776.35	3.29
II	immune system	1673	84	87	2383	793.75	3
III	emotions	1752	84	86	2470	777.53	3.18

* dur(s): measured from the initial syllable to the final syllable in source speech soundtrack

Table 2. Source speech script features for booth tasks.

Table 2 summarizes the features of the source speech soundtracks and texts. The acoustic properties, namely the number of syllables (nsyll), duration in seconds (dur(s)), and speech rate (nsyll/dur), were computed using the Praat software package (Boersma & Weenink, 2023; de Jong *et al.*, 2021).

In the source speech soundtrack, an initial 30-minute blank segment was included for ad hoc preparation, suggesting that informants can undertake any preparation actions within the same duration of time. They were allowed to prepare, review, and introduce any changes, such as checking and even annotating term pronunciation from entries in the master glossary. The informants, whether in the InterpretBank group or Excel group, used the same master glossary (see **§ 2.5.1**). This 30-minute blank segment in the soundtrack ensured that each informant had the same time for preparation. Subsequently, two alert signals (1 and 2) within a three-second interval warned the informants that the time was over and captured their attention for the incoming source speech. An instructional message followed (originally in Chinese):[5]

> *In this section, you will hear a mini-lecture that you need to simultaneous interpret. You will hear the mini-lecture ONCE ONLY. While listening to the mini-lecture, please interpret it into Chinese. When it is over, you will be given ten seconds to finish your interpreting.*

The source speech began three seconds after signal 3. Upon its conclusion, informants had 10 seconds (s) to finalize their interpretation before an ending signal. That is, the system still recorded for 10 s after the source speech ended. A representation of components of an *original source speech* recording soundtrack can be found in **Figure 4**.

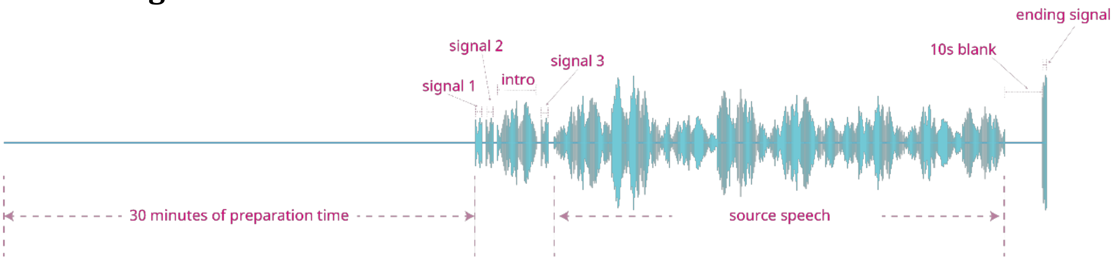


Figure 4. Components of source speech soundtrack.

2.3 Application variables

In interpretation, managing multilingual glossaries is a cognitively challenging task, particularly when dealing with large volumes of multilingual information. In view that interpreters often show reluctance or reject the use of CAI tools in SI

[5] In Chinese script: 下面你将听到一段英文录音。请将听到的内容传译成中文，录音只播放一次。在录音结束播放后,你将有十秒钟额外的时间来完成你的传译。

tasks (Will, 2020), and given MS Excel's popularity for managing terminology among interpreters (Woesler, 2021), we selected MS Excel for the control group. Meanwhile, due to the academic interest in examining InterpretBank, as discussed in **§ 1.5.3**, InterpretBank was chosen for the experimental group.

2.3.1 Excel

Microsoft Excel is part of the MS Office Suite. It is pre-installed on most PCs, and it is the most popular tool for translators to work with glossaries because it is both inexpensive and easy to use (Matis, 2010). MS Excel is flexible for quick data modification and for adding or deleting new entries, and it is widely compatible across operating systems. Tarasenko & Amelina (2020) argues that MS Excel is not a good choice for a terminology management tool and Yang (2021) highlights several drawbacks: it offers only passive queries, it displays only the specific searched results without related information; it tends to be slow, especially with large files, due to its linear search approach; and it restricts item display to rows and columns, limiting its efficiency as a comprehensive terminology management solution.

Still, the use of MS Excel as a glossary-building tool is prevalent among interpreters. Jiang (2013) reports that approximately one out of every four medium-level interpreters adopted Excel for glossary preparation. In a review of terminology management tools, Costa *et al.* (2017) observes that Excel files are widely accepted as an output format in various terminology management systems, for instance, Interplex, Lingo, and AnyLexic. This compatibility underscores the importance of Excel in the field of interpretation. The preference for Excel over more specialized tools, despite its limitations, could be attributed to factors like familiarity, accessibility, and broad compatibility with other terminology management systems. It allows for seamless integration with other tools and platforms, facilitating multilingual glossaries for sharing and management. Additionally, the XLS (XLSX) format deserves special attention when structuring terminology material, as it is a common format for importing data into most CAI and CAT tools in "cloud-based CAT such as XTM Cloud, Wordfast Anywhere, MemSource, MateCat, etc." (Tarasenko & Amelina, 2020, p. 1017). The widespread adoption of this format facilitates the sharing and exchange of glossaries within professional communities. Its compatibility ensures that glossaries can be easily incorporated into other CAI and CAT tools.

2.3.2 InterpretBank

InterpretBank is a comprehensive terminology and knowledge management software available in macOS and Windows systems. The key components include features to extract terms from documents, read and select terms for glossaries, provide translation suggestions for terms, and automatic speech recognition. According to the results of Prandi's (2020) survey, InterpretBank is by far the most widely used CAI tool in interpreter training among the 25 institutions surveyed, being presented to students more than any other CAI tool. In this study, we used InterpretBank in version 8 as representative of advanced terminology management in the CAI tools in the market at the time of data collection, as current version 9 was then not yet released.

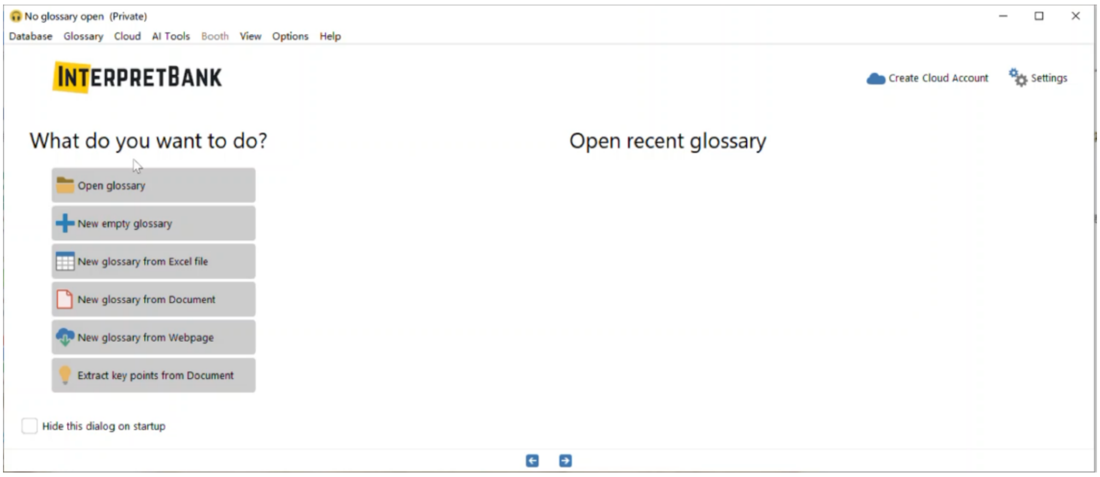
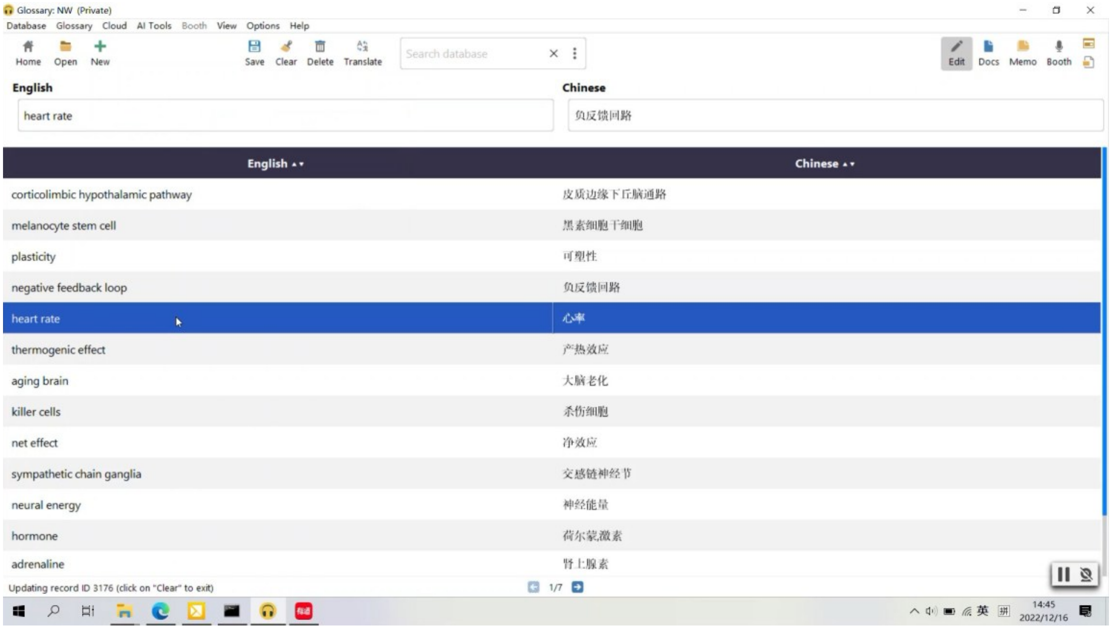

No glossary open (Private)
Database Glossary Cloud AI Tools Booth View Options Help
INTERPRETBANK
Create Cloud Account
Settings
What do you want to do?
Open recent glossary
Open glossary
New empty glossary
New glossary from Excel file
New glossary from Document
New glossary from Webpage
Extract key points from Document
□ Hide this dialog on startup

Figure 5. InterpretBank welcome view.

As illustrated in **Figure 5**, InterpretBank offers a welcome view including quick ways to access some frequent functions. On the left, those functions orient users on how to source their glossaries—either starting from scratch, importing from local documents, or utilizing online resources (such as webpages or keyword searches). On the right, the glossary last used would be shown in the section of the *open recent glossary.* This section enables users to instantly access items they have recently viewed or edited. However, the *welcome view* does not lay out all the key parts of InterpretBank. It is rather an intuitive starting point, orienting users effectively through the software's interface.



English: heart rate

Chinese: 负反馈回路

English	Chinese
corticolimbic hypothalamic pathway	皮质边缘下丘脑通路
melanocyte stem cell	黑素细胞干细胞
plasticity	可塑性
negative feedback loop	负反馈回路
heart rate	心率
thermogenic effect	产热效应
aging brain	大脑老化
killer cells	杀伤细胞
net effect	净效应
sympathetic chain ganglia	交感链神经节
neural energy	神经能量
hormone	荷尔蒙激素
adrenaline	肾上腺素

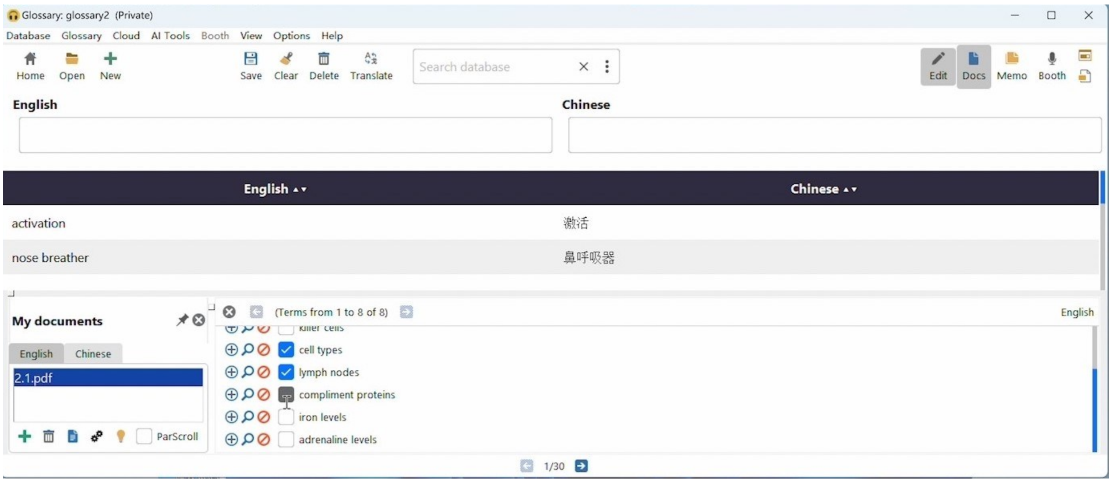
Updating record ID 3176 (click on "Clear" to exit)

1/7

Figure 6. InterpretBank's edit mode.

Four modes of the InterpretBank interface were immediately relevant in this study, *edit, docs, memo,* and *booth,* as shown in **Figure 6**. An *edit mode* is used when editing selected terms and their translations, which offers several editing functions, such as adding entries, deleting entries and duplicate terms, and sorting entries in alphabetical order. Specifically, InterpretBank offers translation suggestions for terms in the entries, enabling users to select their preferred translation. A *docs mode* (**Figure 7**) allows users to use this way to express all file formats (.xlsx, .pdf, etc). In the glossary tasks of the present study, informants were provided episode texts (see **§ 2.2.1**) with two file formats (.docx and .pdf) containing identical content. Informants could choose to use either format for the assigned tasks.

There are three typical sources for creating glossaries in InterpretBank: documents, webpages, and topic keywords entered by the user. The first source is document files. Users can import documents into InterpretBank, which then uses an automatic term extraction algorithm to compile a bilingual glossary from the document content. The second source is webpages. Users can provide a web link, and InterpretBank will download the text from that webpage and extract relevant terms to create a bilingual glossary. The third source involves users entering topic keywords directly into InterpretBank. Based on these keywords, the platform generates a multilingual glossary containing related terms. Please note that newer versions may have modified, improved or changed these features.



English ▲▼	Chinese ▲▼
activation	激活
nose breather	鼻呼吸器

2.1.pdf

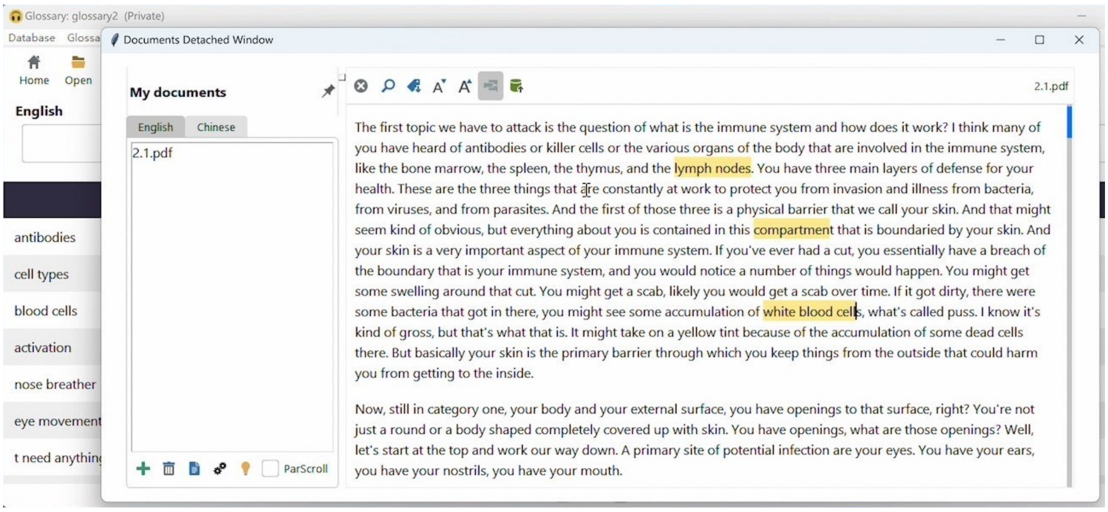
□ killer cells
☑ cell types
☑ lymph nodes
□ compliment proteins
□ iron levels
□ adrenaline levels

Figure 7. Automatically terminology extraction.

The *docs mode* provides *standard* and *smart* terminology extraction modes, based on InterpretBank's algorithms. The standard mode extracts all monolingual terms from the document deemed relevant through an algorithm, whereas the smart extraction mode excludes the terms deliberately marked to be avoided. These modes provide a list of terms extracted from the imported files, and users need to select the monolingual terms they wish to retain from the generated term list (see **Figure 7**). Before registering the selected terms, users can choose to automatically generate their translations based on electronic dictionaries linked to InterpretBank, which provides translation suggestions through access to portals such as

Wikipedia, MyMemory, and Bing (Costa *et al.*, 2017). The automated term extraction can save time, but it also places demands relative to the users' domain knowledge. After all, users need to select the terms from a potentially long term-list that can span several pages, without any context beyond the source text.

InterpretBank also offers a "manual" extraction feature (see **Figure 8**). Users can work with a document in InterpretBank and select terms from the texts by highlighting them. This "manual" approach also supports term extraction from parallel documents (original language + its translation). This function was not used or tested in our study, for one of our goals was to comparatively investigate the informants' information-seeking behavior, and the use of parallel documents might have introduced considerable variation.



The first topic we have to attack is the question of what is the immune system and how does it work? I think many of you have heard of antibodies or killer cells or the various organs of the body that are involved in the immune system, like the bone marrow, the spleen, the thymus, and the lymph nodes. You have three main layers of defense for your health. These are the three things that are constantly at work to protect you from invasion and illness from bacteria, from viruses, and from parasites. And the first of those three is a physical barrier that we call your skin. And that might seem kind of obvious, but everything about you is contained in this compartment that is boundaried by your skin. And your skin is a very important aspect of your immune system. If you've ever had a cut, you essentially have a breach of the boundary that is your immune system, and you would notice a number of things would happen. You might get some swelling around that cut. You might get a scab, likely you would get a scab over time. If it got dirty, there were some bacteria that got in there, you might see some accumulation of white blood cells, what's called puss. I know it's kind of gross, but that's what that is. It might take on a yellow tint because of the accumulation of some dead cells there. But basically your skin is the primary barrier through which you keep things from the outside that could harm you from getting to the inside.

Now, still in category one, your body and your external surface, you have openings to that surface, right? You're not just a round or a body shaped completely covered up with skin. You have openings, what are those openings? Well, let's start at the top and work our way down. A primary site of potential infection are your eyes. You have your ears, you have your nostrils, you have your mouth.

Figure 8. Manual extraction option in InterpretBank.

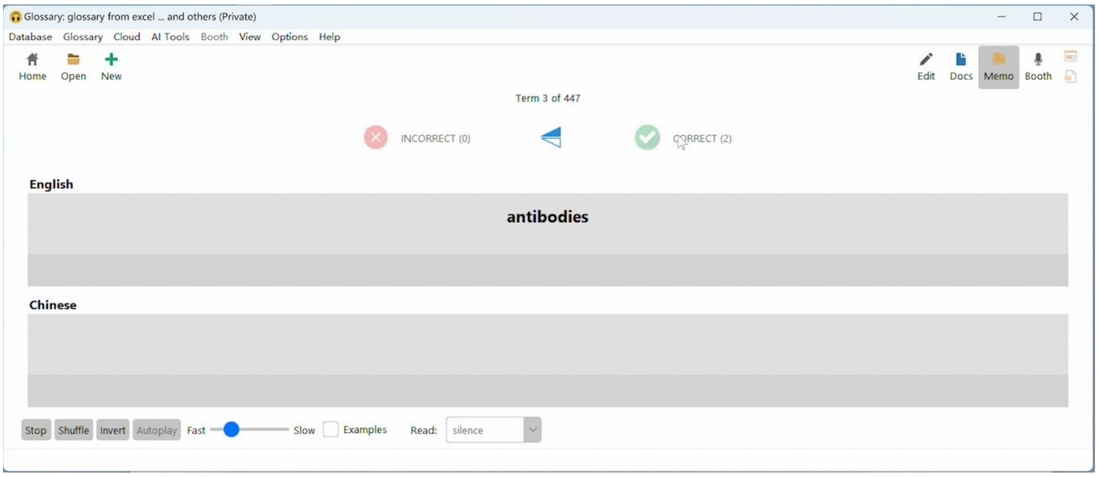
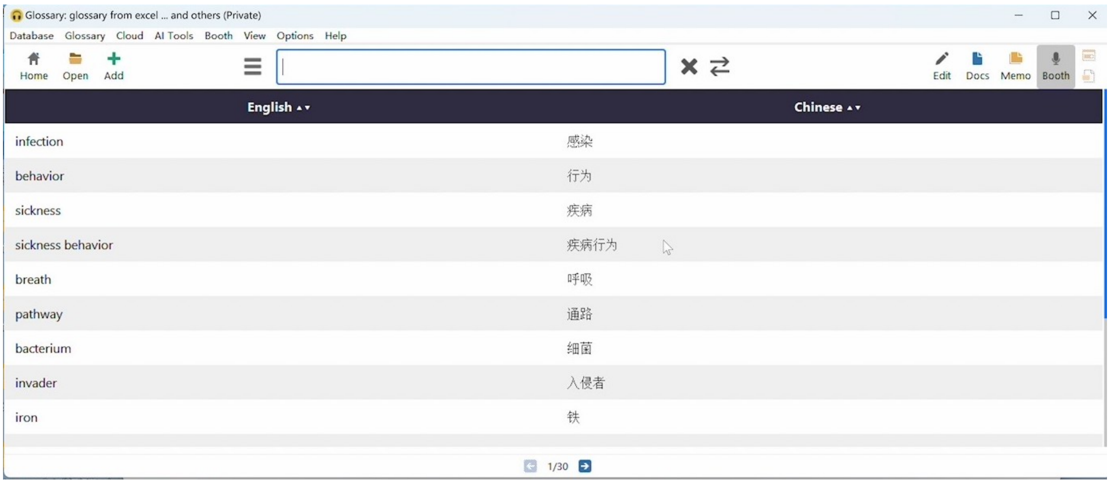


Figure 9. Screenshot of memo mode.

Another mode involved in this study is *memo* (see **Figure 9**), which is a vocabulary flashcard feature. Once users have compiled a camera-ready glossary to be used at the booth, they can activate the *memo* mode, load the glossary, and use it to practice speaking in one language of the relevant pair, prompted by words in the other one (both ways). Either English words show up for the user to speak the chosen Chinese translation, or vice versa. There are two ways to activate word playback: *start manual*, and *autoplay*. *Start manual* means users need to click *incorrect* or *correct* after the translation is shown, compared to the translation displayed at the bottom of InterpretBank with their own translation, and make a judgment before going on to the next word. *Autoplay* means the change of words is automatically controlled by InterpretBank.


Glossary: glossary from excel ... and others (Private)
Database Glossary Cloud AI Tools Booth View Options Help
Home
Open
Add
Edit
Docs
Memo
Booth
1/30

English	Chinese
infection	感染
behavior	行为
sickness	疾病
sickness behavior	疾病行为
breath	呼吸
pathway	通路
bacterium	细菌
invader	入侵者
iron	铁

Figure 10. Booth mode in InterpretBank.

In the *booth mode* shown in **Figure 10**, users can type letters from terms in one of the languages of the relevant pair to into a search window to look for their equivalents in the other one. The default setting for the search window includes a 3-second reset interval following each input. That is, the search window automatically clears all entered contents after a 3-second span with no input.

The ASR function is a cloud-based service. Given the remote data collection setting and the limitation that the ASR function is free of charge for only up to 5 minutes before needing to be restarted, we decided we would not use this function, to avoid complicating the setup and introducing additional complex variables, since Internet connection stability and speeds may vary across informants and our data collection procedures were highly sensitive to the accuracy of timestamps.

2.4 Data collection applications

This study was conducted remotely through three data collection tools: a Webplayer (MS Stream), a customized, python-based keylogger (Pynput), and a screen recorder (TechSmith Capture). The selected data collection tools were deemed appropriate for user-friendly deployment in a remote setting and also to minimize minor errors across various countries and time zones between the researcher and the informants.

2.4.1 Microsoft Stream

Microsoft Stream is a web-based video service application, part of the Microsoft 365 suite. In the present study, we used two of its main functions: audio editing (e.g., automatic speech-to-text transcription) and streaming audio files (no video). Automatic speech-to-text transcription was used for transcribing the source audios (see **§ 2.2.1**), and the audio recordings of each informant's SI delivery in both tasks into Chinese, timestamping all transcripts. Such an automatic speech-to-text transcription editing function can only be accessed by the audio owner, and it is not available to anyone else. Informants were able to access the audios (i.e., source speech soundtracks, see more in **§ 2.2.3**) by sharing links generated by Microsoft Stream. Access was authorized through their email addresses. Once an informant began playing an audio file, they were not allowed to pause, stop, replay it midway, or change the playback speed.

2.4.2 Keylogger

Keylogging entails the act of recording and timestamping the keys pressed on a keyboard (Witte, 2018; Ballier *et al.*, 2019) and, in some keyloggers, the recording of mouse movements as well. Researchers have benefitted from keylogging tools, as they offer comprehensive insights into various processes such as writing (e.g., Miller & Sullivan, 2006; Zhu *et al.*, 2023), pedagogy (e.g., Holm *et al.*, 2022), and lifelogging (e.g., Smeaton *et al.*, 2021). Keystroke analysis, which is a technique involving the analysis of typing patterns and timings, has often been used to evaluate different aspects of cognitive activities. For instance, these analyses have yielded instances where there was a high occurrence of typographical errors in the tasks because it "required too much information to be held in WM for too long…" (G. M. Olson & Olson, 2003, p. 502). This may suggest that high cognitive demands can lead to an increased likelihood of making typing errors, possibly indicating more cognitive effort, thus revealing a cause-and-effect relationship.

In cognitive translation and interpreting studies (CTIS), keyloggers have been employed to understand cognitive processes involved in translating (Aldridge & Fontaine, 2022) and post-editing (Huang & Wang, 2022), but are not limited to these tasks. This study employs keylogging data to infer the informants' cognitive processes while using their computers at the booth, taking into consideration the temporal aspects. To the best of our knowledge, there is no study using keylogging to study interpreting trainees' delivery performance in remote SI with CAI tools. Informants often consult many local and online resources (Sales *et al.*, 2018) when building glossaries. Thus, this research project needed a keylogger capable of recording every keystroke activity on the keyboard, irrespective of the programs the informants engage with when seeking information.

Once launched, most keyloggers run invisibly in the background of word processors, browsers, and other computer software. Since keyloggers often do not require an observer (Bowen & Thomas, 2020), they are unobtrusive and offer a high level of ecological validity. For a recent review of keyloggers, see Rai *et al.*, 2022. There are two main kinds of keyloggers: software and hardware loggers.

Hardware loggers are a physical part that needs to be installed in the computer that is going to be logged, often as a plugin in the keyboard cable. Software loggers are easy to install. Huseynov *et al.* (2020) distinguish four subtypes of software keyloggers: (1) kernel-based; (2) user-space or API-based; (3) form grabbing or form jacking and (4) JavaScript-based. Keylogging is unobtrusive and often can be unnoticed, especially when used to monitor or spy on users without their consent. This is so popular that antivirus spies usually classify all keyloggers as dangerous.

To understand how the logging of data happens, some information is in order about how keyboards work. Every key in the keyboard has a unique value, called *scan code.* When a user presses a key on a keyboard, two scan codes may be generated, one for pressing the key *(keydown),* and another one for releasing it *(keyup).* When a keyboard detects that a specific key has been pressed, it sends the scan code of that key to the central processor. The same happens when the user releases the key. The processor receives the scan code and translates it into a virtual-key code, which is in turn translated into messages that are eventually posted to the appropriate window in the application (Witte, 2018).

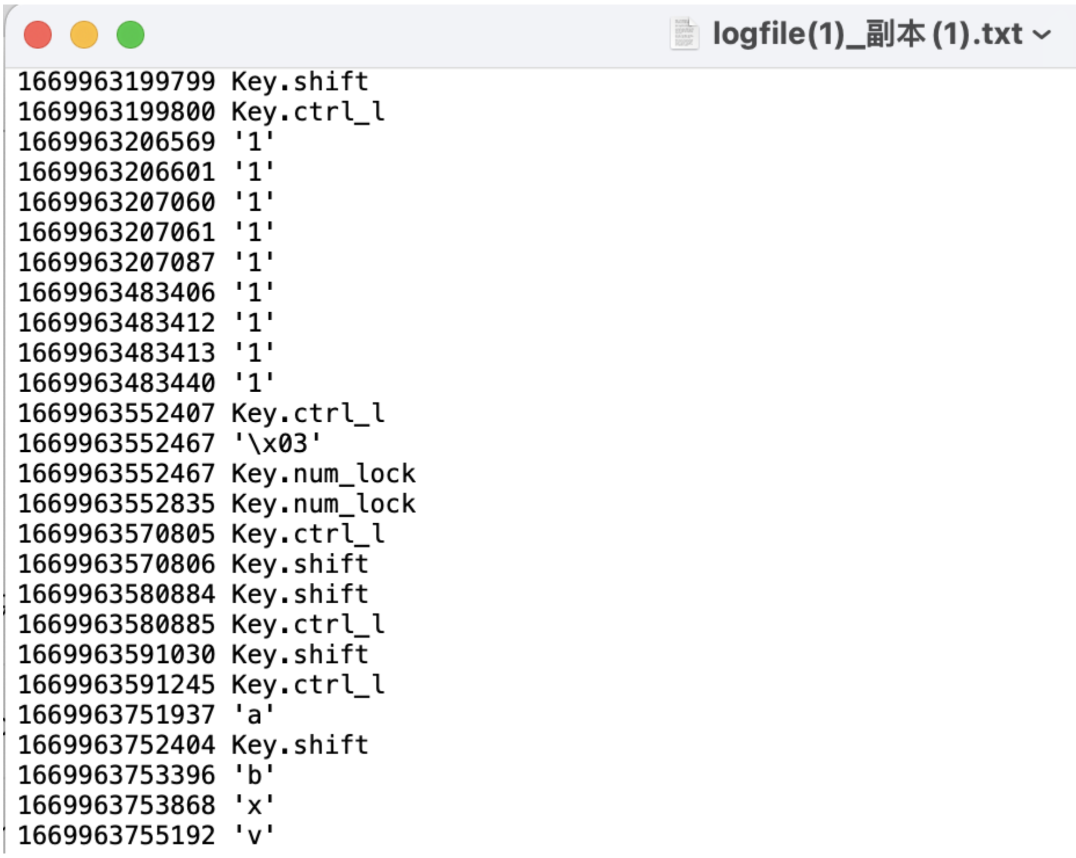
In view that many informants are Windows users but a good share of them use macOS systems, our keylogger had to function effectively on both Windows and macOS platforms. A notable obstacle is the lack of a universally compatible keylogger that can be utilized across operating systems. Non-commercial keyloggers often cater specifically to certain operating systems. For instance, Inputlog (Leijten & Van Waes, 2013) is exclusively designed for Windows, whereas Logger-Man (Hinbarji *et al.*, 2016) is only compatible with macOS systems. Some keyloggers for more than one operating system, like RUI-Recording User Input (Kukreja *et al.*, 2006; Morgan *et al.*, 2013), have discontinued their maintenance, especially of their macOS versions. Furthermore, a few popular research keyloggers, such as Scriptlog (Strömqvist & Karlsson, 2002) and Translog II (Jakobsen, 1999), only record typing activities within very simple word editors built into the keylogger, thus restricting the scope of data collection across different local applications.

Commercial keyloggers are abundant in the market, offering a wide array of choices; for instance, Refog, Spyrix, and Ekran.[6] However, they are primarily designed for parental control, child monitoring, workplace surveillance, and general monitoring purposes. In these contexts, millisecond precision is not a critical requirement. However, in our study, such precision is highly valued for accurately measuring informants' response times with CAI tools during booth tasks and synchronizing keylogging data with screen recordings. Moreover, the use of commercial keyloggers brings forth concerns regarding data security, privacy, and efficiency in practice (Sagiroglu & Canbek, 2009).

Considering that current commercial and non-commercial keyloggers did not meet the requirements for this research project, we chose an open-source keylogger, Pynput (Palmér, 2023), that works within a Python environment. Given that

[6] Refog: https://www.refog.com/
Spyrix: https://www.spyrix.com/
Ekran: https://www.ekransystem.com/

Python environments are compatible with both macOS and Windows operating systems, Pynput can be effectively deployed and utilized by users on both platforms. As Pynput is not a standard executable mode, it requires an easy-copy script for deployment and launching.[7] This in practice erased the differences in operating systems. Pynput is unobtrusive and invisible and stores all typed keys in a log file as a continuous sequence of events, each with a Unix timestamp (see **Figure 11**).



```
1669963199799 Key.shift
1669963199800 Key.ctrl_l
1669963206569 '1'
1669963206601 '1'
1669963207060 '1'
1669963207061 '1'
1669963207087 '1'
1669963483406 '1'
1669963483412 '1'
1669963483413 '1'
1669963483440 '1'
1669963552407 Key.ctrl_l
1669963552467 '\x03'
1669963552467 Key.num_lock
1669963552835 Key.num_lock
1669963570805 Key.ctrl_l
1669963570806 Key.shift
1669963580884 Key.shift
1669963580885 Key.ctrl_l
1669963591030 Key.shift
1669963591245 Key.ctrl_l
1669963751937 'a'
1669963752404 Key.shift
1669963753396 'b'
1669963753868 'x'
1669963755192 'v'
```

Figure 11. Sample logfile from Pynput.

The Unix timestamp is a numerical value representing the number of seconds elapsed since the Unix epoch, defined as 00:00:00 UTC on January 1, 1970. Excluding leap seconds (adjustments to account for variations in the Earth's rotation speed), it can can represent times between January 1, 1970, and approximately January 1, 2091 (Dyreson & Snodgrass, 1993). This timestamp is universally used for recording date and time, decoding to the specific date and time when data is used (Cabral & Minku, 2023). Due to the computing characteristics of the Unix timestamp, various programming languages have extended its precision: Java incorporates both seconds and nanoseconds, while Python extends to milliseconds. In this study, we employed Pynput within a Python environment, ensuring that the Unix timestamp is maintained with millisecond precision. For example, the Unix timestamp *1669963199799* can be converted into *12/02/2022 06:39:59.799*

[7] The deployment steps for Pynput: https://rebrand.ly/geyjxb1

for data analysis.[8] This spared us the need to convert hours, minutes, seconds, and time zones, making it an optimal option for recording keystroke events in remote data collection with informants located in different world locations. So, we can identify variations in keystroke dynamics across time, for each informant.

To activate Pynput the informants were required to type a few commands—in Terminal, for macOS users, or the Windows Command Prompt (CMD) for Windows users—so they were also made aware that the keylogger was running. We thought it was a good tradeoff in that informants would perhaps also feel reassured that their privacy was respected because they remained in control.

Pynput is a typical user-space keylogger. It hooks keyboard APIs inside the active running application and registers keystroke events as if it was a normal piece of the application instead of spyware. The keylogger receives a scan code each time the user presses a key and records it. That is, Pynput only intercepts keydown, which was enough for our purposes. Since everyone in the study used a QWERTY layout, and the basics of the working architecture of registering keystrokes in each operating system were identical, the impact of OS choice on data results is minimal.

Chinese uses a logographic writing system, while English and many other romance languages use an alphabetic writing system. Chinese users cannot type Chinese characters directly using a QWERTY keyboard. Chinese Romanized transcription (pinyin) input methods convert a sequence of Roman alphabet characters into the corresponding Chinese characters. Chinese typists enter the pronunciation of Chinese characters (pinyin sequence) in their QWERTY keyboard. Keyboard conversion applications display a list of Chinese characters mapping to this pronunciation. As they type, keyboard conversion applications such as Sogou Input suggest relevant characters and phrases based on common language patterns, thanks to a predictive text algorithm. As more letters are entered, predictions are refined. Typists choose the correct characters from a list of suggested options. Typists complete entire words or phrases by selecting suggested combinations, and the letters on the screen are then converted into the corresponding Chinese characters. In brief, typists directly type English characters but require pressing *Enter* to confirm, subtly affecting typing efficiency and effectiveness. Although this represents the personal habits of a small group, it also reflects the behavioral characteristics of a certain user base. Chinese input tools like Sogou Input, QQ Pinyin, and Baidu Input are popular software applications for entering Chinese characters with standard QWERTY keyboards that allow direct typing of English characters.[9]

2.4.3 Screen recording

Screen recording has been used in CTIS for at least a decade (Enríquez, 2013; Shreve *et al.*, 2014; Angelone, 2021; Zhong & Xin, 2021). Hvelplund (2019)

[8] The conversion steps have been made available in https://zenodo.org/doi/10.5281/zenodo.10520444

[9] Sogou Input: https://shurufa.sogou.com/
QQ Pinyin: https://qq.pinyin.cn/
Baidu Input: https://srf.baidu.com/

integrated eyetracking and screen recording data to argue that digital search behavior is an intrinsic component of the translation process, among various other subtasks. Further exploring this integration, Cui & Zheng (2022) examined linguistic and extralinguistic consultations combining eyetracking and screen recording to explore cognitive resource allocation and information-processing patterns in English–Chinese translation. No study seems to have used screen recording to analyze informants' behavior in interpreting tasks from a cognitive perspective.

TechSmith Capture is a free screen capture program that has gained popularity in the wake of the COVID-19 pandemic due to its increased use in remote academic training and workshops. The program features, including its ease of use and user-friendly interface, make it an attractive choice for individuals and organizations. TechSmith Capture allows users to record audio from a headset microphone along with the screen recording (see **Figure 12**).

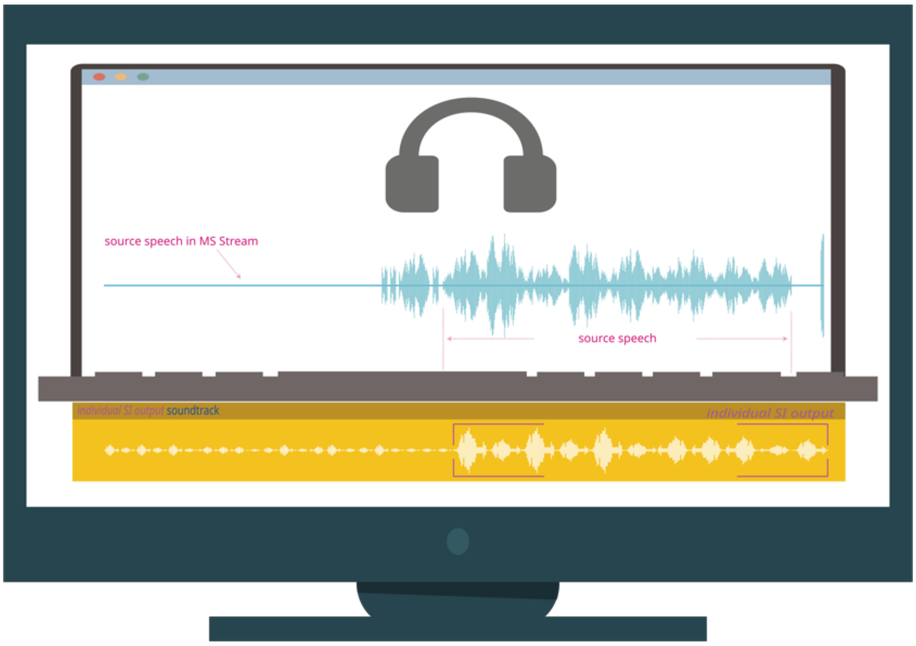


Figure 12. Voice-over screen recording.

Upon completion, users can save their voice-over screen recording file locally. Additionally, TechSmith Capture streamlines the post-recording process by uploading the screen recordings with audio to a personal cloud space. Once the video is uploaded, a URL is generated and copied to the user's clipboard for easy sharing. This feature not only saves the informants' time when transferring large media files across countries, but it also provides added security for the entire recorded file when the informant is sharing the file with the researcher. The automatic upload feature eliminates the need for manual file transfer, making it a convenient and efficient option (Chicca, 2022; Lewis, 2022). Micro errors may occur during screen capture due to variations in monitor refresh rates, latencies in connections, and each computer's processing speed. To minimize these errors, informants were advised before tasks to test and ensure stable internet connections, disable irrelevant applications, and maintain a quiet, comfortable environment. Despite these measures, such micro errors are inherent and uncontrollable, possibly reducing but not extinguishing potential variations and confounders in the situated nature of tasks.

2.5 Study design

This study adopted a mixed-method pretest and posttest approach and was split into four rounds to ensure a comprehensive analysis of informants' documentation behavior over time.

- *Cycle I:* this initial round was the pretest. It served as a baseline for all informants, with no group distinction, providing initial raw data against which subsequent changes could be measured in the following cycles.
- *online training session:* specifically designed for the InterpretBank group, this minimal treatment session aimed at equipping informants with the skills and knowledge for employing InterpretBank in glossary compilation and term retrieval with InterpretBank in booth tasks. Excel group received a generic training session to maximize similarities
- *Cycle II:* this round was the first posttest and followed the introduction of control and experimental groups. It aimed to capture the informants' reactions to using a new tool.
- *Cycle III:* similar to Cycle II, this round involved both control and experimental groups but with no specific tool requirement. It worked as a second posttest, further evaluating the informants' learning effects over time, particularly their progress in performance and the effect of the new tool, now with no surprise.

Each round was separated one week from the next one. Except for the online training session, each data-collection session (i.e., Cycles I, II, and III) comprised a glossary task and a booth task. Throughout these cycles in four rounds, the focus was on capturing the evolving dynamics of the informant's glossary compilation in SI preparation and term retrieval behavior during SI rendering.

In the present study, raters classified each informant's recording into one of six categories, ranging from *bad* to *excellent*, also labeled with numbers (see **§ 2.5.5.3**). Afterward, holistic assessment by PhD students specializing in cognitive approaches to interpreting was checked to determine both reliability and inter-rater reliability. Holistic assessment relies on intuitive evaluation without explicit guidelines. By involving PhD students as raters who assume the role of informants, we aimed to gain insights into the factors these young researchers consider when evaluating interpreting quality holistically. While we reported and discussed the holistic results of informants' renditions (see **Table 29** in **§ 4.4**), these holistic results (over-)simplified the assessment by averaging raters' numerical categories. Representing quality with numerical labels itself is an assumption that may not fully capture the nuances of the assessment process. Using number labels assumes a mathematical relationship and computing the numbers simplifies the assessment of informants' SI performance quality. However, we need to justify whether such a mathematical relationship truly reflects raters' decision-making processes. To address this, we applied an intra-subject analysis to examine the criteria influencing raters' assessment decision-making in the holistic assessment (see **§ 2.5.5**).

This additional analysis serves two purposes: First, it allows us to validate the assumptions underlying the numerical categorization used in the main study's holistic assessment results. Second, it provides an opportunity to explore the decision-making criteria employed by raters when conducting holistic assessments. Understanding these criteria can inform and refine future assessment methods, contributing to a more comprehensive understanding of interpreting assessment (i.e., qualitative and quantitative approach) as a whole. This was not our main goal in this project, but no claim can be made about changes in quality due to using a tool without addressing how such quality is to be assessed. Additionally, it is important for replication and reproduction purposes.

2.5.1 Glossary tasks

In this study, informants were required to engage in a glossary task that took them 2½ hours on average. The task involved extracting terms from a text similar but not identical to the source speech they would later face in the booth task. Informants in Cycle I were free to adopt the term extraction method they preferred, e.g., manual extraction by reading and selecting a term from the source text, automatic term extraction from local applications or online services, or a combination thereof. As mentioned in **§ 2.2.1**, the source texts were (manipulated) speech transcripts of an online podcast and retained oral language style and elements. The topic of the text was the same as the respective booth task. Hvelplund (2019) and Chen *et al.* (2021) have shown that increased topic familiarity and cultural background knowledge can improve interpreting quality.

After completing this task, the informants submitted their individual glossaries in .xlsx files. The researcher then compiled a master glossary out of all individual glossaries that included all terms chosen by at least two participants. The rationale behind employing a master glossary was to minimize the impact of glossary compilation skills on glossary use at the booth. InterpretBank always selects the same terms for all users so, in practice, the master glossary included all terms that InterpretBank extracted automatically, plus the terms not selected in this way that at least two informants had chosen, whether from the InterpretBank or the Excel groups. Additionally, each master glossary also included the 33 potential problem triggers, which only partially overlapped with the extracted terms, for it was in each case a slightly different text. The master glossary included 95 terms in Cycle I, 96 terms in Cycle II, and 97 terms in Cycle III.

Informants received the master glossary 30 minutes prior to their booth task.[10] The glossary, presented in an MS Excel file, was organized into two columns: one for English and the other for Chinese translations. The informants were then

[10] The master glossaries for the booth tasks can be found in https://doi.org/10.17605/OSF.IO/6XF7Y

invited to review it and, if so inclined, freely modify their own copies. In Cycle II, the Excel group was free to review the glossary using their preferred tools but the informants in the InterpretBank group entered the master glossary into InterpretBank for activities like entry verification and editing, and memorizing terms. In Cycle III, all informants were permitted to use any tools or methods they chose to prepare the master glossary, again with a window of 30 minutes immediately before commencing the booth task.

Individual, autonomous glossary-building compilation was repeated in Cycles II and III. In Cycle II, each group used their assigned tool, whether InterpretBank or MS Excel, to compile their glossaries. In Cycle III, both groups were free to choose their tools. This strategic choice made it possible to evaluate the informants' preferences. It hence provided prospects into whether the informants would adhere to InterpretBank or revert to their original ways and tools.

2.5.2 Booth tasks

We used the source speech soundtracks to control the progress of booth tasks. Informants just clicked the beginning of the source speech soundtrack and followed it until the ending signal of the recording. Before the SI tasks in the booth tasks, the first 30 minutes of a source speech soundtrack were silent, allowing informants to prepare ad hoc for the incoming SI tasks. Prior to that, we distributed the digital version of the master glossary, which (reminder) we compiled based on the individual glossaries and the 39 potential problem triggers (see **§ 2.5.1**).

Each booth task lasted approximately 13 minutes. No informant had an audience or a boothmate. During booth tasks, in Cycle I, all informants were free to use tools for term retrieval. In Cycle II, during the SI task, the Excel group informants could access Excel for term consultation, and the InterpretBank group informants could access InterpretBank's booth mode for term retrieval. In Cycle III, informants were free to use any tools for possible term consultation. The behavior of all informants was synchronically recorded—TechSmith Capture registered both their computer activities as displayed on their screens and their SI renderings, and Pynput logged keystroke events.

After each booth task, we downloaded the screen recordings (with embedded rendering audio recording), which were automatically uploaded to TechSmith Capture, from individual sharing links generated by the software. After the booth tasks in Cycles II and III, the informants were welcome to share their opinions on their SI performance and CAI tools through surveys. Also, at the end of each booth task, their edited master glossary versions were collected.

In brief, apart from the tools used in the different groups, the original individual glossaries were not used in the booth tasks. Rather, customizable master glossaries were used as sources of possible term consultations in the booth tasks. We did not explicitly explain the replacement of the glossary and its motivation but let them focus on compiling their own individual glossaries to avoid neglecting the tasks in Cycles II and III, because they would probably have the expectation of receiving an improved glossary before the tasks that they could customized.

2.5.3 InterpretBank training

An online training session of InterpretBank was held at a time point between the close of the booth task in Cycle I and the start of the glossary task in Cycle II. The 2½-hour online training session was designed to familiarize informants with the various functions of the platform. One key aspect covered during the training was glossary compilation.

The training session also covered manually selecting terms and adding them to the glossary. This is useful for terms that are not automatically extracted by InterpretBank or for terms that need to be added to the glossary for a specific interpretation task. Another aspect covered was the use of the booth mode, which allows users to access InterpretBank for term retrieval during SI tasks. This allows for quickly finding definitions or translations of terms while interpreting.

All training contents were delivered through live demonstrations. After the live demonstrations, informants had one week to practice with InterpretBank, while also being provided with pre-recorded video tutorials explaining its key functions covered in the training. Informants were free to ask any questions related to the topics covered. After the training, informants were given a self-trial assignment to test their familiarity with using InterpretBank for glossary compilation and term retrieval in an SI task. The training was only conducted online but was made available to informants in the InterpretBank group for them to review it at will. For the Excel group, a multimodal information consultation workshop was provided. The training content covered Google search syntax, image search, a brief introduction to OpenAI, and e-book consultation. The workshop also lasted 2½ hours as well and was recorded too. During the training phase, each group's treatment was exclusively limited to the informants assigned to their respective groups. After all data collection procedures were completely finished, the recorded training videos of each group were shared among all informants. This allowed informants from one group to access the training content of the other group, enabling them to benefit from each other's training materials. *Noblesse oblige*.

2.5.4 Survey

We surveyed all informants about the use of tools after Cycle III to capture their attitudes toward using either MS Excel or InterpretBank across the cycles. However, in the case of InterpretBank informants, who were in focus, we made an additional survey after Cycle II to collect their impressions right after introducing the tool. The survey for the Excel group comprised three sections (see **Appendix D**). The first section focused on general opinions, including a self-assessment of interpreting performance and two questions at the end about attitudes toward the CAI tools used. The second section dealt with glossary tasks, featuring four statements exploring preferences for term retrieval tools (mobile or PC), checking source term pronunciations, verifying translation reliability, and using automatic term extraction. The third section contained four statements about booth tasks, probing the user's reliance on memorizing terms, the use of self-made or existing glossaries, the necessity of CAI tools, and the need for CAI tool training. Responses to the glossary task statements used a five-point Likert scale from *totally disagree* to *totally agree*. Similarly, responses to the booth task statements ranged from *never* to *always* on a five-point scale.

There were slight differences in the survey arrangements for the InterpretBank group. **Appendix E** aimed to apprehend the effect of new tasks and also capture the informants' attitude changes from Cycle II to Cycle III. Like the Excel group's survey, it consisted of three sections. The first covered general opinions, including a self-assessment of interpreting performance and two questions about attitudes toward using InterpretBank. The second section focused on glossary tasks, with three statements on automatic extraction, manual extraction, and attitudes toward glossary compilation convenience with InterpretBank. The third section presented three statements about booth tasks: the need for term retrieval, attitudes toward retrieval efficiency, and reducing pressure during interpretation. Responses to these statements were also measured using a five-point Likert scale, ranging from *totally disagree* to *totally agree* for the second section, and from *never* to *always* in the third section. After one year of the study, a follow-up survey (see **Appendix F**) was conducted specifically for the InterpretBank group so as to determine whether informants use InterpretBank in their daily workflow and, if so, the main functions they use.

In sum, procedures across cycles were identical. Three distinct but similar speeches were interpreted across cycles available upon request. The InterpretBank group attended an online training workshop on the use of InterpretBank and engaged in self-directed trials to familiarize themselves with the use of InterpretBank during SI. The informants had compiled and sent their own glossaries in advance on similar texts. They later received and adjusted a compiled master glossary 30 minutes before each booth task. Afterward, they activated the screen recording and keylogging software, right before they began the SI task. Then they interpreted a single-play English speech recording and stopped interpreting at a designated ending signal, 10 s after the source speech ended. C2 and C3 only differed from C1 in that about half of the informants used InterpretBank and the other half, Excel.

2.5.5 Holistic assessment

The holistic evaluation of the audios entailed recruiting raters and preparing the materials, as well as the steps for conducting the holistic assessment.

2.5.5.1 *Raters* Quality raters were Chinese PhD students with extensive training in conference interpreting who will probably try to pursue an academic career. Five Chinese PhD candidates (here nicknamed Félix, Jules, Luc, Maxime, and Quentin) working on interpreting tasks from CTIS perspectives were recruited as volunteers for this study.[11] All of them were Chinese L1 speakers with English as their L2 (F:M: 3:2, age range 26–38). No formal profiling was performed of raters to

[11] Again, gender was not considered a variable through this project, and the pronouns do not necessarily represent the gender of the informant.

avoid undermining their self-confidence. A minimal guarantee of professional prowess, scholarly interest, and language command was indirectly supported by the fact that they succeeded in obtaining very competitive fellowships in the UK, Beijing, and Shanghai. Prior to this study, they were invited to fill in an intent form **(Appendix G)** collect the PhD candidates' sociodemographic information, such as age, gender, affiliation, and other relevant details. All of them declared that they thoroughly and completely listened to all audio files assigned to them. The last question of the survey generated a unique Self-Generation Code for each PhD candidate based on their answers, which was used to maintain anonymity. This setting was the same for informants in the main study (see **§ 2.1**). Therefore, raters are anonymous with tagged codes.

By the time they participated in this study, the five raters must have had their own criteria for assessing interpreting quality, even if implicit. Such criteria are probably rooted in their own experience as trainees and as recipients of interpretations. It would be unreasonable and impractical to ask them to abandon or modify their own criteria and use instead predefined rubrics whose definition, understanding, and application may be uneven anyway. With this in mind, we opted for a holistic assessment approach instead of using specific rubric criteria. Our goal was to prompt them to intuitively and holistically assess the recordings, possibly like an expert audience, primarily concerned with the completeness of the delivered information but knowledgeable of other factors impacting quality. We also analyzed quantitative performance indicators of the recordings *ex post facto*. This approach to assessment lets us discern individual differences between raters, as a function of the impact of single quantitative indicators on their assessments, while at the same time, it avoids any distortion of their individual criteria.

2.5.5.2 *Audio preparation* As a reminder, 22 informants provided three recordings of their SI renderings (Cycles I, II, and III). As a result, there were 66 audio files to be rated. These recordings, derived from original speeches on three distinct topics (see **§ 2.2.3**), had an average duration of 13:11 minutes each, after excluding non-SI related content. Having so many files to assess would discourage volunteers, so we opted to reduce the number of recordings each evaluator would assess. Additionally, to evenly assess the quality of SI renderings and to determine the inter-rater reliability among the five raters, the study combined stratified sampling and randomization to avoid order effects, which was implemented in two steps. Each rater was randomly assigned evaluating either 45 or 46 audios.

The first step involved taking five SI renderings each from Cycle I, Cycle II, and Cycle III. These files—let us call them *5C1* (meaning 'five from Cycle I')*, 5C2,* and *5C3*—were chosen based on our subjective judgment (see **Figure 13** for step 1), and they aimed to present a range of quality, from the worst to the best of each cycle. 5C1 files were internally randomized. This reordered set of 5C1 was then fixed and consistently presented as the initial set for evaluation by each rater. The remaining 10 files to study inter-rater reliability (i.e., 5C2 and 5C3) were also evaluated by all five raters. 5C2 and 5C3 files were thoroughly shuffled and distinctly marked by us for us, but not for the raters, who would not know. After extracting 5C1, 5C2, and 5C3, there remained 17 audios in each cycle. Consequently, this resulted in 17 × 3 = 51 audio files per cycle (see **Figure 13** for step 2). These files underwent at separate treatment. A steady number of randomly chosen audios from each cycle, together with all 5C2 and 5C3 files, were then assigned for evaluation by three different raters in fully randomized orders. The specific strategy was as follows.

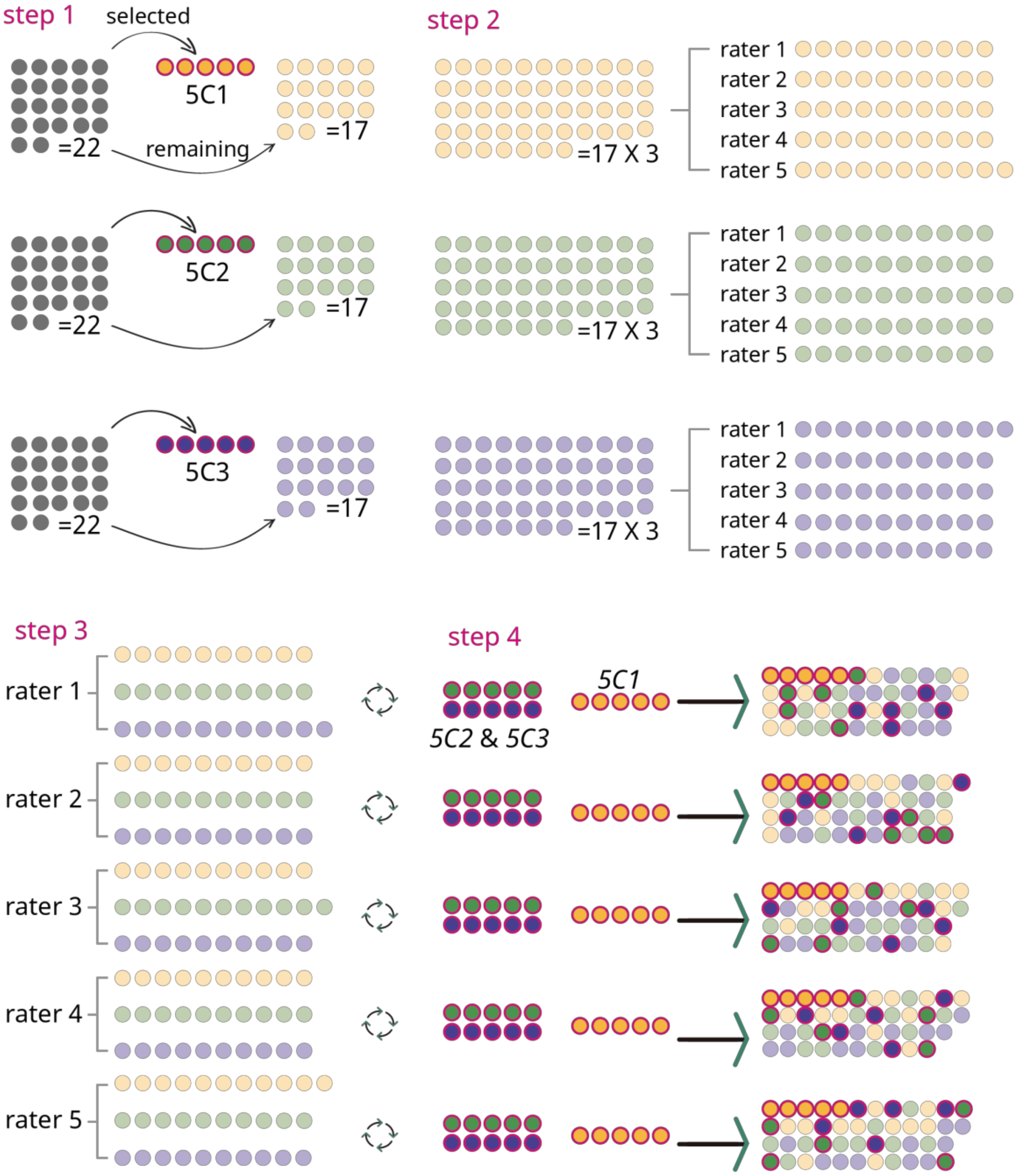


Figure 13. A simplified randomized audio procedure.

The 51 audio files within each set were randomly but evenly allocated among the five raters, adhering to two conditions. First, each audio file was to be assessed by three different raters to ensure diverse perspectives. Second, no audio file was to repeat the exact choice of raters and order within the rating sitting (see below), meaning that each rater evaluated a unique set of files from each set. In this allocation for 51 audios in each cycle, one rater may receive 11 audios from a specific cycle while the remaining four received 10 each in this cycle (see one possible allocation in **Figure 13** for step 3). To prevent any rater from consistently receiving the largest set across all cycles, we made slight adjustments to the number of audios each rater received in each cycle. For example, a rater assigned 11 audios in Cycle I would not receive 11 audios in Cycle II or Cycle III. After the allocation, each rater ended up with 30 or 31 audio files total from Cycle I to Cycle III, apart from the 5C1, 5C2 and 5C3 files. This approach can be done by a Python script for Cycle I, which was then similarly applied to Cycles II and III.[12]

The final step involved full randomization of each rater's set of 30 or 31 audio files together with the shared 10 5C2 and 5C3 audios. Finally, the set of 5C1 files was placed in the first five slots of each rater's set of audios to assess. (see **Figure 13** for step 4).

In brief, each rater was assigned a set beginning with the fixed order 5C1 for each rater. This was followed by a randomized assortment of 30 or 31 audio files from C1, C2, and C3, combined with an additional 10 files from 5C2 and 5C3, bringing the accumulated 40 or 41 audios. Notably, each audio from 5C2 and 5C3 was randomly interspersed within these 40 or 41 audios. Consequently, each rater evaluated a total of either 45 or 46 audio files. 5C1, 5C2, and 5C3 are especially marked by us but raters cannot tell from audio. The set of 5C1 was placed at the beginning of the evaluation to mitigate the raters' new tasks effect while also serving as reference points for calculating inter-rater reliability, along with audios from 5C2 and 5C3.

2.5.5.3 *Rating procedures* The rating task consisted of assessing 45 audio files, each one about 13 minutes long. That is, just listening to all audio files completely amounts to a total of 10 hours. The whole task may thus take between 12 and 15 hours. The raters were given three weeks to complete 45 or 46 SI rendering audio recordings. The rationale behind this time frame was to avoid overexposure and fatigue but to ensure that the raters kept a steady approach to their assessments. However, they would always start with a fixed set of five audios (i.e.,5C1). These five audios were placed in the same fixed order at the beginning of *sitting* 1 for every rater, as part of parameters for inter-rater reliability. As a result, *sitting* 1 comprised a total of six assessments: the five fixed-order audios followed by one additional audio. Therefore, each rater evaluated eight *sittings* in total, with each sitting contained within a separate survey on the PsyToolkit platform. Each sitting comprised no more than six assessments (see **Figure 14**).[13] PsyToolkit is a free-to-use platform to conduct survey-based data collection and cognitive psychological experiments. It offers an easy-to-understand interface, with programming scripts that allow researchers to control the elements and layouts of

[12] The python script (taking Cycle I for example) can be found in https://doi.org/10.5281/zenodo.10520424

[13] PsyToolkit: https://www.psytoolkit.org/

questionnaires. Researchers can use code to freely manipulate the layouts and the orders of questionnaire items.

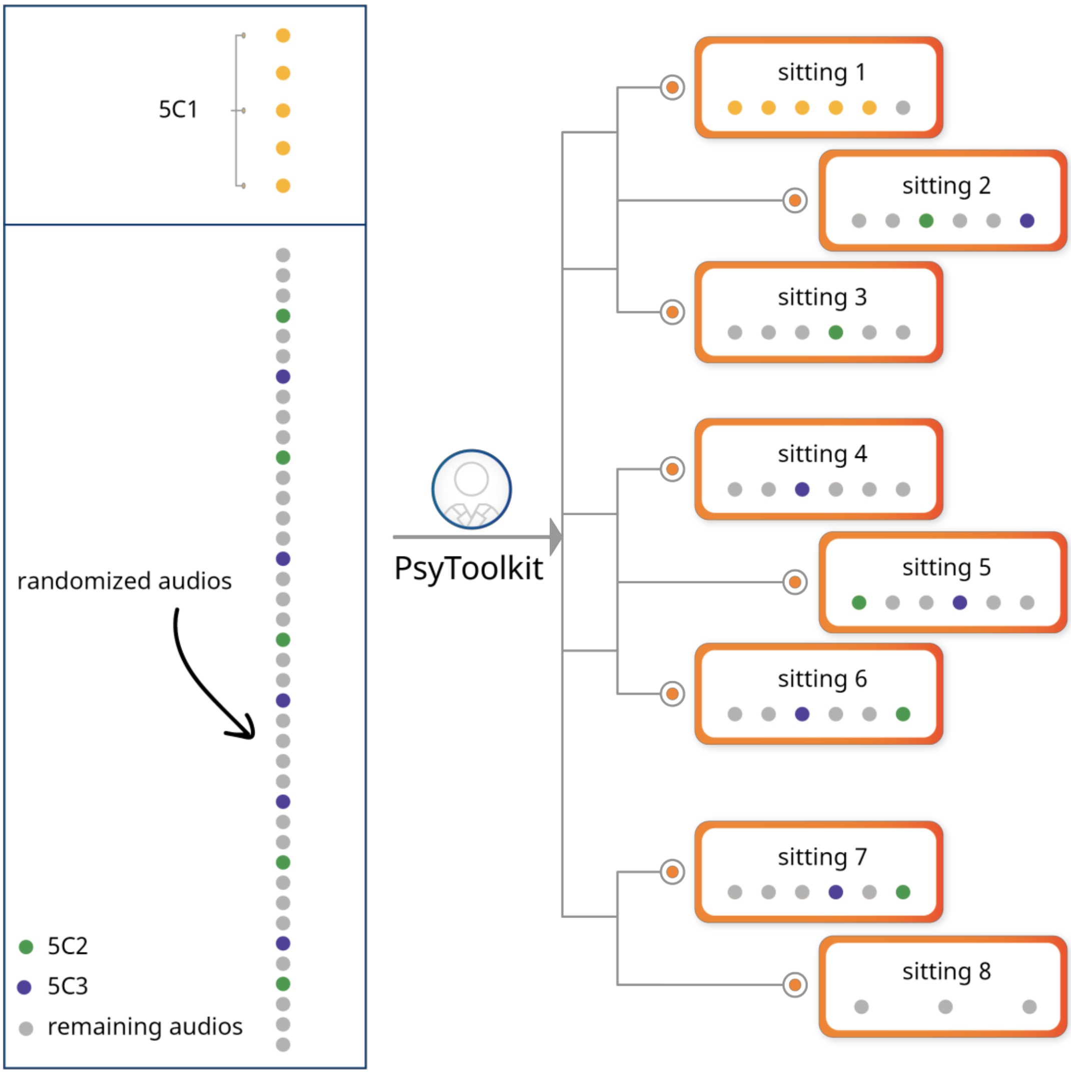
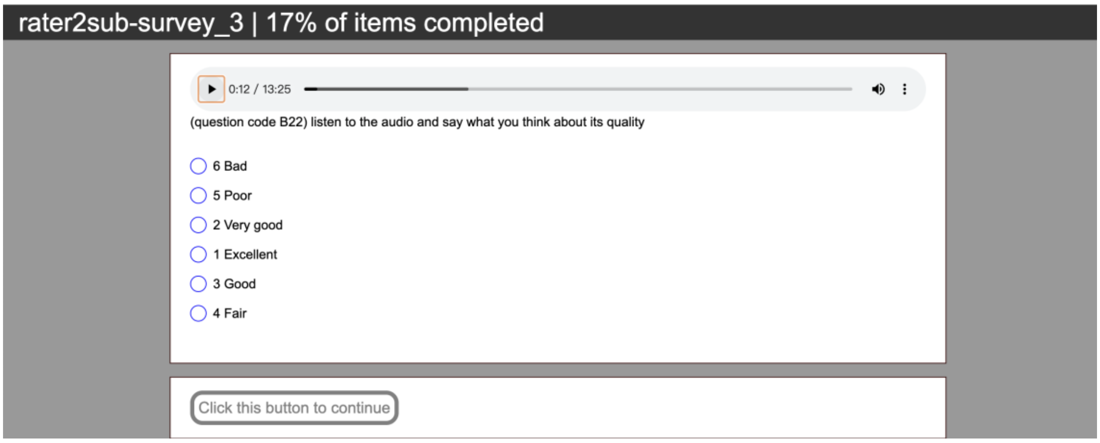


Figure 14. Individual survey structure.

Given the challenge of assessing all audios in one shot, the sitting was divided into a series of rating sections or *sittings* (see below). Each informant was provided with eight *sittings*, with a total number of 45 or 46 audios distributed evenly across these *sittings*. Given that the set included 5C1 followed by 40 or 41 randomized audios, this division was designed such that each *sitting* contained no more than six audios, The number of *sittings* was directly linked to the number of audios in each sitting. With the average length of one audio being 13.11 minutes, a sitting of six would total just over an hour. Reducing the number of audios per sitting to five would increase the total number of *sittings*, potentially overwhelming the raters and passing undue burden onto them. Under this circumstance, the first seven sittings contained an equal number of audios (6), while the last sitting varied slightly in number, depending on whether the total number of audios randomly assigned to the rater was 45 or 46.

Raters had to categorize each recording separately and classify their rendering quality into six categories—*bad, poor, fair, good, very good,* and *excellent*—based on their intuitive assessment, using their own personal criteria. They were in fact discouraged from using anybody else's system, particularly any rubric-based system or cheat sheet itemizing quality aspects. The raters were further instructed to evaluate a whole sitting in one sitting and to do it exactly in the assigned order. The raters could assess all audio at their own pace, with the option to re-listen to any audio if necessary. However, they were required to listen to each completed audio and assess it before moving on to the next. This evaluation approach aimed to mitigate potential biases such as serial evaluation and other effects that might influence the raters' performance, as discussed by Muñoz & Conde (2007) and Conde (2009, 2012a, 2012b). There were, intentionally, no further guidelines or instructions.



rater2sub-survey_3 | 17% of items completed

(question code B22) listen to the audio and say what you think about its quality

○ 6 Bad
○ 5 Poor
○ 2 Very good
○ 1 Excellent
○ 3 Good
○ 4 Fair

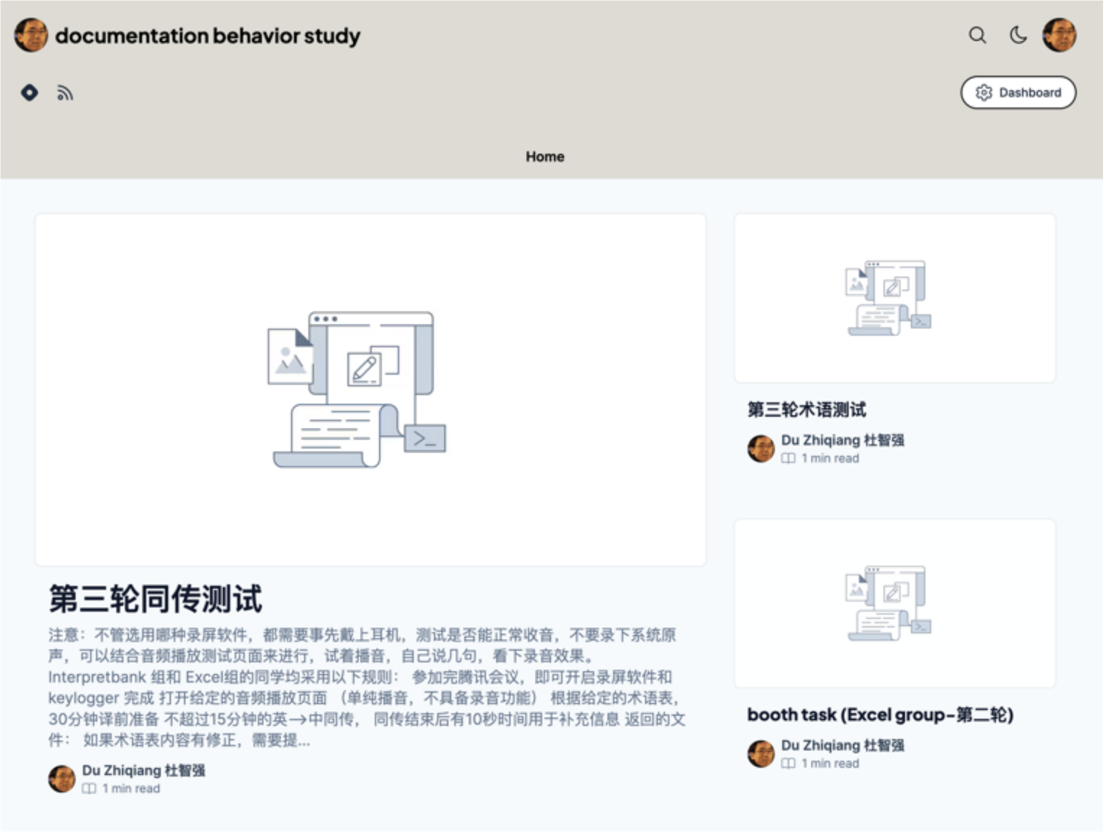
Figure 15. Assessment of one audio in the PsyToolkit survey.

The holistic evaluation process was implemented as a survey, where raters submitted their ratings for each recording after they finished listening to it. The survey was conducted online using the PsyToolkit platform (Stoet, 2010, 2017), a free tool designed for online psychological research. The PsyToolkit platform was set to automatically randomize the order of the audio within each *sitting*, except for the first five, to reduce the new-task effect. Besides, the sequence of six categories for each audio was also randomized automatically by the PsyToolkit platform (see **Figure 15**). In other words, each time an informant accessed a *sitting* as a survey, the sequence of informants' audio recordings presented would differ.

People tend to rely on labels rather than numbers or symbols to determine category boundaries (Gervits *et al.*, 2023), so the number next to the label *bad (6), poor (5), fair (4), good (3), very good (2),* and *excellent (1)* is a cue to help participants choose the right category of their choice instead of simply assigning a numerical value to each audio. Assessment choices were automatically randomized by PsyToolkit (see **Figure 15**). Changing the position of the choices was meant to reduce reliance on the numerical assignment of quality scores, mitigate the effect of learning, and ensure that raters willingly chose their judgment options and did not click one out of routine.

2.6 Data collection

Due to the international travel restrictions of Covid 19 pandemic when the data collection period, the raw data had to be collected remotely, but this approach was adopted not solely due to the pandemic. Online data collection is a means to expand the scope of research in both the number and diversity of informants (Mellinger, 2015; Rodd, 2024). In this case, informants from three top Chinese universities were recruited to participate in the study, and this would have been much more difficult in a face-to-face approach, which would require travel and, more importantly, that all informants coincided not only in time but also in space. Furthermore, compared to lab-based experiments, remote data collection allows broader possibilities for researchers to replicate and reproduce the study. Informants were instructed to complete tasks in a quiet environment without disruptions, to avoid breaking the task flow with distractions.



第三轮同传测试

注意：不管选用哪种录屏软件，都需要事先戴上耳机，测试是否能正常收音，不要录下系统原声，可以结合音频播放测试页面来进行，试着播音，自己说几句，看下录音效果。Interpretbank 组和 Excel组的同学均采用以下规则： 参加完腾讯会议，即可开启录屏软件和keylogger 完成 打开给定的音频播放页面 （单纯播音，不具备录音功能） 根据给定的术语表，30分钟译前准备 不超过15分钟的英→中同传， 同传结束后有10秒时间用于补充信息 返回的文件： 如果术语表内容有修正，需要提...

Du Zhiqiang 杜智强
1 min read

第三轮术语测试

Du Zhiqiang 杜智强
1 min read

booth task (Excel group-第二轮)

Du Zhiqiang 杜智强
1 min read

Figure 16. Project website.

A dedicated website was created to disseminate information on this project to all informants alike (see **Figure 16**).[14] This website worked as a central information hub for announcements and updates, ensuring that all informants were informed and updated about aspects of the project implementation. The website included a comprehensive introduction to the project, a detailed timeline outlining each

[14] Project website: https://dzq1007.hashnode.dev/

round, requirements for the data-collection apparatus, instructions for installing data-collection applications, and clear guidelines for session procedures. Informants were encouraged to regularly consult the website to stay abreast of the latest developments and ensure smooth and similar participation in the project. This digital resource was instrumental in facilitating effective communication and coordination among participants.

A stable internet connection was a prerequisite. Following detailed instructions made available on the website, the informants were guided to set up their individual data-collection environment (e.g., keylogger (Pynput) and screen recorder (TechSmith Capture)). This preparation involved equipping their PCs with Pynput and TechSmith Capture, ensuring that these tools were operational for all tasks. Before each task, informants were required to activate Pynput and TechSmith Capture. Additionally, specifically for the booth tasks, they needed to turn on the audio recording function within TechSmith Capture to register their own renderings. The consistent use of the keylogger and TechSmith Capture software was crucial across tasks to ensure reliability in data collection.

The analysis faced challenges due to informants being in different remote locations, leading to variations in the launch of screen recording and keystroke logging tools. We emphasized the importance of maintaining honesty and ethical conduct among informants and, in fact, no anomalies were found in the data from these tools. However, the geographical dispersion of informants caused slight timing differences in launching screen and keystroke recordings, which will be discussed further in the upcoming sections. Screen movements of informants in both glossary and booth tasks were captured using TechSmith Capture and then either uploaded to MS Stream or transcribed with millisecond precision timestamps (see **§ 2.4.1**) or machine transcription with millisecond precision timestamps. We then reviewed the transcripts from MS Stream and edited the transcripts while replaying and listening to the recorded video files. Of course, in remote data collection, there are certain confounders beyond our control. For instance, data collection heavily relies on web-based communication, and unexpected glitches or delays may occur. We asked informants to report any glitches or delays they might notice. Also, since data analysis depends on timestamps at the millisecond level, results are sensitive to these time measurements. Although the differences in screen refresh rates between different PCs are minimal, the potential for micro errors cannot be ignored. Variations in web browsers might also cause delays in the playback of MS Stream-based source speeches.

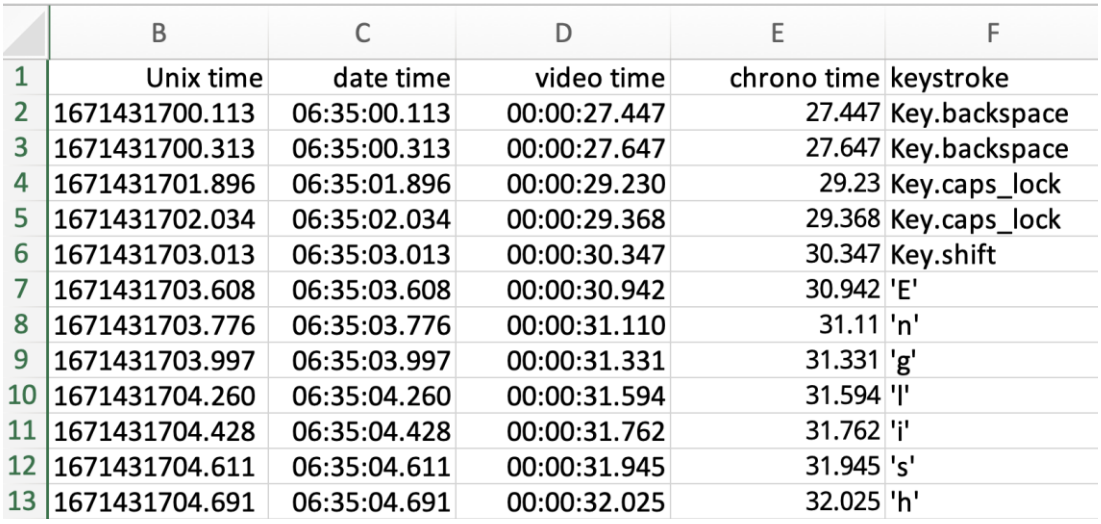
2.7 Data analysis

The raw data from individual behavioral logs collected online had to be cleaned and synchronized into a singular, multimodal timeline for future comparison and discussion. Raw data was gathered from screen recording files featuring SI rendering audio from TechSmith Capture, keystroke logs from Pynput, transcripts from informants' SI renderings, assessment results from raters, and survey responses.

2.7.1 Data cleaning procedures

To further humanize our approach and respectfully portray our informants when presenting results, we assigned neutral nicknames alphabetically ordered, ranging from Alex to Lee (A to L) for InterpretBank informants, and from Morgan to Val (M to V) for Excel informants. This step underscores our commitment to treating informants not merely as data sources but as individuals whose contributions are crucial to our research.

2.7.1.1 *Procedures for glossary task* The final step to cleaning raw data from glossary tasks entailed compiling all the information, including video time, keystroke, behavior, environment, application/service, function, keywords, keywords type and notes onto a single Excel sheet. Based on the conversion steps explained in **§ 2.4.2**, the Unix timestamps in keystroke logs were converted into date time, as shown in **Figure 17.**[15]



	B	C	D	E	F
1	Unix time	date time	video time	chrono time	keystroke
2	1671431700.113	06:35:00.113	00:00:27.447	27.447	Key.backspace
3	1671431700.313	06:35:00.313	00:00:27.647	27.647	Key.backspace
4	1671431701.896	06:35:01.896	00:00:29.230	29.23	Key.caps_lock
5	1671431702.034	06:35:02.034	00:00:29.368	29.368	Key.caps_lock
6	1671431703.013	06:35:03.013	00:00:30.347	30.347	Key.shift
7	1671431703.608	06:35:03.608	00:00:30.942	30.942	'E'
8	1671431703.776	06:35:03.776	00:00:31.110	31.11	'n'
9	1671431703.997	06:35:03.997	00:00:31.331	31.331	'g'
10	1671431704.260	06:35:04.260	00:00:31.594	31.594	'l'
11	1671431704.428	06:35:04.428	00:00:31.762	31.762	'i'
12	1671431704.611	06:35:04.611	00:00:31.945	31.945	's'
13	1671431704.691	06:35:04.691	00:00:32.025	32.025	'h'

Figure 17. Conversion table from Unix time to chrono time.

Next, keystrokes had to be identified and aligned with their corresponding video times in screen recordings. *Video time* refers to the timestamp shown in a video file, measured in *hours: minutes: seconds. milliseconds.* Zooming in frame by frame to the millisecond using Adobe Premiere Pro—a popular video editing application—it was possible to determine, for instance, that that the keystroke *E* in the Excel cell F7, appeared at video time 00:00:30.942. To enhance alignment precision, we used the Inter Keystroke Interval (IKI), the time between consecutive keystrokes. By applying the IKI data, Excel calculates the exact timestamps for each keystroke event, aligning them with the screen recording's timeline (i.e.,

[15] The deployment steps can be found in https://zenodo.org/doi/10.5281/zenodo.10520423

video time). However, calculating video times is not straightforward; for example, we can determine the accumulated seconds spent on specific events like online translation searches, but calculating the percentage of time dedicated to translation searches relative to the total glossary task duration is challenging due to potential errors caused by cell format changes. To mitigate this, we introduced an additional time layer, which we dubbed *chrono time*, which converts video times into *second.milliseconds* format. For instance, an event appearing at 00:01:32.025 in the video corresponds to 92.025 s as a single floating-point number in *second.milliseconds* format.

With Adobe Premiere Pro, the researcher annotated each behavior observed in the screen recordings for each informant with a video-time timestamp. This annotated information was then added to the spreadsheet (see also **§ 2.7.2.1**). Following the aggregation of this spreadsheet data, observed behavior events and keystroke events have been synchronized with the video timestamps. We were then able to proceed to a qualitative content analysis of the compiled information. The aim here was to distill and standardize the most significant elements (e.g., the time spent using a particular tool or the frequency of consulting specific services) to facilitate a quantitative analysis of various information-seeking behaviors in glossary tasks.

2.7.1.2 *Procedures for booth tasks*

Data cleaning for the booth tasks involves a two-tiered approach to behavior log data alignment: the first layer refers to individual data alignment, followed by alignment across informants. Our initial focus centers on individual-level alignment.

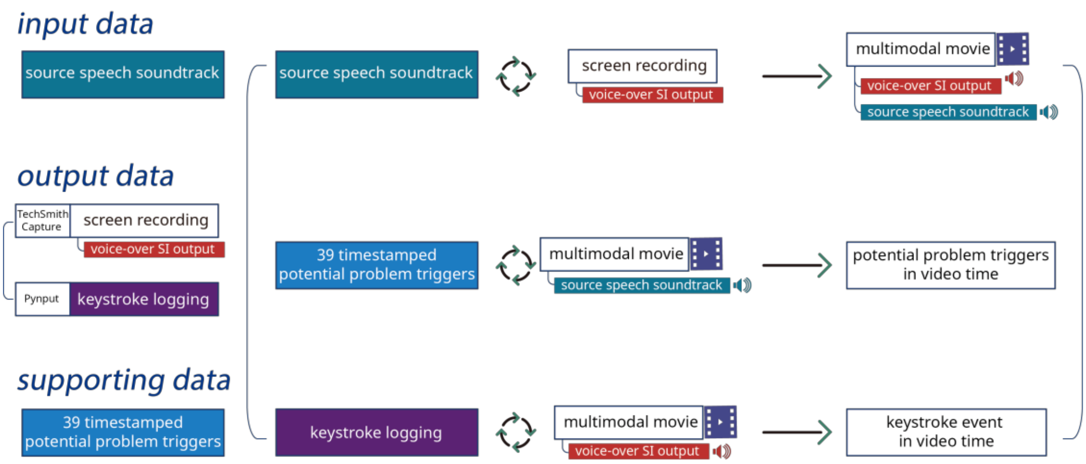


Figure 18. Individual multimodal data synchronization.

In contrast to glossary tasks, individual booth task data cleaning entails the overall synchronization of multimodal raw data. All elements for each informant were compiled into a spreadsheet in an Excel file. The data cleaning process was divided into two sections: (1) synchronization and (2) compilation of multimodal elements along the same timeline (video time). In **Figure 18**, we present a comprehensive view of individual multimodal data synchronization. The input data comprises the *source speech soundtrack*. The output data consists of screen recordings captured using TechSmith Capture, along with the corresponding SI rendering audio recordings, and keystroke logging data detailing every stroke event during the booth task. Additionally, these 39 problem triggers were timestamped as supporting data derived from the transcript of the *source speech soundtrack* (see **§ 2.2.3**). The problem triggers captured how informants responded to these triggers when using InterpretBank. The rationale behind these alignments is that a single data type alone cannot provide a comprehensive narrative. For example, each single keystroke event lacks context. Without contextual information about which dictionary website informants are using, we cannot make informed statements about their behaviors.

The alignment is three-tiered. First, the source speech soundtrack is aligned with the *screen recording*. This leads to two aligned files dubbed*:* one with SI output audio, and the other with the source speech soundtrack. To investigate how informants in the InterpretBank group address problem triggers, we aligned the 39 timestamped problem triggers (for the selection of potential problem triggers in **§ 2.2.2**) and the *multimodal movie with the source speech soundtrack*. This alignment allowed us to identify terms present both in the source speech soundtrack and the 39 timestamped problem triggers, which were then synchronized in the video timeline. The last alignment involves synchronizing keystroke logging data with another *multimodal movie with voice-over SI output* audio, ensuring that each keystroke event is represented in the video timeline. Next, we will provide a detailed explanation of how to proceed with these three alignments.

Firstly, we aligned the *source speech soundtrack* with the screen recording (see **Figure 19**). This process starts by importing the *screen recording* video and the *source speech soundtrack* into Adobe Premiere Pro (see **Figure 20**). This application offers several convenient features for data cleaning in this study: it enables precise zooming into the video at millisecond-level accuracy, allows annotations at selected timestamps, and adds a new soundtrack to the video file. Moreover, it flexibly adjusts the soundtrack placement, ensuring alignment with the screen recording video.

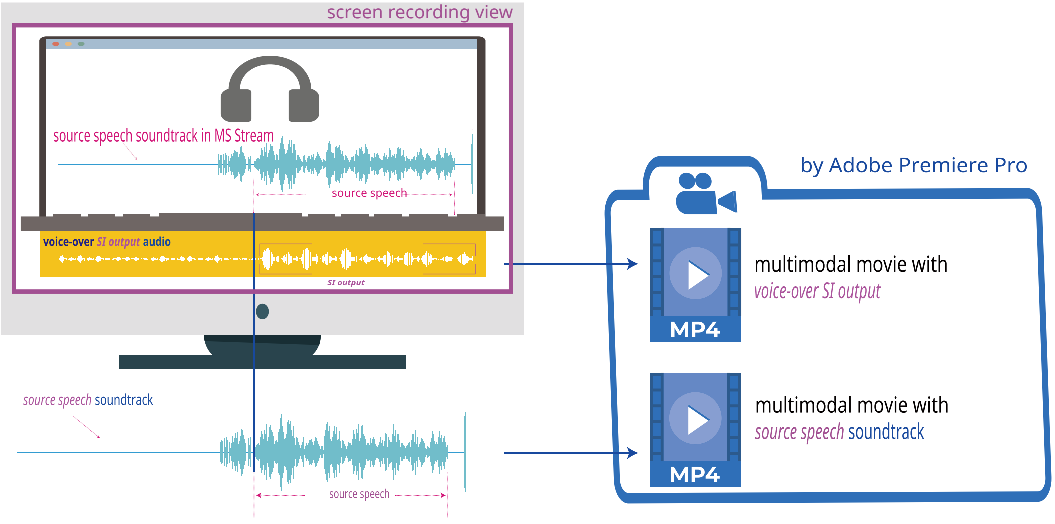
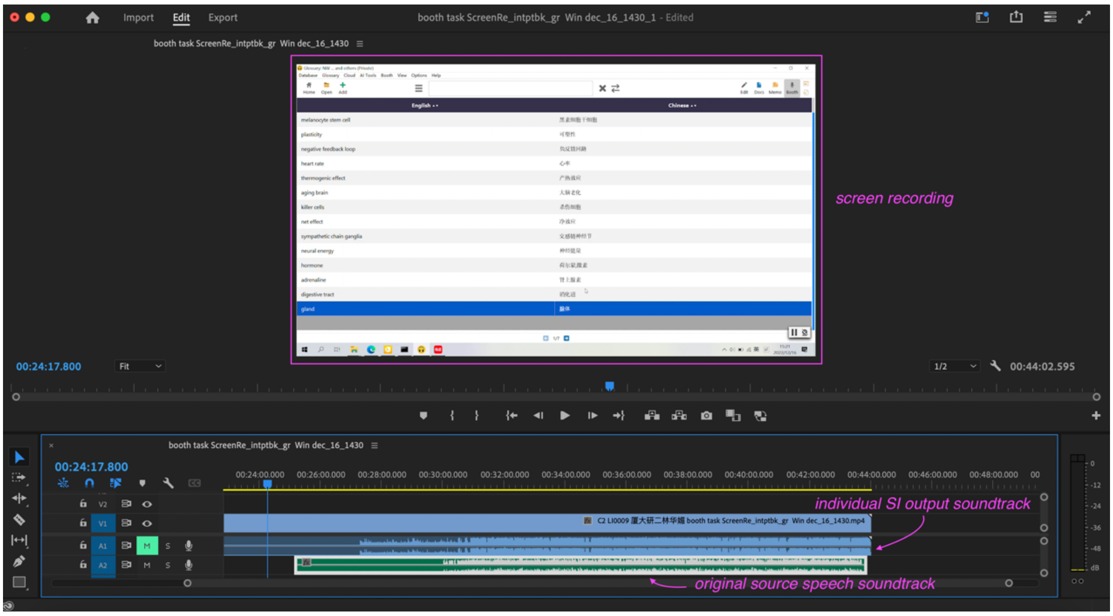


Figure 19. The first alignment workflow.



booth task ScreenRe_intptbk_gr Win dec_16_1430_1 - Edited
screen recording
00:24:17.800
00:44:02.595
individual SI output soundtrack
original source speech soundtrack

English	Chinese
melanocyte stem cell	黑素细胞干细胞
plasticity	可塑性
negative feedback loop	负反馈回路
heart rate	心率
thermogenic effect	产热效应
aging brain	大脑老化
killer cells	杀伤细胞
net effect	净效应
sympathetic chain ganglia	交感链神经节
neural energy	神经能量
hormone	荷尔蒙激素
adrenaline	肾上腺素
digestive tract	消化道
gland	腺体

Figure 20. Importing Video and Audio into Adobe Premiere Pro.

There are multiple options for anchoring this alignment process (strategies A, B, and C) and the choice should be based on each informant's individual circumstance. Strategy A involves identifying the start of the *source speech soundtrack* in the *screen recording* in Adobe Premiere Pro. This can be done by observing the informant's action as they click to play the speech in MS Stream where the *source speech soundtrack* is located. If this event was visible in the screen recording, the timestamp or video frame at this moment was annotated in Adobe Premiere Pro as the anchor point. The process then entails a frame-by-frame examination to pinpoint the exact millisecond. This may reduce the delay caused by confounding variables, establishing a precise anchor point. Subsequently, the beginning of the *source speech soundtrack* is mapped to the timestamp of this anchor point.

If the click-and-play of the original source speech in the web player is not visible in the screen recording—e.g., because the informant focused on the task and launched TechSmith Capture after clicking the playback in MS Stream—alternative strategies (B and C) are available to identify another suitable anchor point. Strategy A relies on frame observation, while strategies B and C are based on auditory features. In some instances, the *source speech soundtrack* may be faintly audible in the *screen recording*, possibly due to headset settings or computer audio output configurations. This faint signal, though minimal, can be sufficient for the precise positioning of an anchor point. In strategy B, we listened to the screen recording (i.e., the voice-over *SI output* audio) to pinpoint a distinct word from the *source speech* at its onset timestamp. This word is then used as a new anchor point, aligning it with the corresponding moment in the *source speech soundtrack* in Adobe Premiere Pro. Strategy C is similar to B, but it focuses on identifying signal sounds (as detailed in **§ 2.2.3**) within the individual screen recording. Once these sounds are located, they are mapped to their respective positions in the *source speech soundtrack* using Adobe Premiere Pro.

Fortunately, these three strategies covered all the cases observed in booth tasks, ensuring alignment of the *source speech soundtrack* with the *screen recording*. This re-aligned screen recording or multimodal movie includes two audio tracks: the sound of the screen recording (i.e., the voice-over SI output) and the *source speech soundtrack*, which started the video timeline.

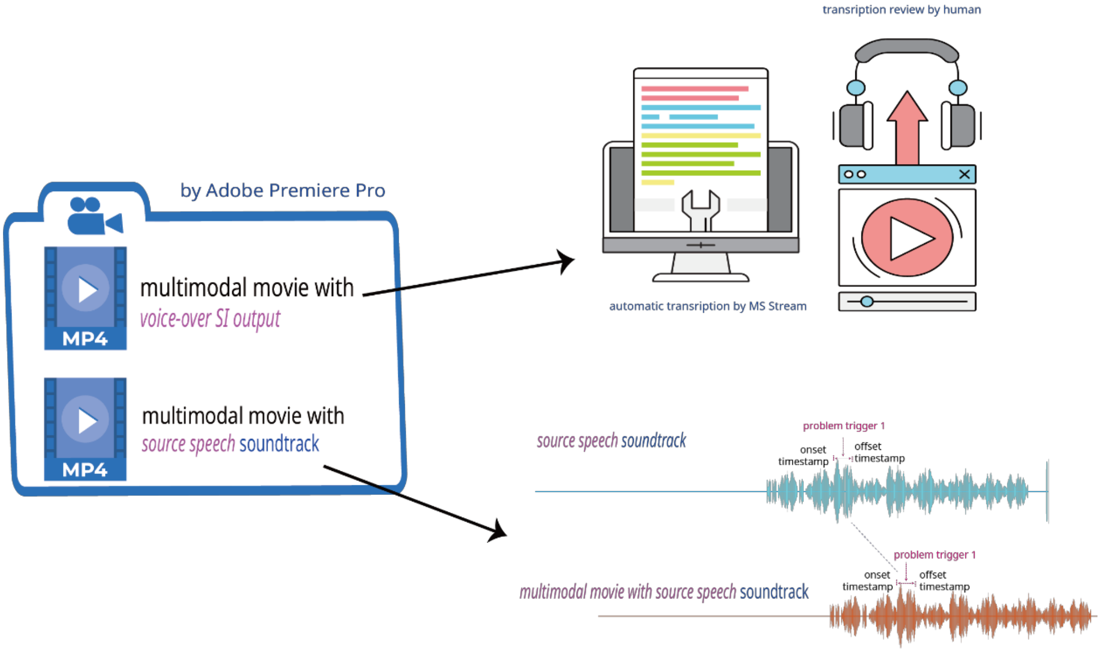
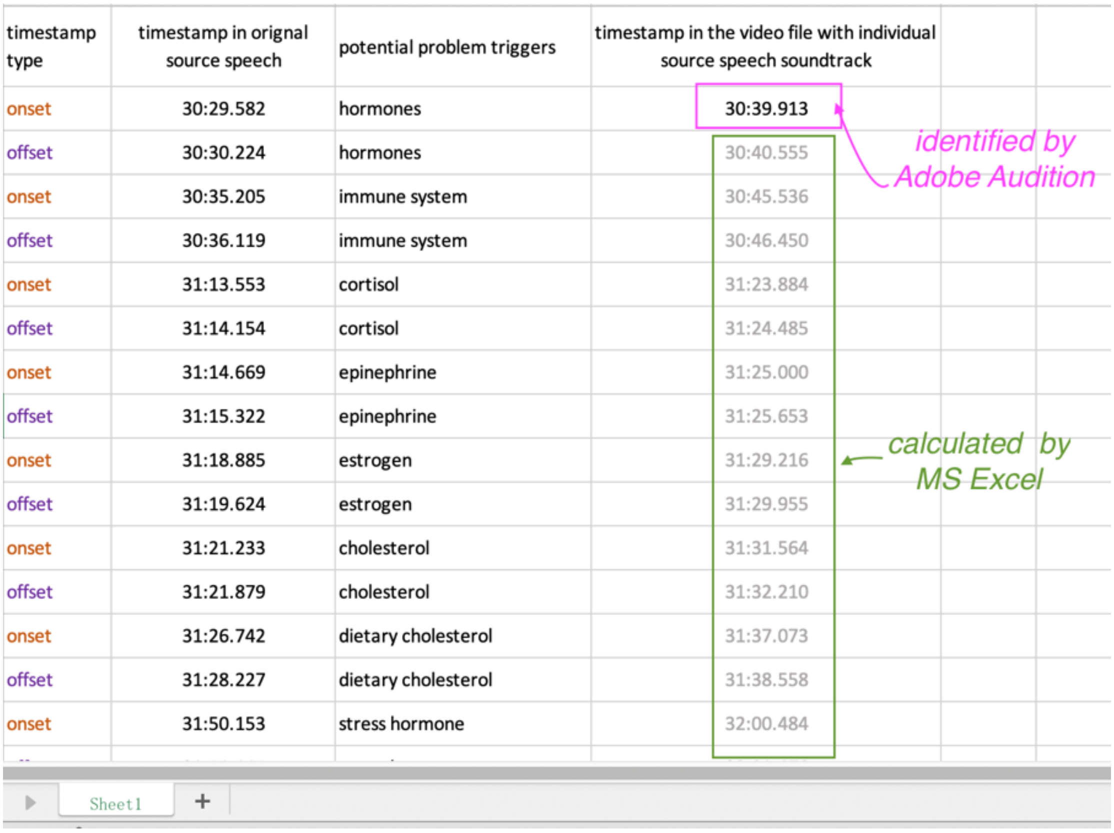


Figure 21. The second alignment workflow.

After the last step, we proceeded to the second alignment (see **Figure 21**). Adobe Premiere Pro was used to export two multimodal movies: one with voice-over *SI output* and the other containing the *source speech soundtrack*. The video file with the voice-over *SI output* audio was uploaded to MS Stream for automatic transcription of the SI renderings. This auto-generated transcription is automatically segmented based on the pauses in the SI rendering, and the transcription requires human verification and adjustments for accuracy. This process then involves reviewing and modifying the transcription within MS Stream, and then aligning these changes with the play and pause functions to ensure precision. Finally, the transcript with timestamps for each segment was exported, and these segments were adjusted to correspond with the chunks in the original source speech script (see **§ 2.2.3**).

Another aspect of the second alignment process involves identifying potential problem triggers within the video timeline of the multimodal movie with the *source speech soundtrack*. To achieve this, we used Adobe Audition 2024, a professional digital audio workstation application known for multitrack recording, mixing, precise editing, and audio mastering. We imported the *multimodal movie with the source speech soundtrack* into Adobe Audition 2024. In tandem with this, we created an Excel spreadsheet listing 39 terms, including three groups of repeated terms selected as potential problem triggers, as detailed in **§ 2.2.2**. Each term in the spreadsheet was marked with onset and offset timestamps, noting that these timestamps specifically refer to those within the *source speech soundtrack*. Adobe Audition was then utilized to determine the onset timestamp of the first trigger in the *multimodal movie with the source speech soundtrack,* aligning it with the corresponding timestamp in the *source speech soundtrack*. This process is similar to the use of IKI for calculating the timestamps of consecutive keystrokes in **§ 2.7.1.1**. Excel was employed to calculate precise timestamps for both the onset and offset of each problem trigger, aligning them with the timeline of the *multimodal movie* (i.e., video time). In other words, the Excel spreadsheet computed onset and offset timestamps for all triggers present in the *multimodal movie with the source speech soundtrack*, based on the observed time differences (see **Figure 22**).



timestamp type	timestamp in orignal source speech	potential problem triggers	timestamp in the video file with individual source speech soundtrack
onset	30:29.582	hormones	30:39.913
offset	30:30.224	hormones	30:40.555
onset	30:35.205	immune system	30:45.536
offset	30:36.119	immune system	30:46.450
onset	31:13.553	cortisol	31:23.884
offset	31:14.154	cortisol	31:24.485
onset	31:14.669	epinephrine	31:25.000
offset	31:15.322	epinephrine	31:25.653
onset	31:18.885	estrogen	31:29.216
offset	31:19.624	estrogen	31:29.955
onset	31:21.233	cholesterol	31:31.564
offset	31:21.879	cholesterol	31:32.210
onset	31:26.742	dietary cholesterol	31:37.073
offset	31:28.227	dietary cholesterol	31:38.558
onset	31:50.153	stress hormone	32:00.484

Figure 22. Partial alignment list showing timestamps using MS Excel.

The third alignment involved aligning keystroke events with *multimodal movie with voice-over SI output* audio (see **Figure 23**). In a fashion similar to data cleaning in the glossary tasks (see **§ 2.7.1.1**), keystroke data, including Unix timestamps, were imported into an Excel spreadsheet. These Unix timestamps were pre-processed and converted to a video time format, annotated as hh:mm:ss.000. Then, we imported a *multimodal movie with SI output* into Adobe Premiere Pro to identify a single keypress event of a visible character in the screen recording video. The character chosen as an anchor point for this step needs to be both in the keystroke log and within the screen recording, and the timestamp must correspond with the exact moment it appeared in the video. To enhance alignment precision, we used the time between consecutive keystrokes or IKIs as we did in the glossary task.

Using the IKI data, Excel calculated the exact timestamps for each keystroke event, aligning them with the video timeline. This process results in recoding prior keystroke timestamps into new, tailored keystroke timestamps, now synchronized with the video.

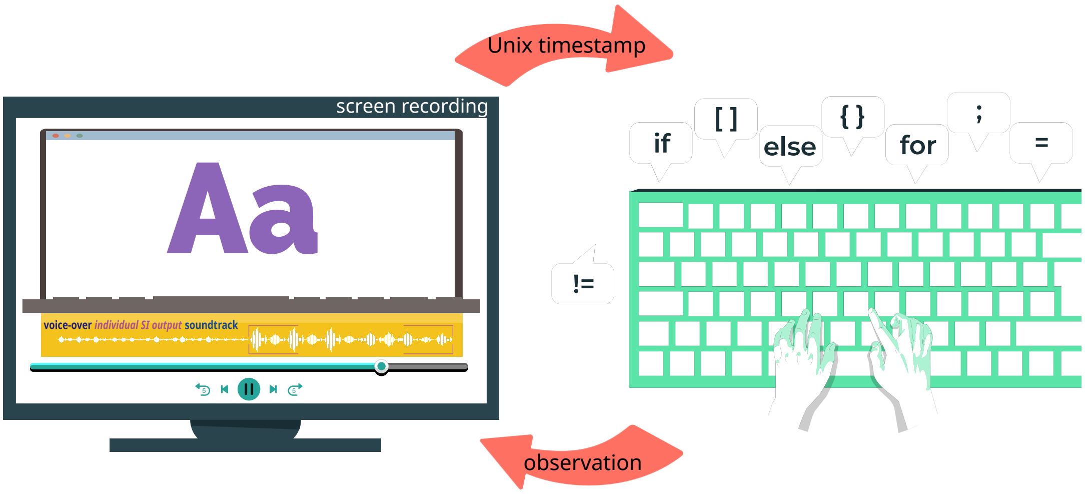


Figure 23. The third alignment workflow.

So far, for each informant, we had successfully finished three alignments. All the multimodal data from each informant including source speech, video recordings, the SI rendering transcripts, keystroke events, and potential problem triggers had been converted or shown in the same video timeline. In other words, all collected data from primary informants had been compiled into a single spreadsheet, sharing a unique timeline specific to each informant. Now it could be organized chronologically for further analysis, and it was easy to visualize and understand. It consists of sequential behavior observations from different tools, and it eases direct temporal links between expected behaviors across different modalities. For instance, by examining the spreadsheet, we can observe the offset timestamp for a specific problem trigger, followed by a term retrieval action. The informant's typing—whether the letter sequence is accurate or includes corrections—is also clearly recorded and available for analysis. This integration provides rich contextual information, allowing nuanced inferences and hypotheses about informants' behaviors. The chronological data structure aids in developing behavioral redescriptions, which will be discussed in the following sections. Besides, the video time is expressed in *hh:mm:ss. milliseconds* (00:00:00.000) can be easily converted into just *seconds. milliseconds* (ss.000) for easy computations.

Now, we progressed to the second level of data cleaning, addressing asynchronicity in log data from different informants within a single cycle. This asynchronicity may arise from differences in informants' geographical locations and the varied sequence in which data collection tools are launched before booth tasks. Synchronizing this log data in a universal timeline is vital for comparing performance within groups. For example, to analyze the usage duration and timing of InterpretBank by its group members, a universal timeline is a must.

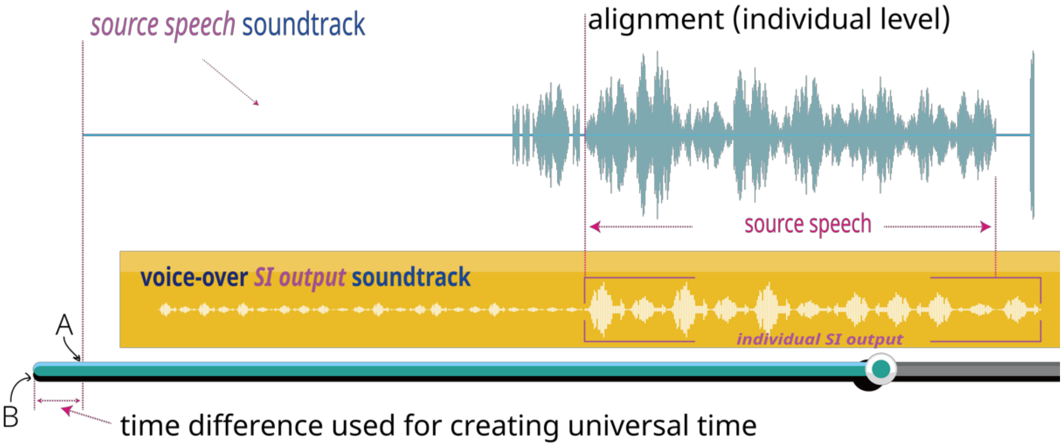

source speech soundtrack
alignment (individual level)
source speech
voice-over SI output soundtrack
individual SI output
A
B
time difference used for creating universal time

Figure 24. Computation for universal time.

In the prior (individual) level of alignment, we obtained two *multimodal movies* with the same timeline but different soundtracks: the *source speech soundtrack* and the *SI output soundtrack*. For clarity, these two soundtracks are combined and illustrated in **Figure 24**. Due to varying starting times of SI renderings across informants, the lengths of the SI output soundtracks differ, but the length of the source speech soundtrack, which shares the timeline with the SI rendering, is consistent across informants. To synchronize the timelines, we first identified the first timestamp of the source speech soundtrack (referred to as Point A in **Figure 24**) and the zero time at the start of the video (referred to as Point B in **Figure 24**). By calculating the time difference between Points A and B, we can adjust any event timestamp on this timeline. Subtracting this time difference from an event's timestamp in this timeline yields a new timestamp, aligning all informants on the same timeline starting from the source speech soundtrack. This method can generate a virtual universal time, ensuring all informants are synchronized from the onset point. This simple computation is not required for all variables in the following data analysis but is applicable in specific situations, e.g., when considering *search durations* and *dropped chunks* (see **§ 3.4.4**).

2.7.1.3 *Punctual events and span constructs* We have described the strategies and techniques to synchronize all raw data into a single, unique timeline for each informant. Our record contains aligned behaviors, each one of them depicting a different angle of a complex action, so it can be called an *ethogram*. The term is not widely used in CTIS but it is more popular in ethology and it is also used in ethnographic tools, such as BORIS.[16] Since our ethograms include information from

[16] BORIS: https://www.boris.unito.it/

different language modes, they are multimodal. Hence, we will call them *multimodal ethograms*.

Based on the synchronized data, the next step is to code the behavior for analysis in glossary and booth tasks. *Punctual events* are actions occurring at specific time points that have a minimal duration rather than extending over temporal intervals. These events can be thought of as snapshots in time, capturing a single moment of behavior. On the other hand, a *span construct* refers to behaviors that occur over a longer period that can be understood as continuous behaviors between two *punctual events*. The data coding was based on the assumption that when at least two such punctual events are understood to be related to each other, they can be described as behavioral spans, which calls for a set of span constructs (see **Figure 25**).

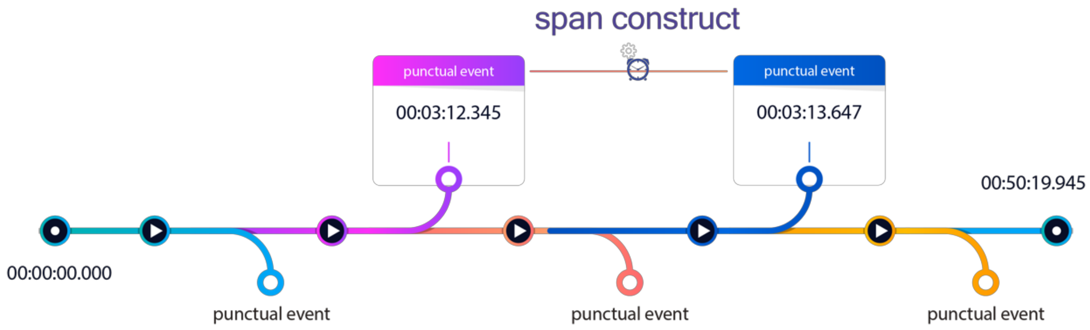


Figure 25. Representation of punctual events and span construct.

To study the impact of CAI tool use from a cognitive situated perspective, the distinction between punctual events and span constructs is important. Both punctual events and behavioral spans may be studied in relation to each other, but spans easily lend themselves to be studied as hosting punctual events, very much like the notions of *figure* and *ground* work in the Gestalt psychology (Peterson & Skow-Grant, 2003; Cacciamani *et al.*, 2012). A punctual event may be studied in isolation, without considering the neighboring events where it is interspersed or juxtaposed. Two contiguous or overlapping punctual events may be understood as serial or parallel depending on the adopted time frame. For instance, simultaneous interpreting is so considered because we do not analyze the task in terms of milliseconds when the different subtasks might be thought of as sequential, rather than simultaneous or isochronic. This is *chunking,* the procedure of adding elements into higher units or else breaking a continuous flow into smaller parts. Here the researcher used the first (synthetic) approach, which we will call *behavioral redescription* (Muñoz, personal communication). It is not new. It is only explicit, to foster rigor.

Behavioral redescription aims to build higher, meaningful constructs out of registered data, to allow for richer analyses. When a pair of punctual events are interpreted as thresholds defining a span, other punctual events occurring within such thresholds can be interpreted as happening simultaneously along the continuous action defined by the behavioral span. Behavioral spans may also partially or totally overlap with other behavioral spans. This allows for a more varied analysis than temporal and contingency relations. Overall, this approach of detailed behavioral redescription fosters reproducibility, and it also provides a better fit for the agentive, interactive, and multitasking view of interpreters in cognitive translatology. Of course, the interpretation of data as either punctual events or continuous behavioral spans depends on the way such recorded and coded behaviors are understood and defined. Different researchers may interpret one and the same recorded behavior in different ways, depending on their own perspectives and biases toward informants' information-seeking behaviors. We have striven to maintain an interpretive modesty and hence sought to create spans only when an intersubjective agreement seemed very likely. Data is available upon request.

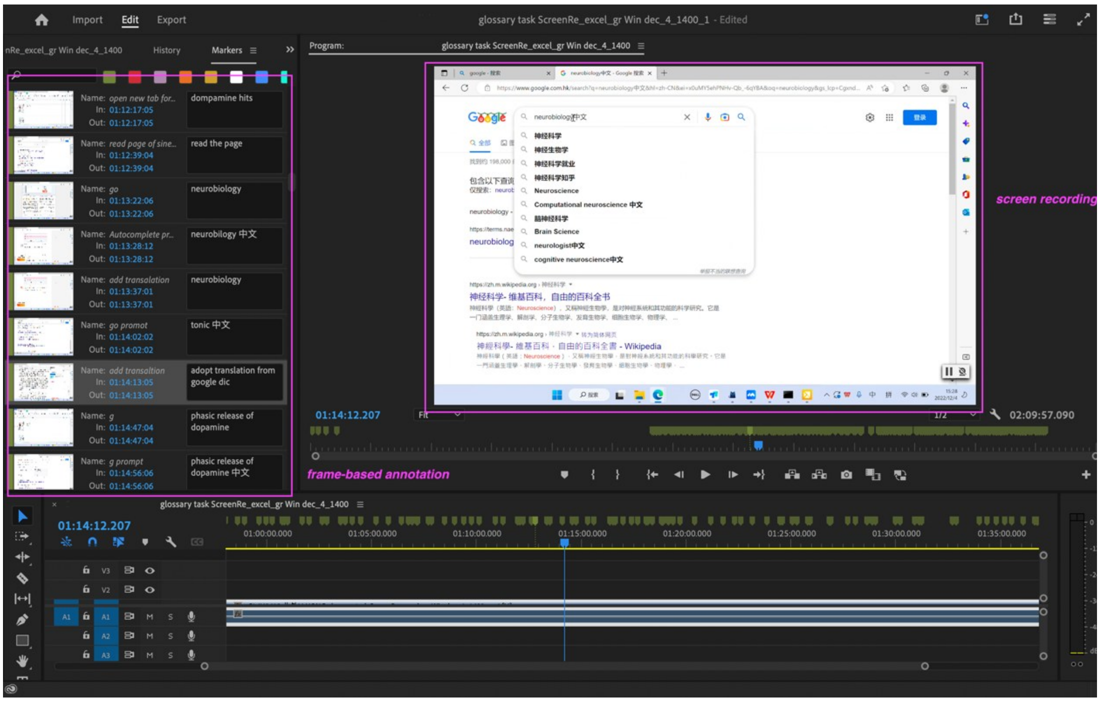
Time gaps—periods where no behavior, such as mouse movements or switching windows, was registered in the *ethogram*—can be considered spans too, between two registered events in keystroke logging and screen recording. Even though it is not possible to directly know what was happening in the informants' minds during these spans, inaction may be meaningful as well. Depending on surrounding hints, *time gaps* may be interpreted as related to confusion, uncertainty, mind wandering, or anxiety. In the present study, *time gaps* in glossary tasks were coded as blanks and were thought to potentially provide insights into the informant's cognitive activities.

Time gaps in interpreting output—together with fillers, re-starts, and the like—are traditionally approached as disfluencies, thereby ignoring that some of them may respond to intentional, strategical behaviors by interpreters. For instance, not all *fillers* need to be mental blackouts. Some pauses are rhetorical, and not all re-starts need to be *false starts*. We attempted to describe observations without pre-judging their causes.

In particular, as in the Task Segment Framework (Muñoz & Apfelthaler, 2021, 2022) for written translation analysis, the number and length of these *time gaps* or *blanks* are hypothesized to change with cognitive effort in tandem with coded behaviors. A larger pool of data may lead to identifying regularities in span lengths and positions for individual informants, perhaps even general tendencies. This is beyond the reach of this project, but it is mentioned because we claim that blank periods should not be dismissed as insignificant, but rather may become valuable sources of information about the informants' cognitive activities.

The other, symmetrical chunking strategy—breaking... and the like—has breaking down behaviors identified at a high level of abstraction, such as importing and exporting documents, technical troubleshooting, and the like, has been noted but not yet broken down into more granular components because, again, there were beyond the reach of this project, which targets other goals. Hence, some behaviors identified during the study were coded as *blanks, bumps*, and *respites*. They are defined and discussed in **§ 2.7.2.2**, and might provide rich, additional information on cognitive processes.

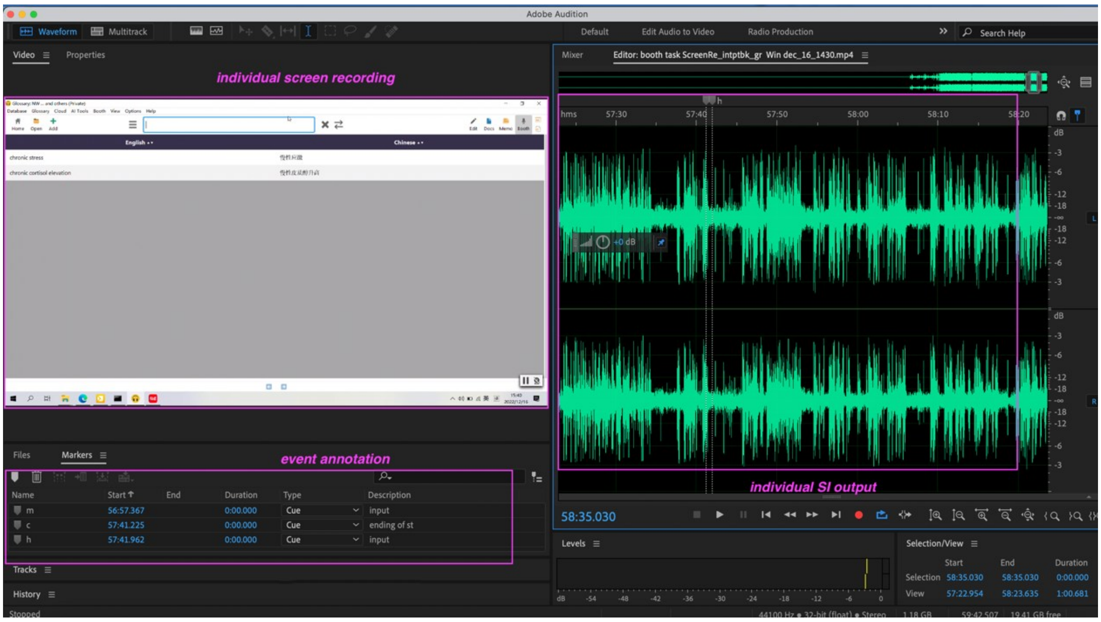
2.7.1.4 *Coding behaviors* Transcribing audio data is relatively straightforward and now often semiautomatic. In contrast, screen-recorded video data mainly consists of non-verbal information, such as cursor movements and software operation, which demand that coding be performed manually and in several steps.



Import
Edit
Export
glossary task ScreenRe_excel_gr Win dec_4_1400_1 - Edited
History
Markers
Program:
glossary task ScreenRe_excel_gr Win dec_4_1400
screen recording
frame-based annotation
01:14:12.207
02:09:57.090

Name: *open new tab for...* In: 01:12:17:05 Out: 01:12:17:05	dompamine hits
Name: *read page of sine...* In: 01:12:39:04 Out: 01:12:39:04	read the page
Name: *go* In: 01:13:22:06 Out: 01:13:22:06	neurobiology
Name: *Autocomplete pr...* In: 01:13:28:12 Out: 01:13:28:12	neurobilogy 中文
Name: *add transalation* In: 01:13:37:01 Out: 01:13:37:01	neurobiology
Name: *go promot* In: 01:14:02:02 Out: 01:14:02:02	tonic 中文
Name: *add transaltion* In: 01:14:13:05 Out: 01:14:13:05	adopt translation from google dic
Name: *g* In: 01:14:47:04 Out: 01:14:47:04	phasic release of dopamine
Name: *g prompt* In: 01:14:56:06 Out: 01:14:56:06	phasic release of dopamine 中文

Figure 26. Coded behavior in Adobe Premiere Pro for glossary task.



Name	Start	End	Duration	Type	Description
m	56:57.367		0:00.000	Cue	input
c	57:41.225		0:00.000	Cue	ending of st
h	57:41.962		0:00.000	Cue	input

Figure 27. Workspace in Adobe Audition 2024.

In order to categorize and code behaviors to study glossary compilation, the videos were first imported and tagged using Adobe Premiere Pro to ensure millisecond-level accuracy (**Figure 26**). This was the basis for segmenting the behavioral flow into discrete units. A series of behavioral records were created in Adobe Premiere Pro, based on punctual events and span constructs (see **§ 2.7.1.3**), such as *reading the source text, searching for a translation*, and *selecting terms for a glossary*. All the tagged events were exported and saved in a file with the .csv format. This made it possible to integrate them with keylogging data, facilitating a comprehensive, quantitative subtask analysis, which will be explored in **§ 2.7.1.1**. Adobe Audition 2024 was employed when coding behaviors in booth tasks because it provides millisecond precision, essential for accurately tracking time-sensitive indicators like *ear-key span* and *eye-voice span,* introduced in the next section. The *multimodal movie with the source speech soundtrack* was analyzed on a frame-by-frame basis, to identify and tag each event (**Figure 27**).

A range of SI rendering quality indicators were introduced to describe informants' renditions, which fell into two (customary) dimensions, fluency and accuracy (reviews in Zwischenberger, 2010; Chen *et al.*, 2022; Han, 2022b). This was done with Audacity®, a free, open-source digital audio editor (see **Figure 28**). Audacity allows the precise coding of specific behaviors in a screen-recording video file in the form of the soundtrack with millisecond accuracy. This application offers an intuitive interface for creating multiple layers, with each layer designated for tagging specific codes. This highly customizable feature allows users to define and assign codes and durations for each event.

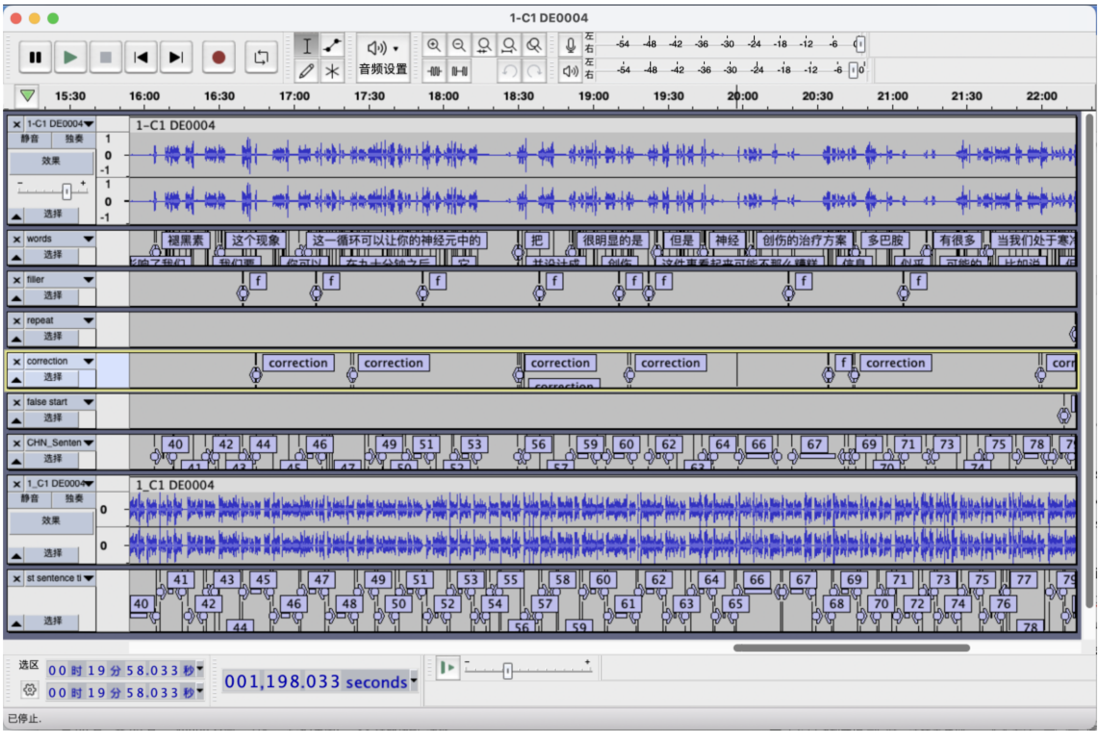


Figure 28. Layers for coding indicators in Audacity®.

Audacity® can automatically extract soundtracks from imported video files and can handle multiple audio tracks simultaneously. In our study, when a *multimodal* *movie with SI output* was imported into Audacity®, we supplemented it by adding the corresponding script of the SI renderings and importing the aligned source speech soundtrack. This allowed us to code each indicator related to fluency or accuracy dimensions by tagging it into different layers within Audacity® (as shown in **Figure 28**). By combining the functionalities of Adobe Audition and Audacity, this method of behavior coding proved to be highly precise and efficient for analyzing behaviors in the booth tasks. It significantly contributes to achieving an explicit, systematic, and transparent behavioral redescription.

2.7.1.5 *Ear-key span and eye-voice span* We developed or reformulated the constructs *ear-key span* (E2K) and *eye-voice span* (I2V) to examine the cognitive effort required during SI tasks with CAI tools, specifically involving the use of InterpretBank (see **Figure 29**) in the present study. These constructs are inspired by the notion of eye-voice span (Su, 2020; Zhou *et al.*, 2021; Chmiel & Lijewska, 2022) to explore the informant's interaction with the task. E2K is defined as the timespan between the end of the soundwave of the relevant source speech utterance (in this case, of the potentially problematic terminological unit, often plurilexical) to *keydown* of the informant's first related keyboard action. I2V is the timespan between InterpretBank displaying a term's translation on the screen to the beginning of the soundwave by the informant vocalizing the corresponding renditions. We assume that informants—working against the clock and expecting a word they need to appear on a certain area of the screen—tend to fixate on it after entering the last keystroke event, which could be retrieved from the Pynput log file. Together, these constructs are deemed adequate to capture behavioral and cognitive dynamics for the following reasons.

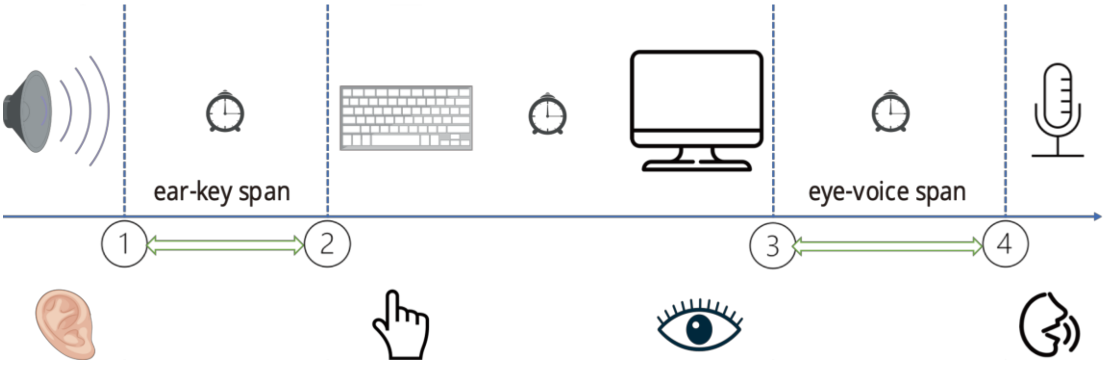


Figure 29. Representation of ear-key span and eye-voice span.

First, the E2K can be easily identified by examining the onset time, as flagged by the end of terms in the wave file of the source speech. The offset time, the moment of the first related keypress, can also be easily determined by examining the Pynput log file. E2K is taken to reflect the first millisecond when a full term can be identified. We will come back to this point when referring to the simultaneity of actions but let us for now enter a terminological note. We will use *predictions* as in recent publications instead of the more traditional term *anticipation* because we will be referring to both positive and negative timespans.

Second, E2K reflects the processing time between term identification, through memory and motor routine activation, to keyboard action. This indicator, which is both spatial and temporal in nature, can provide insights into the informants' situated cognition, mainly through human-computer interaction. Although the interval is very short, it reveals a temporal aspect of how the informants receive stimuli presented by the source speech and possibly increase their cognitive effort due to engaging in simultaneous actions against the clock.

As for I2V, we hypothesize that right after typing a letter, the user is likely to fixate on the screen to see the translations InterpretBank may provide. The time code for the frame showing an entry on the screen is *approximately* the same as the time the translation is uttered. The reader is reminded, once again, that there may be small measurement errors due to variations in screen refresh rates and frame rates of screen recording tools. The typical laptop screen refresh rate is 60Hz (times per second), i.e., 16.7 milliseconds, and the usual default frame rate of screen recording tools is 30 frames per second or 33.3 milliseconds. This is not too small an error, but it is sporadic and counterbalanced with the millisecond accuracy of the data from multiple sources. I2V is not only the time it takes for the informant to read and name the item presented by the machine. The processing latency involved in monitoring the SI delivery and interpreting the target terms by adequately inserting them into the ongoing flow will affect the I2V span.

Ideally, E2K and I2V spans should come in paired, linear sequences. However, not every instance conforms to this pattern. For example, an informant might complete an E2K event by pressing a key upon hearing a term and seeing its translation on the screen, but then opt not to engage in the corresponding I2V event by speaking out the translation. The reasons for dropping the verbal renditions may vary, such as encountering an unexpected translation pop out from InterpretBank search or simply being unable to handle the information, thereby reducing cognitive effort by skipping the rendition. Nevertheless, even an incomplete E2K-I2V sequence can provide meaningful data. More research is needed in this area to understand the underlying reasons and implications of such incomplete sequences. In a task tightly constrained by time, the E2K and the I2V spans may offer valuable insights into the cognitive effort exerted to perform particular actions.

The E2K and I2V spans have a typical temporal order, with E2K occurring before I2V. E2K represents the time from the audio signal to the keypress, so it is always a positive value since we measure the time lag from hearing the signal to pressing the key. On the other hand, I2V spans can be positive or negative. A positive I2V value means the suggested translation appeared on the screen before the utterance, while a negative I2V value means the utterance occurred before the suggested translation was displayed.

The E2K and I2V spans are measured to also assess spillover effects indirectly. Keys may be pressed before the potentially problematic term (which may consist of several words) is completely uttered in the source speech recording, and translations may start being uttered before they appear on the screen. The former scenarios suggest intents of prediction. In the E2K span, a term may become active in memory while listening to a few syllables. In the I2V span, the term might be active in memory and ready to be used, but it may require verification post-utterance. Reversed actions like typing before the term is complete (captured as negative E2K spans) could influence subsequent behaviors due to higher multitasking efforts, compared to SI delivery without searches, possibly affecting the length of the I2V span. These observations allow us to infer the extent of spillover or preprocessing effects, but the scope of this study limits a direct measurement of these phenomena. In any case, these indirect measures might provide valuable insights into the dynamics and consequences of prediction in SI.

2.7.2 Indicators

The indicators for glossary and booth tasks in this project range from intra-subject analysis to inter-group analysis, but the emphasis was on intra-subject analysis. A range of indicators from different dimensions were chosen to circumvent Simpson's Paradox (Carlson, 2024), a statistical issue where combining different groups of data can reverse observed trends, potentially skewing the overall interpretation of results. By using a diverse set of time-related or language production (content)-related indicators (discussed in **§ 2.7.2.2**), we hope that the analysis is truly reflective of actual patterns and prevents such distortions. Although the time cluster indicators belong to fluency, there is a certain parallelism between the differences between accuracy and fluency, and between content and time clusters: fluency and time cluster indicators depend on the rendering, whereas accuracy and content indicators codify the relationship between source speech and informants' renderings (rendition by rendition). Fluency indicators seem to tend to be source speech independent and this may happen sometimes with EVS1 and EVS2 that may be caused by reasons other than cognitive demands. However, EVS measures depend on source speeches to calculate their magnitudes.

2.7.2.1 *Indicators in glossary tasks* Informants were tasked with extracting unfamiliar terms from the source texts and compiling their own glossaries. Screen movements and keystroke events were registered (see **§ 2.7.1.1** and **§ 2.7.1.4**) and each event was coded as a discrete or punctual activity or else as (part of) a span construct , or continuous activity (see **§ 2.7.1.3**). Prior research (Onishi & Yamada, 2020; Lu *et al.*, 2022; Enríquez & Cai, 2023; Gough, 2023) primarily examines how translation students utilize online resources. Yet studies are scarce that examine how interpreting trainees manage novel terms extraction and seek their translations during glossary preparation, incorporating both local tools and online resources. Throughout this glossary task, informants process multimodal data, for instance, source texts, pronunciation audios from online dictionaries, and images from Google search results. To understand task-related problems and behaviors, we set to analyze them in standard ways, but human behaviors are complex, interactive, and context-dependent.

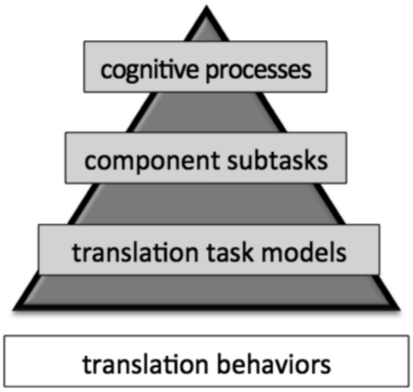


Figure 30. Layered approach to task models (from Muñoz, 2014).

To capture and reflect on the informants' overall performance in glossary tasks, we applied Muñoz's (2014) layered approach to task models. This application went beyond merely breaking down the informants' information-seeking process into observable behaviors. Rather, we tagged these behaviors with various constructs and spans that are indicative of translation behaviors in Muñoz's model (see **Figure 30**, reproduced with permission from Muñoz, 2014, p. 13). Our analysis included extracting subtasks and summarizing them to depict informant behavior in cognitively demanding, naturalistic environments. For example, these actions consist of operations such as

1. *IB- search the term in IB,*
2. *open a new tab for reading*,
3. *paste the term into the search box*.

The behaviors are grouped and refined and behaviorally redescribed into several strategies, understood as components of subtasks. For instance, *searching the term in IB* is labeled as ***source-text term retrieval****; pasting the term into the search box,* as ***search queries****;* and *opening a new tab for reading,* as ***search result review***. This tagging process requires accurate descriptions of behaviors observed in screen recordings and keylogging events during data cleaning. By doing so, we aimed to reduce the likelihood of overlooking variations in behavior that might stem from an analyst's bias. Our goal is to ensure that our tagging or labeling data reliably and accurately reflects the diverse methods our informants used to navigate their tasks. In the analysis, the observations were further refined into corresponding strategies, aiming to capture as broad a range of behaviors as possible. From this analysis, these strategies were categorized into various subtasks and integrated into the task model. For instance, *search result review* and *search queries* are both part of the ***translation search*** subtask, while *source-text term retrieval* is a component of the ***term extraction*** subtask.

This hierarchical model based on successive redescriptions reflects the collective, situated range of behaviors to handle tasks (**Figure 31**). Additionally, from the timestamp of each movement or keystroke event, we can calculate its duration. This temporal dimension allowed us to scrutinize the cognitive activities possibly underlying the dynamic nature of this complex process. The informants integrate various resources, so the order of subtasks is not fixed, resulting in interleaving and also continuous, repetitive behaviors. Abstracted subtasks also recur and alternate, so that the sequences and frequency vary of individual's engagement in subtasks. By observing the informants' behavior within these subtasks, we can dynamically capture conduct dynamics specific to the overlapping subtasks and can therefore hypothesize necessary cognitive processes and combinations thereof likely to be temporarily at work when executing a complex task.

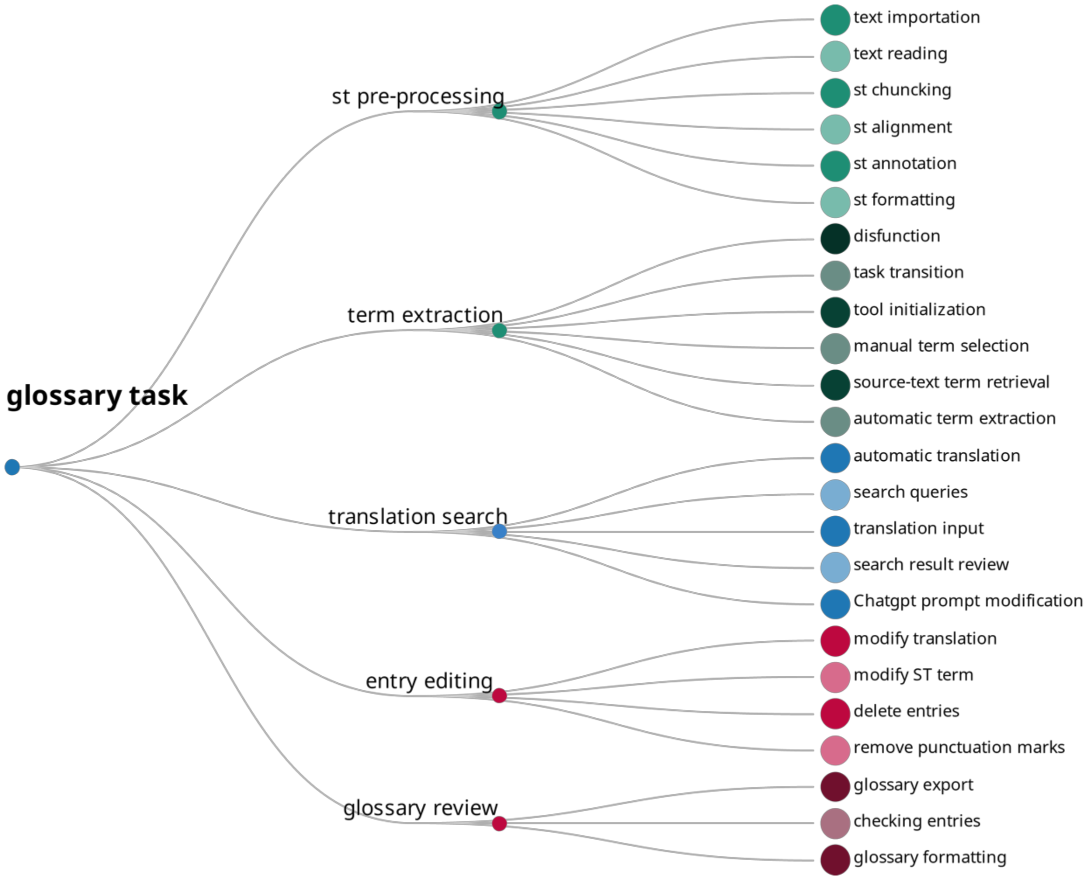


Figure 31. Labels and structures of subtasks and corresponding strategies.

Please note that a task is understood as 'any goal-oriented activity undertaken by an individual or a group' (APA dictionary) and that it is a recurrent notion, so that *subtasks* (not in the APA dictionary) may also be described and conceived of as tasks. For instance, when translating, revision may range from minimal monitoring through being a subtask, part of the translation process, to being a separate task, often performed by people other than the translation drafter, at a different time. The categorization of subtasks and strategies is derived from the behaviors observed in our study. This approach aimed to streamline the very varied individual behaviors and allowed us to compare the use of different CAI tools at glossary compilation in terms of efficiency.

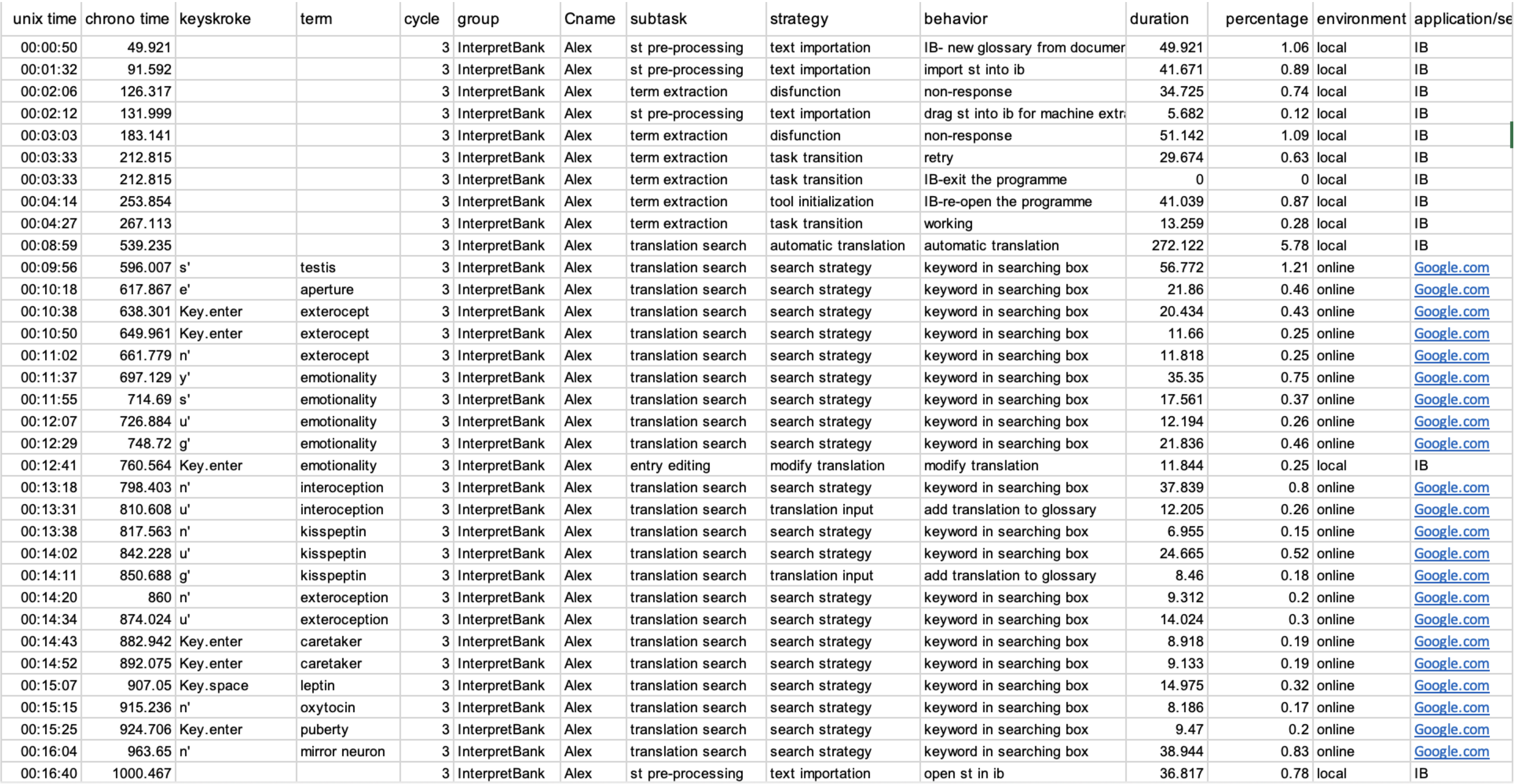
ST pre-processing involves how informants handle the source text. It includes *text importation*, *text reading*, ST *chunking*, ST *alignment*, ST *annotation*, and ST *formatting*. ***Term extraction*** describes how informants rely on either machine extraction, manual extraction, or both. It comprises dysfunction, task transition, tool initialization, *read-first* glossary compilation, source-text term retrieval, and automatic term extraction. *Dysfunction* relates to any bugs or errors encountered in CAI tools. *Task transition* includes actions such as clicking buttons and testing different functionalities within the tools or services. *Tool initialization* indicates steps such as setting up term frequency for automatic term extraction services or typing glossary names. ***Translation search*** focuses on how informants use various resources to find translations for terms. It includes *automatic translation, search queries, translation input, search, result review, and ChatGPT prompt modification*. The ***Glossary review*** describes how informants review the glossary. It does not involve modifying the contents of glossary entries, which are included in the next category. It includes glossary export, checking entries, and glossary formatting. ***Entry editing*** involves how informants modify entry content in the glossary, such as modifying translations, modifying source text terms, deleting entries, and removing punctuation marks.

Due to the complexity of human behaviors, different events may interweave or repeat, not necessarily following a specific pattern. Thus, this categorization groups similar types of behavior events together. Since each event can be measured in a format of chrono time (see **§ 2.7.1.1**), this categorization also implies temporal measurements. This approach lays the groundwork for discussing how term extraction efficiency can be derived from all behavior log data. There is no consensus on the definition of the *term extraction efficiency* (Fulford, 2001). This data, having been timestamped, is easy to compute. We focused on the informants' information consulting behavior, specifically regarding term extraction efficiency. We introduced a series of indicators: *time taken for glossary tasks*, *term counts*, *time per term*, and *diversity rate*. These metrics would quantify term extraction efficiency for group comparison (see **Table 3**).

Time taken for each glossary task is a typical indicator, though it may not always be reliable due to the varying methods of term extraction. For example, while automatic term extraction with InterpretBank can save time, it may require further user intervention actions like removing irrelevant terms identified by the system or deleting duplicates collected through both manual and automatic means. Another indicator related to term extraction is *term counts,* which measures how many terms are ultimately retained in the individual glossary. This indicator may become less and less informative and possibly more and more distorted the more automatized glossary compilation becomes.

The indicator *time per term* refers to the time taken to compile the whole glossary, divided by the number of terms surviving all processing actions. Time per term is a fairer metric for comparing informants using InterpretBank and those using Excel, because InterpretBank informants were expected to tend to have more entries in their glossaries but also to collect them faster. As mentioned, this is a rough estimation, because the extraction process may involve additional actions like deleting, removing, or modifying entries, but assessment is always “manual”, so it yields a fairer comparison. For each informant, we calculate the time per term by dividing the total time taken to compile their glossary by the number of terms in their individual glossaries. Then, to obtain the average time per term at the group level (e.g., for all InterpretBank informants or all Excel informants), we took the mean of the individual time per term values for all informants in that group. We assumed that the group-level average time per term hints at general tendencies of InterpretBank or Excel informants in terms of the time they dedicated to each term. It is assumed to also partially reflect the efficiency of the term extraction process for each group.

The *diversity rate* refers to the percentage of overlap between the individual glossaries of the InterpretBank group and the Excel group (separately). Each informant adopted various tools and services, often employing them interchangeably. Usage percentages for each tool and service were calculated for each informant. This percentage refers to the time taken for each tool and service, divided by the total time devoted to the glossary task. The *diversity rate* of extracted terms from individual glossaries provides basic information about the differences between the two approaches: the Excel glossaries typically represent human choices, while the InterpretBank glossaries are highly influenced by the application functions. InterpretBank’s glossaries can be based on both the informants’ manual and automatic term selection. To some degree, by comparing how InterpretBank’s automatically extracted glossary differs from human selection (whether in Excel or InterpretBank glossaries), we can discern differences between the two approaches to term extraction.



unix time	chrono time	keyskroke	term	cycle	group	Cname	subtask	strategy	behavior	duration	percentage	environment	application/se
00:00:50	49.921			3	InterpretBank	Alex	st pre-processing	text importation	IB- new glossary from documer	49.921	1.06	local	IB
00:01:32	91.592			3	InterpretBank	Alex	st pre-processing	text importation	import st into ib	41.671	0.89	local	IB
00:02:06	126.317			3	InterpretBank	Alex	term extraction	disfunction	non-response	34.725	0.74	local	IB
00:02:12	131.999			3	InterpretBank	Alex	st pre-processing	text importation	drag st into ib for machine extr	5.682	0.12	local	IB
00:03:03	183.141			3	InterpretBank	Alex	term extraction	disfunction	non-response	51.142	1.09	local	IB
00:03:33	212.815			3	InterpretBank	Alex	term extraction	task transition	retry	29.674	0.63	local	IB
00:03:33	212.815			3	InterpretBank	Alex	term extraction	task transition	IB-exit the programme	0	0	local	IB
00:04:14	253.854			3	InterpretBank	Alex	term extraction	tool initialization	IB-re-open the programme	41.039	0.87	local	IB
00:04:27	267.113			3	InterpretBank	Alex	term extraction	task transition	working	13.259	0.28	local	IB
00:08:59	539.235			3	InterpretBank	Alex	translation search	automatic translation	automatic translation	272.122	5.78	local	IB
00:09:56	596.007	s'	testis	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	56.772	1.21	online	Google.com
00:10:18	617.867	e'	aperture	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	21.86	0.46	online	Google.com
00:10:38	638.301	Key.enter	exterocept	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	20.434	0.43	online	Google.com
00:10:50	649.961	Key.enter	exterocept	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	11.66	0.25	online	Google.com
00:11:02	661.779	n'	exterocept	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	11.818	0.25	online	Google.com
00:11:37	697.129	y'	emotionality	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	35.35	0.75	online	Google.com
00:11:55	714.69	s'	emotionality	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	17.561	0.37	online	Google.com
00:12:07	726.884	u'	emotionality	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	12.194	0.26	online	Google.com
00:12:29	748.72	g'	emotionality	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	21.836	0.46	online	Google.com
00:12:41	760.564	Key.enter	emotionality	3	InterpretBank	Alex	entry editing	modify translation	modify translation	11.844	0.25	local	IB
00:13:18	798.403	n'	interoception	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	37.839	0.8	online	Google.com
00:13:31	810.608	u'	interoception	3	InterpretBank	Alex	translation search	translation input	add translation to glossary	12.205	0.26	online	Google.com
00:13:38	817.563	n'	kisspeptin	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	6.955	0.15	online	Google.com
00:14:02	842.228	u'	kisspeptin	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	24.665	0.52	online	Google.com
00:14:11	850.688	g'	kisspeptin	3	InterpretBank	Alex	translation search	translation input	add translation to glossary	8.46	0.18	online	Google.com
00:14:20	860	n'	exteroception	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	9.312	0.2	online	Google.com
00:14:34	874.024	u'	exteroception	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	14.024	0.3	online	Google.com
00:14:43	882.942	Key.enter	caretaker	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	8.918	0.19	online	Google.com
00:14:52	892.075	Key.enter	caretaker	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	9.133	0.19	online	Google.com
00:15:07	907.05	Key.space	leptin	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	14.975	0.32	online	Google.com
00:15:15	915.236	n'	oxytocin	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	8.186	0.17	online	Google.com
00:15:25	924.706	Key.enter	puberty	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	9.47	0.2	online	Google.com
00:16:04	963.65	n'	mirror neuron	3	InterpretBank	Alex	translation search	search strategy	keyword in searching box	38.944	0.83	online	Google.com
00:16:40	1000.467			3	InterpretBank	Alex	st pre-processing	text importation	open st in ib	36.817	0.78	local	IB

Table 3. Sample of behaviors, activities, and subtasks in glossary tasks.

2.7.2.2 *Indicators in booth tasks* To capture and categorize behaviors at different stages, indicators were formulated that were either related to time or content. First, blank intervals or silent spans in the speech flow of the informants were categorized into three sets: under 200 ms *(lags),* between 200 and 600 ms *(bumps),* and above 600 ms *(respites).* The constructs and the terms draw from Muñoz & Apfelthaler (2022). With today's knowledge, any and all intervals shorter or equal to 200ms, or lags, were deemed to amount to noise, so they were ignored. The 200ms is an arbitrary threshold, indirectly supported by most of the reasons argued by Muñoz & Apfelthaler (2022) for choosing that threshold for translation typing. For the SI rendering, we categorized time gaps into length-based subcategories, *bumps* (200–600-ms, see **§ 3.2.4**) and *respites* (over 600 ms, see **§3.2.6**)—based on the notion that simultaneous interpreters are under pressure to avoid communicative breakdowns, and tend to avoid respites longer than 600 ms (Ho, 2021), for instance, by adding fillers to maintain the attention of their audience.

The upper threshold of 600 ms follows Ho (2021), who found that pauses over 600ms were consistently noticed by both trained and untrained native Chinese speakers as gaps. The upper threshold is also the closest round under the pause average across groups (623 ms in Cycle I, 661 ms in Cycle II, and 639 ms in Cycle III). Technically, 650ms was even closer and more round enough, but we opted to accept more potential noise (false hits, Type I error, i.e., a false discovery) rather than omit significant intervals (a type II error, i.e., a missed discovery), adhering to the commonsensical principle of *better safe than sorry,* which seems appropriate in exploratory research.

This selection of the 600 ms upper threshold also inspired the results and methods from cross-lingual generalization. Wehrle (2023) referred to previous work by De Jong & Bosker (2013), Cho & Hirst (2006), and Megyesi & Gustafson-Capková (2002) to establish a 700 ms cut-off for identifying silence thresholds, aligning with findings from these earlier studies. In the Wehrle's research (2023), this threshold exceeded the average pause duration in the current dataset (646 ms across groups), which had autistic informants. People with disorders in the autistic spectrum may be faster or slower speakers than *allistic* (neurotypical) speakers, and they may also display different pausing patterns, which may impact their averages. As a result, a 600-millisecond threshold seemed again a prudent limit to define potentially perceptible pauses. The upper threshold, in brief, divides hypothesized non-mechanical intervals (all those above 200 ms) into those prone to go unnoticed, or bumps (between 200 and 600 ms) and those that tend to be noticed (above 600 ms). We assume that interpreting trainees may have developed a particular perceptual sensitivity or acuity for this threshold, however intuitive, as flagging that their performance may go wrong.

Ear-voice span, or EVS, is the time gap between the acoustic end point of a source speech unit (usually, a sentence) and the acoustic ending point of the aligned chunk in the target output. A well-known indicator, it can be traced back to the very first studies of simultaneous interpreting (Paneth, 1957; Oléron & Nanpon, 1965/2000; Barik, 1973). Christoffels & De Groot (2004) found a significant difference between *shadowing* and simultaneous interpreting but there is no consensus as to how to interpret it, beyond the general notion that larger EVS tend to reflect more cognitive difficulties performing the task. In order to refine the analysis, this study drew inspiration from the notion of *sentence-initial eye-voice span,* as opposed to *sentence-final eye-voice span* from Zhou *et al.* (2021). In our study, *EVS1* symmetrically refers to the time gap between the acoustic starting point of a source speech unit (here, a *chunk*) and the acoustic starting point of the aligned chunk in the target output. *EVS2*, symmetrically, refers to the time gap between the end of the acoustic signal of the source speech unit (a *chunk*) and the ending point of the acoustic signal of each informant's rendition of that unit. Here, chunks are seen as production units that tend to coincide with sentences, which are linguistic units, but some consecutive sentences tend to be interpreted together, so we study them together too (see **Appendix H** for chunks and sentences, and also their differences in green background cells). At the very least, considering both EVS measurements together may support interpretations of the informants' behavior being regular or not (e.g., *EVS1* is similar to *EVS2*, compensatory or not (e.g., faster pace and *EVS2* shorter than *EVS1*). As for content indicators in the booth task, they are the following:

- *fillers*: *"Uh, mm"* and the like.
- *repetitions*: a sequence of at least two words follows an identical sequence immediately afterward (e.g., ABCDCDE...).
- *self-corrections*: adjustments of errors on the fly.
- *false starts*: abruptly interrupted start. It is not necessarily a subtype of repetitions, as false starts often lead to different expressions being uttered as repairs.

These two sets of indicators were described above in terms of the nature of the information they measure, but they can also be approached in terms of what they may flag. Since we were exploring new aspects with new tools, we decided to apply consensual views in interpreting research and thus approach these indicators from the perspective of fluency and accuracy. Considered in this way, the indicators may reveal individual and group differences in both efficiency and accuracy, with or without CAI tools. However, there is a difference between fluency and efficiency, on the one hand, and accuracy and effectiveness, on the other. Whereas fluency and efficiency may be easily reduced to quantitative indicators such as the ones chosen and presented above, accuracy and effectiveness entail a rather wider margin for the interpretation of the observed phenomena on the part of the researcher.

To provide a clear overview of the impact of InterpretBank on the fluency and accuracy in the renditions of Chinese interpreting trainees, we clustered the indicators presented above into two groups: a content cluster and a time cluster. The content cluster encompasses *false starts, self-corrections, fillers, and repetitions,* which are—often, but by no means always—intricately related to the practices of repair in conversation (Hayashi & Yoon, 2010, p. 56). The time cluster focused on the distribution of actions over time whereas the content cluster talks on the nature of activities, whether in absolute terms *(i.e., bumps and respites)* or in relative terms (with respect to the source speech: EVS1 and EVS2).

For accuracy, we adopted two strategies: a local and a global evaluation. First, we categorized the renditions of the potential problem triggers (see **§ 2.2.2**) into four categories that would then be quantified. Second, we resorted to external raters performing a holistic (not itemized or rubric-based) evaluation. To determine the quality of the renditions, they were categorized into four conditions: *correct, adequate, wrong renditions,* and *skipped terms.* The term *correct* refers to the use of the exact and identical rendition provided by the master glossary. *Adequate* refers to the informants' renditions that match the correct terms in the master glossary as to their meaning. That is, they are good renditions that do not match the ones in the glossary. These renditions are both understandable and acceptable from the researcher's perspective—maybe not so to hearers!—but they are simply not the ones previously deemed optimal and hence entered in the master glossaries (as they were when returned to the researcher after the booth task). *Wrong* terms are renditions not included in the master glossary and deemed not understandable or acceptable from the researcher's perspective. Obviously, in many cases, the informants may have thought otherwise. *Skipped terms* refer to the absence of renditions in the informants' production corresponding to the 39 potential problem triggers.

These four conditions were applied to 39 potential problem triggers, including 33 *first-time* terms and 6 *repeated* terms. For each condition in first-time terms, basic statistical analysis was performed to analyze the distribution within the group and the differences between groups. The distribution presents the frequency value of each condition for each informant in the group, and the median value for each group of each condition was calculated. Group differences were compared using the Mann-Whitney U test for within-group and the Friedman test for inter-group analysis from Cycle I to Cycle III. For repeated terms, a group-level comparison shows the percentage rate of each condition for each group over time. This allows for the comparison of the performance related to *repeated* terms (i.e., *rep1* and *rep2* in **§-3.3.5**) for two groups across cycles. As for the global assessment of the audios, it is technically part of the booth indicators, but it is devoted a separate section because it stands out for being performed by third parties and because SI rendering quality may be the result of factors beyond the booth task.

2.8 Summary

This chapter outlines the materials and methodologies employed in the study. The research involved 22 Chinese L1 speakers with English as their L2, all enrolled in MA programs in conference interpreting at top Chinese universities. Their ages ranged from 22 to 34, with an average of 24.7 years. These informants had completed at least two semesters of SI training before joining the study. A mix of macOS and Windows system users, with some having prior experience with InterpretBank, were included in the study.

The study utilized *Dr. Huberman Lab Podcast* series episodes, categorized by specific topics, as source material. These episodes were transcribed using Microsoft Stream and the scripts were manually modified for the study purposes. Episodes were divided into glossary and booth task texts, with particular attention to ensuring consistent speech and style features. A comprehensive approach was used to select potential problem triggers, employing tools like AntConc and BootCaT for term extraction and corpus compilation. This meticulous process involved creating a customized, domain-specific corpus to identify key terms and phrases.

The study compared the use of InterpretBank and MS Excel for managing glossaries in SI tasks. Excel was selected for the control group due to its popularity and accessibility, despite its limitations in terminology management. InterpretBank, representing advanced CAI tools, was chosen for the experimental group. Its features, such as term extraction modes, translation suggestions, and term retrieval support in RSI, were evaluated for their impact on interpreters' performance.

The study employed a mixed-method, pretest-posttest design over three rounds. The first round (Cycle I) established a baseline for all informants. Cycle II introduced the use of InterpretBank and MS Excel in two separate groups. In Cycle III, informants chose their preferred tool. The study design included glossary and booth tasks, with an additional online training session for the InterpretBank group. Informants' performance and preferences were surveyed after the tasks.

Data was collected remotely using Microsoft Stream for audio files, a Python-based keylogger for keystroke recording, and TechSmith Capture for screen recording. These tools were selected for their user-friendliness and compatibility with remote data collection. The data collection process was adjusted to a remote setup due to travel restrictions. Stable internet connections and specific software setups were crucial for the integrity of the data collection.

The collected data underwent thorough cleaning and synchronization processes. Behaviors in the glossary and booth tasks were categorized and coded for analysis. The study introduced novel constructs such as ear-key span (E2K) and eye-voice span (I2V) to explore cognitive effort during tasks. Various indicators were used to analyze efficiency and effectiveness in both glossary compilation and SI output, including, but not limited to, *fillers*, *repetitions*, and *self-corrections*. Holistic evaluations of audio files were also conducted by external raters.

Chapter 3

results

Data was collected over three cycles, each one consisting of one glossary task and one booth task, which bestowed additional temporal attributes upon the log data (see indicators in **§ 2.7.2**). In this chapter, we will interpret data from the glossary and booth tasks. Statistical analyses measure specific aspects of interpreting quality, such as term retrieval and term accuracy, but these variables are only parts (or, rather, aspects) of the interpreting process. Isolating them in an experimental setup for inferential analysis is challenging because such aspects co-occur and sometimes are even one and the same behavior contemplated from different viewpoints (for instance, *false starts* have an impact on *EVS1*). Single variable measures, as isolated in test results, cannot fully represent the high degree of interaction with tools and the environment, but also among different hypothesized cognitive processes, inherent in the interpreting process. Consequently, results should be interpreted as evolving over time within the tasks, to more accurately and realistic reflect the complex behavior we simply describe as *remote, CAI-tool supported simultaneous interpreting.* In view of the many confounders and potential variables unaccounted for, an intra-subject analysis allows greater security in interpreting behavior in naturally varying environments. Indeed, as expected, Excel served as an external support for glossary compilation. Its integration with various local or online tools demonstrated strong compatibility and flexibility, going beyond the simple activity of copying terms from the article and pasting them with their respective Chinese renditions that informants would consult. The strategies for term extraction varied across informants, and glossary consultations also displayed considerable variation.

3.1 Glossaries

Glossary compilation involved manually tagging punctual events and span constructs derived from behavior observations from screen recordings, and from individual glossaries, as described in **§ 2.7.1.1.**

3.1.1 Individual glossary

In the glossary tasks, information-seeking performance in each glossary subtask (see **§ 2.7.2.1**) was analyzed for two groups, and the adaptation of flashcard mode in the InterpretBank group.

3.1.1.1 *Compilation* **Figure 32** depicts a circular dendrogram, representing the hierarchical clustering of informants, engaged in a glossary task in Cycle I. The dendrogram's branching patterns identify clusters of informants who exhibit similar behaviors based on task-related criteria, specifically sub-tasks and their associated activities.

subtask	activities
st pre-processing	text importation
	text reading
	st chunking
	st alignment
	st annotation
	st formatting
translation search	automatic translation
	search queries
	translation input
	search result review
	ChatGPT prompt modification
term extraction	dysfunction
	task transition
	tool initialization
	read-first glossary compilation
	source-text term retrieval
	automatic term extraction
glossary review	glossary export
	checking entries
	glossary formatting
entry editing	modify translation
	modify ST term
	delete entries
	remove punctuation marks

Table 4. Ethogram subtasks and activities in glossary tasks.

The dendrogram branches represent informants by their initials encircled with colored nodes: mustard for InterpretBank and green for Excel, matching their interface colors. We have aggregated the time spent on identical activities for each individual and divided it by individual total glossary time for Cycle I. This calculation yields a percentage, which is represented by a point on the outermost circle of the dendrogram. In other words, the size of the node indicates the proportion of an individual's engagement in specific documented activities within the glossary task in Cycle I. For example, informant Jordan (J) dedicated 33.64% of his time to the *read-first* glossary compilation as one of the activities from the source text. The size of each node on the dendrogram's periphery is the cumulative percentage of individual time spent on activities within its branch, allowing for a visual comparison of strategy engagement. These nodes aggregate at the activity level and coalesce further into subtasks, with the size of subtask nodes being the sum of the sizes of their constituent activity nodes, reflecting the relative time investment in these subtasks in the glossary of Cycle I.

All informants in the InterpretBank group in Cycle II used *automatic term extraction* before *read-first* glossary compilation (i.e., *auto before read-first* in **Table 5**), but in Cycle III, four out of 11 informants (i.e., Blake, Casey, Dana, and Jordan) changed from *automatic term extraction* + *read-first* glossary compilation to only relying on *read-first* glossary compilation.

informant names	Cycle III
Alex	auto before read-first
Blake	only read-first
Dana	only read-first
Erin	auto before read-first
Frankie	auto before read-first
Gale	only read-first
Harley	auto before read-first
Ira	auto before read-first
Jordan	only read-first
Kelly	auto before read-first
Lee	auto before read-first

Table 5. Automatic term extraction in Cycles III.

The percentage of time devoted to *activities* is used to rank informants based on their participation in various sub-tasks, represented in circles with other colors in **Figure 32**, whose shade hints at recursive categorization. The basic subtasks within the glossary task are ***ST pre-processing*, *term extraction*, *translation search***, ***entry editing*,** and ***glossary review*.** Larger sub-task dots indicate higher shares of time devoted to that specific activity and subtask. Within each activity, represented by lighter-colored dots, informants are arrayed in descending order of their individual time percentages. For instance, within the ***translation search*** subtask, the informants range from Kelly (K, 94.62%) to Blake (B, 5.82%) in decreasing order of time spent. Basic subtasks comprise, in turn, several behaviors, represented in similar but lighter colors. For instance, ***translation search*** (dark purple, on the right-hand side) breaks down into *search queries, translation input,* and *search result review,* represented in lighter purple circles, logically smaller.

Data shows that the most time-consuming subtask within the glossary task in Cycle I was ***translation search*** —particularly the execution of various *search queries*—closely followed by ***term extraction***. ***ST pre-processing*** and ***glossary review*** consumed roughly equal amounts of time, while ***entry editing*** required the least. Interestingly, informants from the InterpretBank group generally invested more time compared to those from the Excel group, across most activities. Within *search queries,* common behaviors include *copy and paste, hypothetical equivalence testing,* and the *use of bilingual N-grams as search terms*. A comprehensive exploration of these activities would significantly expand the project's breadth, but it is beyond the scope of this project and is penciled as pending for future research.

In Cycle II (see **Figure 33**), ***translation search*** emerged again as the most time-consuming subtask, with *search queries* particularly time-consuming. Notably, the top four informants spending the most time on this task—*Ira (I), Frankie (F), Erin (E),* and *Lee (L)*—were all from the InterpretBank group. Noticeably, *automatic translation* was predominantly employed by InterpretBank informants, with Informant Gale (G) leading at 71.25%. No *automatic translation* was found among informants in the Excel group. Informants from the Excel group generally devoted more time to *translation input*. Within the term extraction subtask, Excel informants also spent much time on the *read-first* glossary compilation. For instance, Sidney (S) allocated 71.10%; Oakley (O), 43.64%; and Quinn (Q), 36.35% of their time to this task. This suggests that the Excel informants primarily focused on *translation input* and *read-first* glossary compilation, with less time allocated to other activities, possibly due to the influence of the *search queries*. In comparison, the InterpretBank informants invested more time in other behaviors, that is, they displayed a broader engagement across subtasks.

Figure 34 represents the behaviors of informants in the glossary task in Cycle III. It corroborates the patterns observed in Cycles I and II, with ***translation search*** and ***term extraction*** as the most time-consuming subtasks among all the subtasks. Within their subcategorized behaviors, the implementation of *search queries* and *read-first* glossary compilation continued to dominate in terms of timeshare. Excel informants spent time predominantly in ***translation search*** and ***term extraction***, indicating a more concentrated approach in these specific areas. The Excel informants predominantly focused their efforts on *search queries*, *translation input*, and *read-first* glossary compilation. Interestingly, Excel informants Taylor (T), Peyton (P), and Val (V) dedicated substantial proportions of their time to *search queries* (88.13%, 74.58%, and 71.61%, respectively). In contrast, InterpretBank informants displayed a more diverse time allocation, with higher shares of other subtasks (i.e., ***glossary review, entry editing,*** and ***ST pre-processing***). Irrespective of the group, *search queries* significantly impact time allocation, underscoring its central role in the glossary task process. Using the data cleaning methods described in **§ 2.7.1.1**, each observed behavior was tagged and mapped to the activities and subtasks described in **§ 2.7.2.1** and **Table 4**.

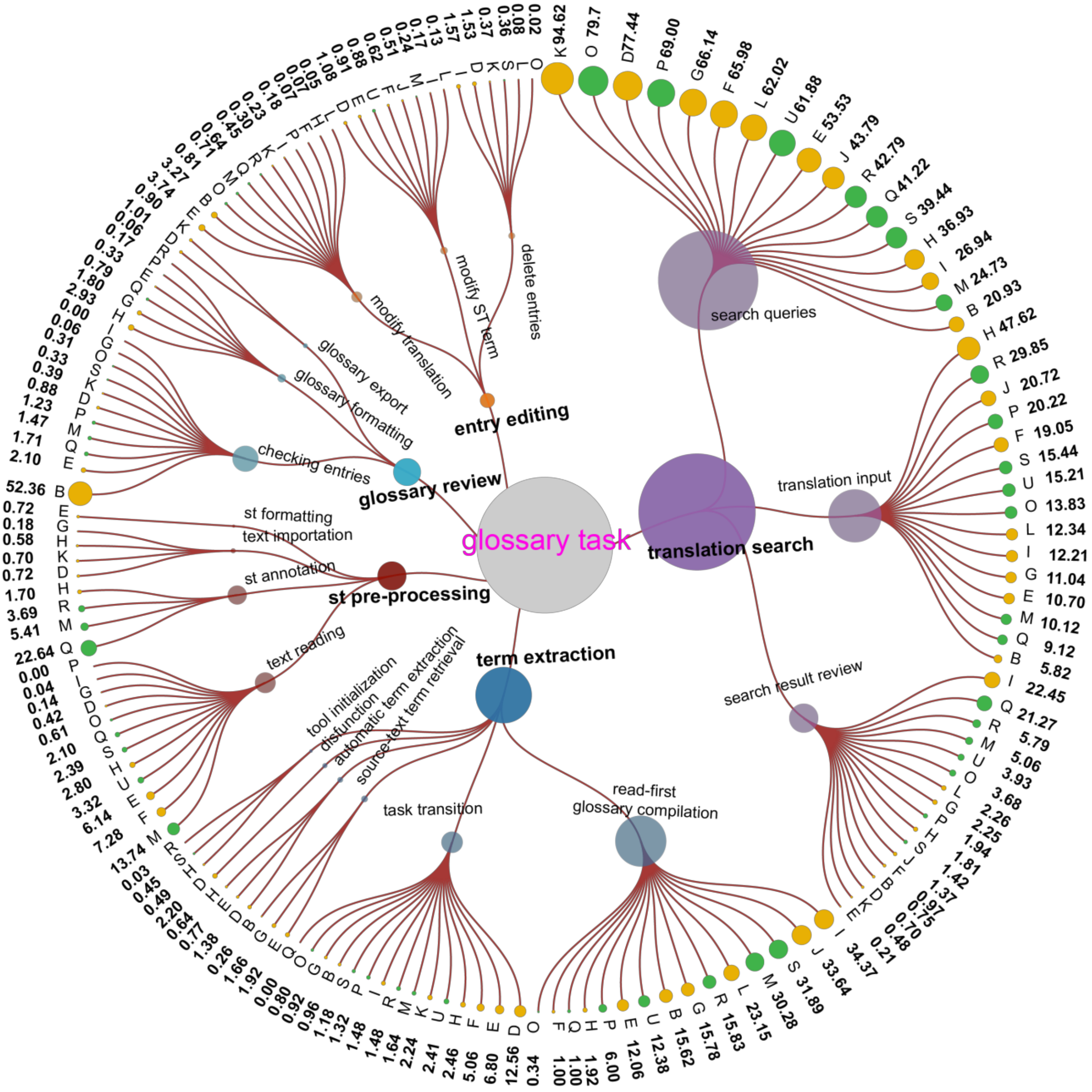


Figure 32. Behavior of informants in the glossary task of Cycle I.

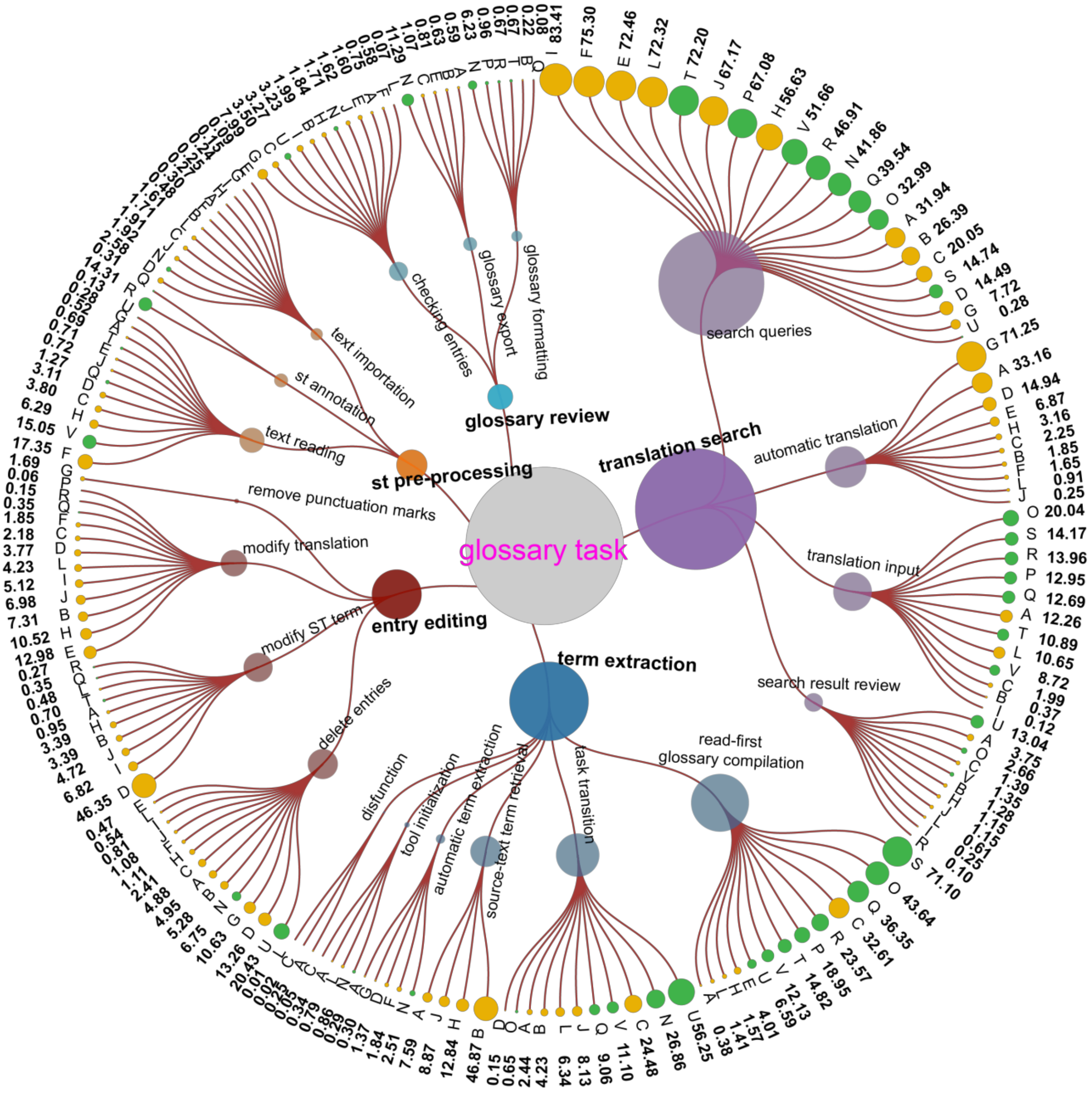


Figure 33. Behavior of informants in the glossary task of Cycle II.

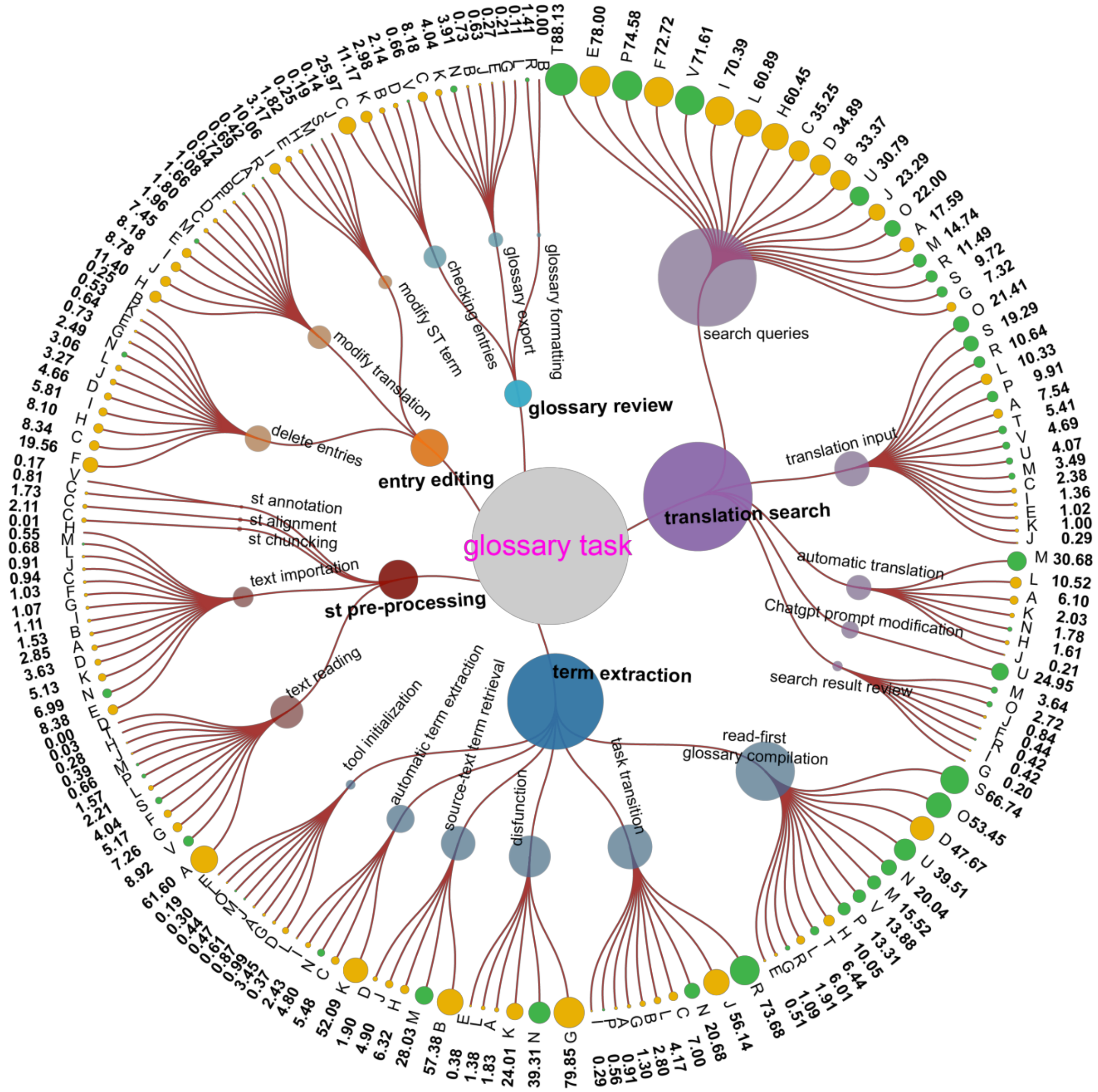


Figure 34. Behavior of informants in the glossary task of Cycle III.

3.1.1.2 *Sources of consultation* This section explores the applications and services used for term retrieval. **Figure 35** presents a layered sunburst diagram that highlights individual tool usage. Each concentric ring denotes a hierarchy level. The innermost ring codifies the *group*, while the subsequent rings represent each informant. InterpretBank informants, from Alex to Lee, are depicted in yellow tones, and the Excel informants, from Morgan to Val, are in green.

In this diagram, the outermost ring denotes *application/service*, with each segment's size corresponding to the time percentage dedicated to that tool, relative to the total time spent on the glossary task. Light blue segments represent the use of local (offline) tools and light orange segments, the use of online tools. Segments with important shares are labeled with the tool or service. For instance, while informant Harley predominantly used Interpreters' Help as evidenced by the substantial allocation of time to this tool. Smaller segments, indicating lesser-used applications or services, are not labeled for obvious space restrictions. These minor segments were occasional or trial uses of tools for term retrieval. Nevertheless, while these instances constitute a small percentage of overall tool usage, they point to the informants' exploratory behavior and help understand the broader picture of tool use.

Figure 36 illustrates group-level tool and service interactions in Cycle I through a chord diagram. Each group is represented with their usual color-coded segments along the ring—green for Excel and yellow for InterpretBank. From these segments, arrows depart that reach applications and services represented as gray segments. The arrows map out the connections between groups and the tool they used, with each ribbon's endpoint corresponding to a particular application or service. The width of each arrow codifies the timeshare allocated to each specific application or service. For example, the Excel group (green ribbon) predominantly uses MS Excel, followed by Google and MS Word. Interestingly, the InterpretBank group (yellow ribbon) displayed the most extensive use of MS Excel, suggesting a higher reliance on this tool. The reader is reminded that in Cycle I informants were not grouped into different cohorts using either Excel or InterpretBank, but they could rather use the tools of their choice (e.g., Excel, InterpretBank, or any other tools), as opposed to information sources which they were allowed to freely consult through all three cycles.

Figure 37 presents a sunburst diagram that illustrates the tool used by individual informants in Cycles II and III, aiming to explore contrasts and patterns in their approach to glossary tasks. The diagram is divided into two semi-circles, with the left representing Cycle II and the right representing Cycle III. Moving outward from the center, the first ring is marked by two segments: Cycle II and Cycle III. The second ring from the center has four segments, each colored by group: the green segment is labeled *InterpretBank group*, and the orange segment is labeled *Excel group*. Each semi-circle has two color-coded segments representing the two groups. The third ring contains segments with the informant names of the corresponding groups, colored in different shades of the overall group colors to help readers locate the corresponding data for the same informant in the opposite cycle.

The different yellow or green shades in the segments of group members are meaningless, they only seek to ease location and identification. The outermost visible ring contains segments, each representing the local applications and online services used by each informant. The size of each segment represents the percentage of time devoted to the tool. This percentage is calculated by dividing the cumulative duration used by the application or service by the total personal glossary time for Cycles II and III. While most segments are labeled with the names of applications and services, some remain unlabeled due to their minimal percentage, prioritizing diagram readability and clarity (see more in **Table 6**).

In **Figure 37**, we observed that most informants consistently utilized InterpretBank, except one informant, Casey, who ceased using InterpretBank in Cycle III. Focusing on segment size, in Cycle II, among the 12 informants using InterpretBank, Kelly, Gale, Dana, and Casey allocated a significant portion of their time to InterpretBank, while the others showed less usage, allocating more time to other applications and services such as search engines (e.g., so.com, google.com, baidu.com) and local applications (e.g., Oulu app, Youdao app, MS Word). Oulu app and Youdao app are popular free bilingual dictionaries among Chinese users. In Cycle III all informants continued using InterpretBank except for Casey. Among these 11 users, Alex, Dana, Gale, John, and Kelly showed an increased proportion of time spent on InterpretBank compared to Cycle II, indicating a slight change in usage patterns. The rest continued to show scare use of InterpretBank, preferring other applications and services, including search engines (e.g., so.com, google.com, cn.bing.com) and bilingual dictionaries (e.g., Oulu app, Youdao app, MS Word), which was similar to Cycle II.

In both Cycles II and III, the Excel group relied heavily on external applications and services. They used a variety of external tools, for instance, Youdao Dictionary, cn.bing.com, Google.com, and Bing.com Additionally, they spent a significant amount of time using MS Excel. This may suggest that they frequently switched between MS Excel and external tools in their work. Possibly, actions involving summarizing data in Excel took up a considerable amount of time. This pattern persisted into Cycle III. Moreover, in Cycle III, Excel group member Alex adopted InterpretBank in the glossary task.

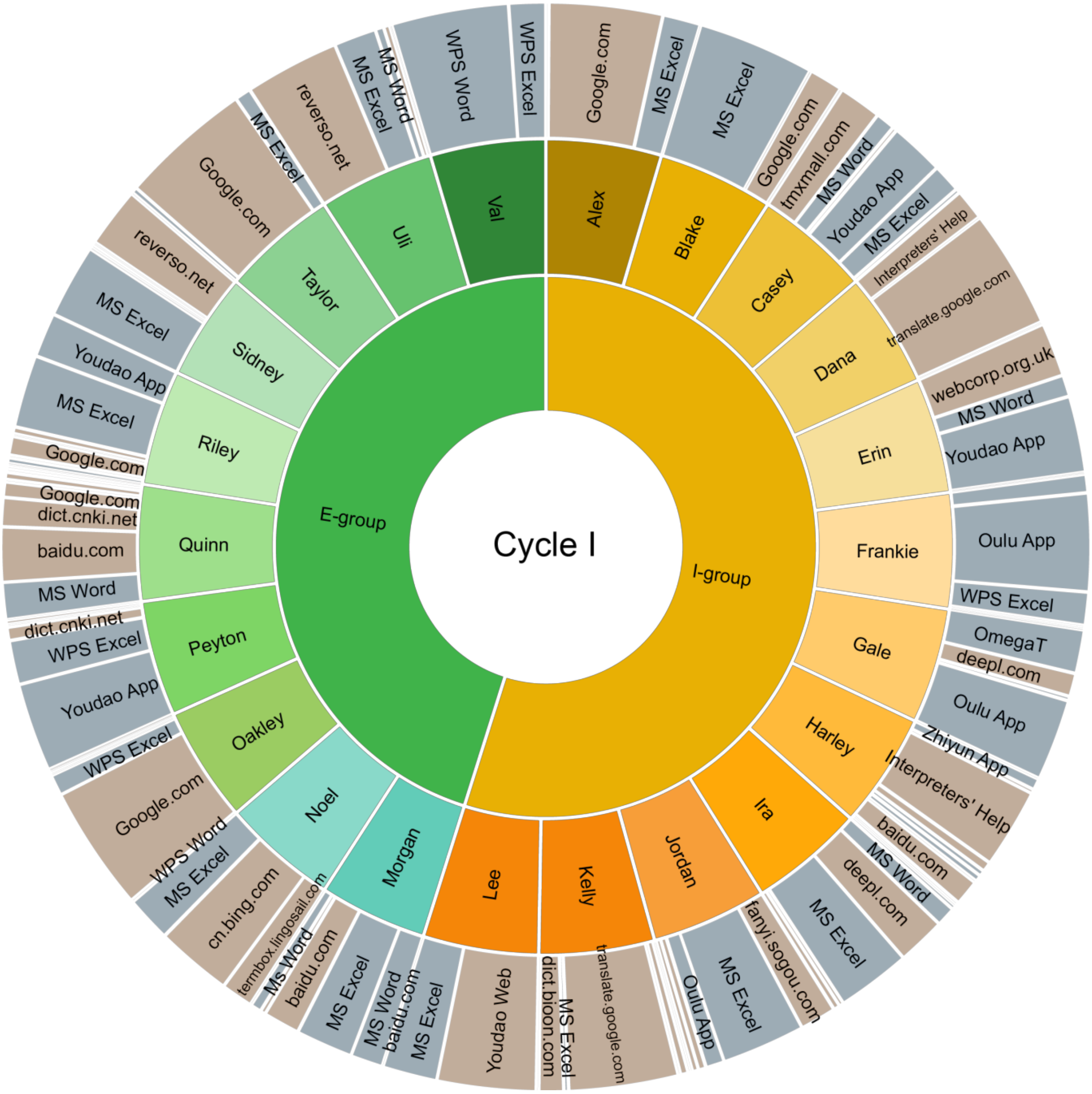


Figure 35. Application/service percentage in Cycle I.

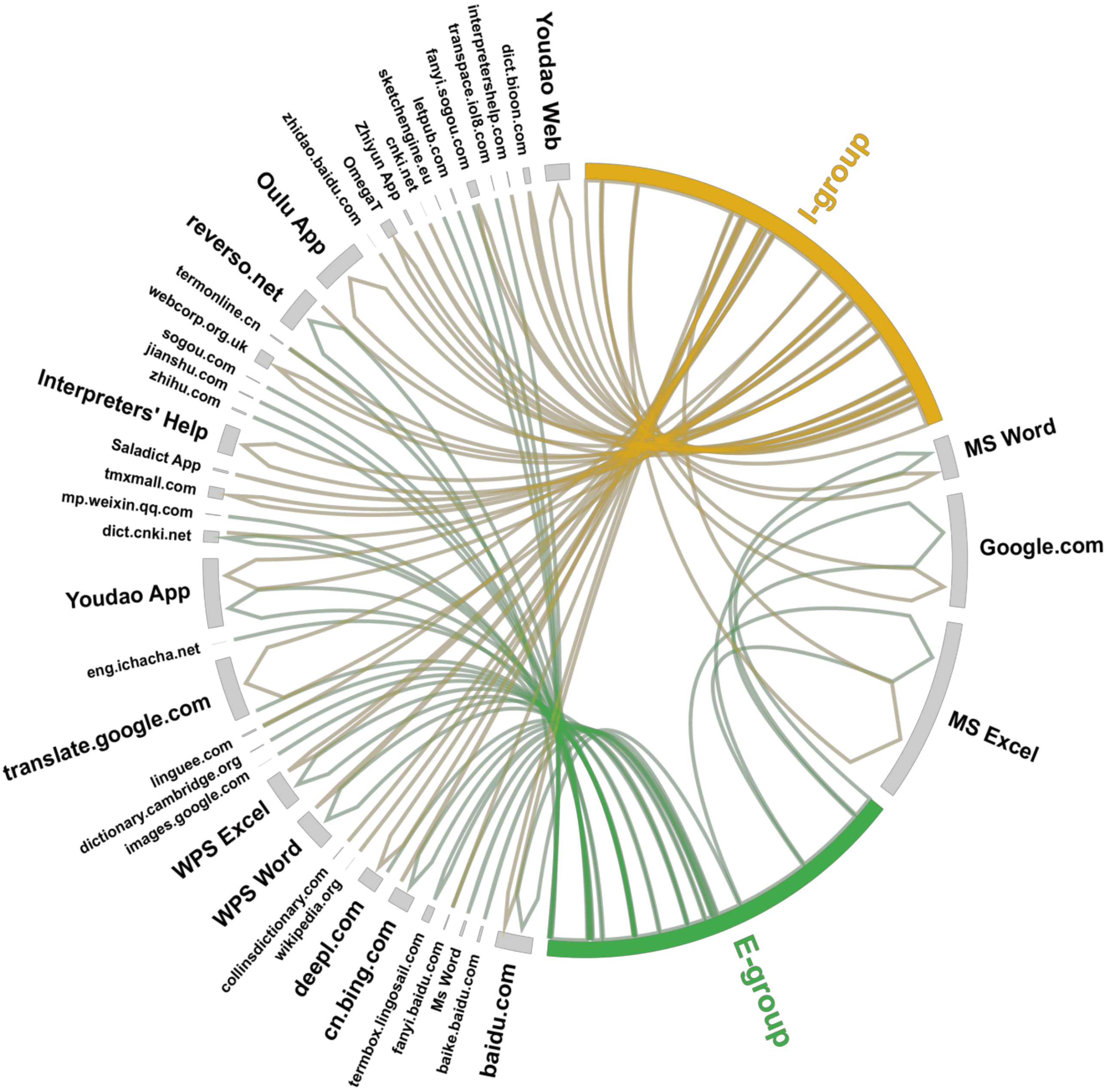


Figure 36. Group-level application/service percentage in Cycle I.

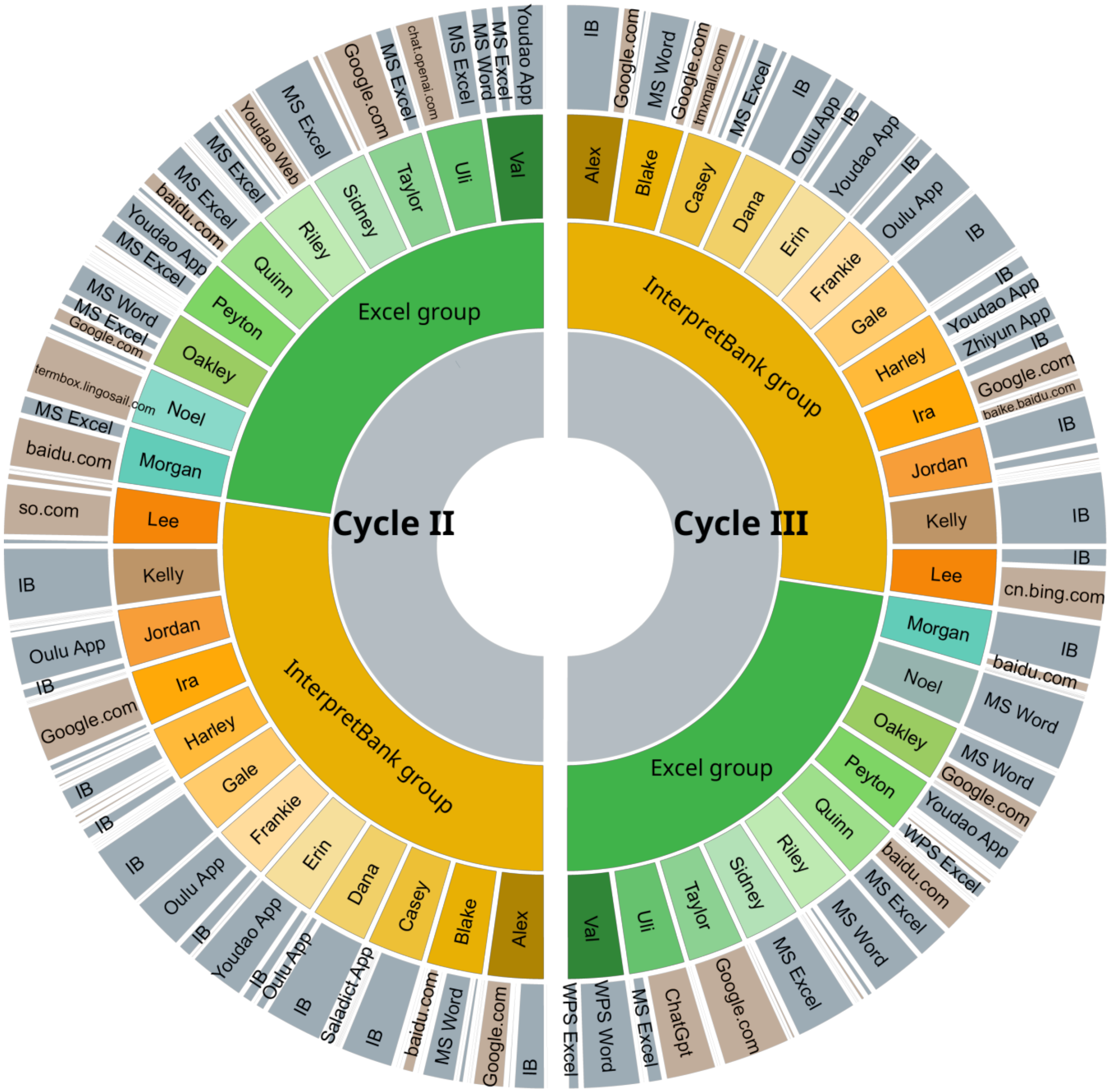


Figure 37. Application/service percentage in Cycles II and III.

Cycle	group	environment		application/service	duration(s)	percentage (%)
II	XL	local	28238.3	MS Excel	16842.3	35.61
	47300.8 s		59.70%	MS Word	6864.2	14.51
				Youdao app	4531.8	9.58
		online	19062.5	others	6458.7	13.65
			40.30%	baidu.com	5175.4	10.94
				Google.com	4678.1	9.89
				Youdao Web	2750.3	5.81
	IB	local	38156.4	IB	18106.9	33.13
	54659.1 s		69.81%	Oulu App	9414.9	17.22
				Youdao app	6365.4	11.65
				MS Word	2803.3	5.13
				others	1466.0	2.68
		online	16502.7	so.com	6103.6	11.17
			30.19%	others	5609.8	10.26
				Google.com	4789.3	8.76
III	XL	local	31564.1	MS Excel	8867.5	19.58
	45279.4 s		69.71%	MS Word	8842.9	19.53
				WPS Word	5863.3	12.95
				IB	4554.5	10.06
				WPS Excel	2005.7	4.43
				others	1430.2	3.16
		online	13715.2	Google.com	5881.1	12.99
			30.29%	baidu.com	3721.4	8.22
				others	2238.7	4.94
				chat.openai.com	1874.1	4.14
	IB	local	38856.0	IB	21303.4	43.23
	49284.2 s		78.84%	Youdao app	6244.3	12.67
				Oulu App	5872.0	11.91
				others	3062.3	6.21
				Zhiyun App	2373.9	4.82
		online	10428.2	cn.bing.com	5133.2	10.42
			21.16%	Google.com	3287.6	6.67
				others	2007.4	4.07

Table 6. Group-level application/service in Cycles II and III.

Table 6 focuses on the use of tools by groups, not individuals, and depicts the use of applications and services by groups from Cycle II to Cycle III. For instance, in Cycle II, the Excel group informants cumulatively spent 47300.8 s in the glossary task. Of that time, 28238.3 s (59.70%) were spent on *local applications*, and 19062.5 s (40.30%) on online services. For local applications, the top three most time-consuming were MS Excel (16842.3 s, 35.61%), MS Word (6864.2 s, 14.51%), and Youdao app (4531.8 s, 9.58%), with MS Excel accounting for 35.61% of the total 47300.8 seconds. Similarly, for online services, the top three most time-consuming were *others* (6458.7 s, 13.65%), baidu.com (5175.4 s, 10.94%), and Google.com (4678.1 s, 9.89%), with the others category accounting for 13.65% of the total time, 47300.8 s (see below for a detailed explanation of the *others* category).

We first calculated the total time spent on the glossary task for each group by summing the time spent by each informant. Next, we calculated the duration in seconds spent on each local application or online service by each informant. This was similar to the process used to create dendrograms in **§ 3.1.1.1**. We then summed these durations within each group. Then we split the data into two categories: local applications and online services (see **Table 6**). We labeled the accumulated duration for each category within each cycle in each group and followed it with a percentage. This percentage shows the proportion of time spent on local applications or online services in the total group time spent on the glossary task for each group. For each environment (local or online), we listed the applications and services used in each category, along with their corresponding duration and percentage. The percentage was calculated by dividing the duration for an application or service by the total time spent on the glossary task for each group in each cycle. Finally, we calculated the average percentage for each tool (local application and online service) within each group. For instance, in Cycle II, the average percentage of applications and services for the Excel group was 4.76%. Therefore, any application or service within the local and online categories whose percentage was lower than 4.76% was combined into the *Others* category. The complete list of applications and services in the *Others* category, along with their percentages, is shown in **Appendix I**. In **Table 6**, the percentage column uses color to highlight the three tools with the highest percentage for each environment.

Table 6 shows that both the Excel group and the InterpretBank group relied heavily on local applications, spending significantly more time on them than on online services. This may be due to the specialized nature of local applications (such as dictionaries), the direct presentation of needed information, or personal preference. Specifically, for local applications, the Excel group spent the most time on MS Excel and MS Word in both Cycles II and III. This may suggest that these tools are essential for their work in the glossary tasks. The InterpretBank group spent the most time on InterpretBank, Youdao, and Oulu in both cycles. These two applications are focused on translating words, so it can be inferred that the informants did not rely on InterpretBank's translation results, but instead searching translation for their own. In terms of online services, the Excel group and the InterpretBank group had a high percentage of *others* set of tools in Cycle II, ranking first and second, respectively. This shows that there was a large variation in the online services tools used by the two groups. In Cycle III, however, both groups shifted to search engines, such as cn.bing.com and Google.com. This may suggest that they were adapting their tool choices to the demands of the tasks at hand.

3.1.1.3 *Glossary contents* **Figure 38** presents a multi-set bar chart detailing term counts in the individual glossaries compiled by each informant from the two groups over the three cycles. The chart is organized in descending order based on the average number of terms. InterpretBank Informants have yellow bars, while Excel informants have green bars.

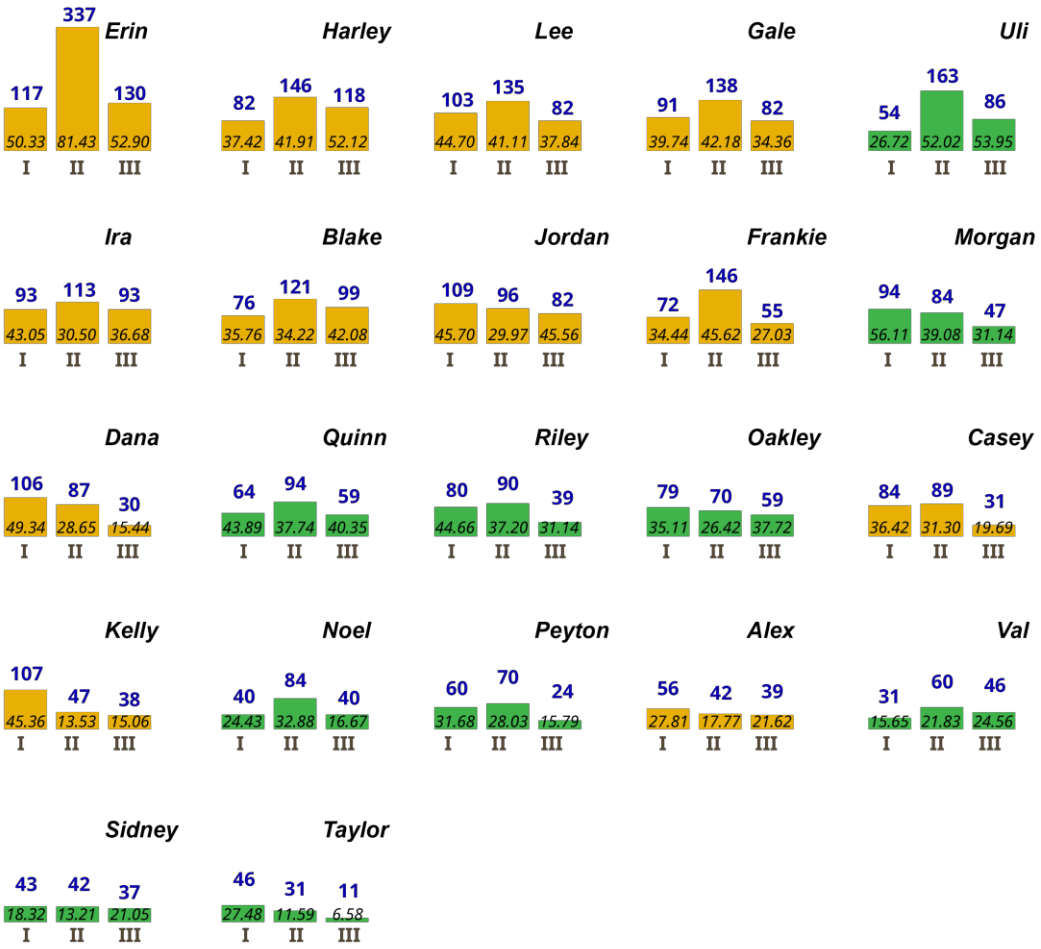


Figure 38. Term counts in individual glossaries, master glossaries, and diversity rates.

Figure 38 shows the number of terms compiled by each informant after the glossary tasks, displayed across cycles (numbers above bars) to compare them to the master glossary's entry counts (Cycle I, 95 entries; Cycle II, 96; Cycle III, 97) in terms of diversity rate (inside bars). For instance, Erin compiled 117 terms in Cycle I, and only one out of two terms (50.33%) were present in the master glossary (95 term counts). That is, half of the terms were not chosen by any other informant. As a reminder, no informant used InterpretBank in the first Cycle. Interestingly, Erin had comparable results in the third Cycle. The InterpretBank informants generally compiled more terms than those in the Excel group. Eight out of the ten informants with more terms belong to the InterpretBank group. The informants Erin, Harley, Lee, and Gale stand out for compiling the most terms on average. The chart reveals diverse patterns of term compilation among the informants over time. Several, including Erin, Harley, Frankie, Oakley, and Noel exhibited fluctuations in their term counts across cycles. The higher concentration in Cycle II might be due to the topic of the speech, but also because of the introduction of InterpretBank which, once again, was not used in Cycle I. Informants are split for the reader to be able to compare their baselines. Riley, Casey, Kelly, Alex, Sidney, and Taylor's term count steadily declined. In this case, InterpretBank and Excel informants are on a tie, so it might be an effect of getting used to or more confident about task demands.

3.1.2 Glossary review (*Memo* mode)

We explored the application of InterpretBank's *Memo* mode, which functions similarly to flashcards. Although the *Memo* mode was introduced before the Booth tasks, it reflects the informants' glossary review action. The use of *Memo* mode was captured through screen recording within a fixed 30 minutes for ad-hoc glossary preparation. In this stage, each informant was given the same 30-minute time frame to review the master glossary. The data in **Table 7**, calculated based on the seconds spent in Memo mode, was then normalized by dividing it by 30 minutes (equal to 1800 s) to yield percentage values. Informants who used Memo mode activated it with *start manual*, both in Cycles II and III, rather than *autoplay* (see the introduction of *Memo mode* in **§ 2.3.2**).

name	*Memo* duration (s) in Cycle II	percentage	*Memo* duration (s) in Cycle III	percentage
Alex	865.0	48.06	988.5	54.92
Blake	N/A		1345.0	74.72
Casey	427.8	23.77	N/A	
Dana	N/A		N/A	
Erin	623.0	34.61	N/A	
Frankie	N/A		596.4	33.13
Gale	459.7	25.54	N/A	
Harley	1049.1	58.29	N/A	
Ira	495.1	27.5	325.5	18.08
Jordan	503.7	27.98	1222.7	67.93
Kelly	N/A		N/A	
Lee	333.7	18.54	426.2	23.68

Table 7. Memo mode usage in Cycle II and Cycle III.

In Cycle II, varied engagement with the Memo feature was observed among the informants. Harley used it most (58.29% of the preparation time), closely followed by Alex (48.06%). On the lower end, and apart from four informants—Blake, Dana, Frankie, and Kelly—who did not use it at all (one-third of InterpretBank informants), Lee's share of time spent engaged in the memo feature reached only 18.54%. The medium-level users included Erin, Ira, and Jordan, who engaged with the mode for 34.61%, 27.50%, and 27.98% of their preparation time, respectively, while Casey and Gale utilized it for 23.77% and 25.54%. In Cycle III there were shifts in the use of Memo. Alex increased her engagement to 54.92%. Remarkably, Blake, who had not used the mode in Cycle II, dedicated a significant 74.72% of her preparation time to it.[17] Frankie, another new user, allocated 33.13% of her time to the mode. Conversely, Ira's engagement decreased to 18.08%, while Jordan's usage markedly rose to 67.93%. Lee's usage remained consistent at 23.68%. The number of informants not using the mode increased to six, or half of the participants (6 out of 12).

3.2 Fluency analysis

The collected data comprises a range of indicators, including but not limited to *false starts, corrections, fillers, repetitions, bumps, respites, EVS1,* and *EVS2*. **Table 8** portrays group aggregated values, and includes statistical measures such as Mean, Median, Mode, Standard Deviation (SD), Minimum, Maximum, Shapiro-Wilk *W*, and p-value for each variable and cycle. As a reminder, two informants are missing from the analysis, one from each group, because they chose to use the opposite tool in Cycle III.

									Shapiro-Wilk	
	cycle	group	Mean	Median	Mode*	SD	Min.	Max.	W	p-value
false starts	I	**XL**	4.0	3	3	3.6	0	13	0.83	0.03
		IB	3.3	4	0	2.6	0	8	0.93	0.43
	II	**XL**	3.6	3	1	3.5	0	12	0.85	0.05
		IB	3.5	3	1	2.7	0	9	0.94	0.46
	III	**XL**	3.9	3.5	2	2.7	0	10	0.92	0.39
		IB	2.7	2	4	2.7	0	9	0.85	0.05
Self-corrections	I	**XL**	12.6	11	9	7.0	4	28	0.86	0.07
		IB	8.4	8	8	2.0	5	11	0.91	0.23
	II	**XL**	13.5	12	12	7.9	2	29	0.96	0.78
		IB	11.8	10	8	8.8	3	36	0.74	0.00
	III	**XL**	11.8	10.5	10	6.0	3	23	0.97	0.89
		IB	9.3	8	6	5.4	2	20	0.93	0.43
fillers	I	**XL**	34.8	19.5	29	50.7	4	172	0.63	0.00
		IB	12.8	10	21	11.6	0	35	0.91	0.25
	II	**XL**	34.2	23.5	2	39.0	2	137	0.71	0.00
		IB	11.8	11	3	11.4	1	42	0.78	0.01
	III	**XL**	28.1	21.5	21	27.3	0	96	0.82	0.02
		IB	10.7	6	0	11.1	0	34	0.86	0.06
repetitions	I	**XL**	4.6	4	3	2.1	2	8	0.88	0.15
		IB	5.2	4	4	2.2	3	11	0.76	0.00
	II	**XL**	3.7	3.5	5	2.5	0	9	0.95	0.61
		IB	4.3	4	4	1.1	2	6	0.92	0.29

[17] Gender was not considered a variable through this project, and fake informants' names strived to be gender neutral. The pronouns do not necessarily represent the gender of the informant.

									Shapiro-Wilk	
	cycle	group	Mean	Median	Mode*	SD	Min.	Max.	W	p-value
	III	XL	3.2	2.5	1	2.7	0	8	0.91	0.26
		IB	2.9	2	0	2.9	0	10	0.86	0.06
bumps	I	XL	83.6	72	53	31.2	53	134	0.84	0.04
		IB	98.8	77	73	41.8	54	172	0.77	0.00
	II	XL	97.2	93	57	37.9	57	192	0.82	0.02
		IB	120.7	117	73	27.3	73	163	0.94	0.50
	III	XL	98.0	92	91	29.2	53	149	0.96	0.78
		IB	108.0	110	105	21.0	54	135	0.84	0.03
respites	I	XL	166.7	160	141	22.9	141	208	0.85	0.07
		IB	161.9	157	111	27.5	111	224	0.92	0.29
	II	XL	179.4	179.5	188	18.5	147	210	0.98	0.95
		IB	176.1	174	119	32.9	119	238	0.99	1.00
	III	XL	175.2	175.5	164	21.6	142	210	0.98	0.96
		IB	160.9	159	136	21.2	136	201	0.94	0.57
EVS1	I	XL	5.184	5.223	3.559	1.0	3.559	6.777	0.97	0.85
		IB	4.565	4.585	3.514	0.6	3.514	5.467	0.96	0.81
	II	XL	4.689	4.371	3.376	0.8	3.376	5.697	0.87	0.10
		IB	6.010	5.564	4.001	2.3	4.001	11.663	0.79	0.01
	III	XL	4.502	4.502	3.124	0.7	3.124	5.421	0.95	0.70
		IB	4.393	4.253	3.636	0.7	3.636	5.852	0.92	0.29
EVS2	I	XL	4.433	4.359	2.91	1.1	2.91	6.175	0.95	0.67
		IB	3.772	3.55	2.545	0.8	2.545	5.174	0.94	0.57
	II	XL	3.748	3.686	2.071	1.0	2.071	5.561	0.99	1.00
		IB	5.347	4.484	2.978	2.7	2.978	12.415	0.74	0.00
	III	XL	3.937	3.736	3.04	0.7	3.04	4.953	0.92	0.36
		IB	3.815	3.384	3.018	1.0	3.018	6.255	0.79	0.01

* More than one mode exists, only the first is reported

Table 8. Fluency indicators across cycles for two groups.

The Excel group exhibited a slight decrease in the mean number of *false starts* from Cycle I to Cycle III (for individual differences, see **§ 3.2.1** and **Figures 39 to 41**). The Shapiro-Wilk test indicated that the data were not normally distributed in Cycles I and II but became more so in Cycle III. The InterpretBank group also showed a decline, but their data was generally more normally distributed across all cycles.

For *self-corrections,* the Excel group showed a relatively stable mean number, and their data were normally distributed in Cycles II and III (Shapiro-Wilk, see individual details in **§ 3.2.2** and **Figures 42 to 44**). The InterpretBank group experienced an increase in the mean number of *self-corrections* in Cycle II and they decreased to original numbers in Cycle III and their data were generally not normally distributed, particularly in Cycle II, where the p-value was significantly low.

Regarding *filler*s, the Excel group's mean numbers remained fairly constant across cycles and the data were not normally distributed. The InterpretBank group's mean number of *filler*s consistently declined from Cycle I through Cycle III, and their data were more normally distributed (for individual data, see **§ 3.2.3** and **Figures 45 to 47**). Additionally, both groups showed a declining trend in the mean number of *repetition*s, from Cycle I to Cycle III (see also individual details in **§ 3.2.4** and **Figures 48 to 50**).

The mean number of *bumps* increased in the Excel group from Cycle I to Cycle III, whereas it first increased and then decreased in the InterpretBank group As for *respites*, both groups exhibited an increase until Cycle II, followed by a decline in Cycle III, forming a bell-shaped curve. The data for the two groups in both of these variables were mostly normally distributed, except for a few cycles where the p-value indicated otherwise (for individual differences, see **§ 3.2.5** and **Figures 51 to 53** for *bumps*; see **§ 3.2.6** and **Figures 54 to 56** for *respites*).

For *EVS1* and *EVS2*, the data were generally more stable and normally distributed across all cycles for both groups, although there were some fluctuations (for individual differences, see **§ 3.2.7** and **Figures 57 to 59** for *EVS1*; see **§ 3.2.8** and **Figures 60 to 62** for *EVS2*).

In summary, in terms of average number, the Excel group produced noticeably more *false starts*, *self-corrections*, *fillers*, and *respites*, than the InterpretBank group in all cycles. The Excel group produced fewer *bumps* than the InterpretBank group in all cycles. The data for most fluency indicators is not normally distributed. At this stage, it remains to be seen whether these differences in descriptive statistics are significant.

Before we address individual differences, let us see the overall bird's eye view of their data in all fluency variables (see **Table 9**).

group	Cycle	name	false starts		fillers		self-corrections		repetitions		dropped sentences		bumps		respites		*EVS1*		*EVS2*	
			count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	second	*z*	second	*z*
IB	I	Alex	**2**	*-0.50*	**21**	*-0.09*	**8**	*-0.51*	**4**	*-0.10*	**20**	*2.35*	**75**	*-0.70*	**157**	*-0.53*	**4.526**	*-0.29*	**4.393**	*0.17*
		Blake	**6**	*0.90*	**15**	*-0.27*	**6**	*-0.81*	**4**	*-0.10*	**4**	*-1.51*	**154**	*1.63*	**224**	*2.10*	**4.906**	*0.01*	**4.384**	*0.17*
		Casey	**4**	*0.20*	**107**	*2.56*	**16**	*0.69*	**12**	*2.90*	**17**	*1.63*	**35**	*-1.89*	**153**	*-0.69*	**4.096**	*-0.63*	**2.577**	*-1.08*
		Dana	**0**	*-1.19*	**1**	*-0.70*	**5**	*-0.96*	**4**	*-0.10*	**13**	*0.66*	**73**	*-0.76*	**163**	*-0.30*	**4.485**	*-0.32*	**3.361**	*-0.54*
		Erin	**0**	*-1.19*	**26**	*0.07*	**8**	*-0.51*	**7**	*1.03*	**6**	*-1.02*	**73**	*-0.76*	**151**	*-0.77*	**4.585**	*-0.24*	**3.320**	*-0.57*
		Frankie	**4**	*0.20*	**4**	*-0.61*	**8**	*-0.51*	**6**	*0.65*	**13**	*0.66*	**85**	*-0.41*	**179**	*0.33*	**4.021**	*-0.69*	**4.229**	*0.06*
		Gale	**1**	*-0.84*	**10**	*-0.43*	**9**	*-0.36*	**11**	*2.53*	**19**	*2.11*	**54**	*-1.32*	**111**	*-2.34*	**5.090**	*0.16*	**3.550**	*-0.41*
		Harley	**5**	*0.55*	**5**	*-0.58*	**7**	*-0.66*	**5**	*0.28*	**16**	*1.39*	**161**	*1.84*	**156**	*-0.57*	**5.467**	*0.46*	**3.104**	*-0.72*
		Ira	**8**	*1.59*	**35**	*0.34*	**11**	*-0.06*	**4**	*-0.10*	**14**	*0.91*	**76**	*-0.67*	**149**	*-0.85*	**3.514**	*-1.10*	**2.545**	*-1.10*
		Jordan	**5**	*0.55*	**21**	*-0.09*	**11**	*-0.06*	**5**	*0.28*	**11**	*0.18*	**87**	*-0.35*	**175**	*0.17*	**4.908**	*0.02*	**4.376**	*0.16*
		Kelly	**4**	*0.20*	**3**	*-0.64*	**11**	*-0.06*	**3**	*-0.47*	**16**	*1.39*	**77**	*-0.64*	**170**	*-0.02*	**3.867**	*-0.82*	**3.057**	*-0.75*
		Lee	**1**	*-0.84*	**0**	*-0.73*	**8**	*-0.51*	**4**	*-0.10*	**8**	*-0.54*	**172**	*2.16*	**146**	*-0.96*	**4.846**	*-0.03*	**5.174**	*0.71*
XL	I	Morgan	**6**	*0.90*	**24**	*0.00*	**13**	*0.24*	**5**	*0.28*	**14**	*0.91*	**56**	*-1.26*	**181**	*0.41*	**5.869**	*0.78*	**4.447**	*0.21*
		Noel	**0**	*-1.19*	**55**	*0.96*	**4**	*-1.11*	**4**	*-0.10*	**15**	*1.15*	**59**	*-1.18*	**141**	*-1.16*	**5.247**	*0.29*	**3.595**	*-0.38*
		Oakley	**5**	*0.55*	**9**	*-0.46*	**28**	*2.49*	**6**	*0.65*	**8**	*-0.54*	**126**	*0.80*	**208**	*1.47*	**5.289**	*0.32*	**5.136**	*0.69*
		Peyton	**3**	*-0.15*	**6**	*-0.55*	**8**	*-0.51*	**8**	*1.40*	**6**	*-1.02*	**81**	*-0.53*	**204**	*1.31*	**6.777**	*1.51*	**6.175**	*1.41*
		Quinn	**2**	*-0.50*	**15**	*-0.27*	**12**	*0.09*	**4**	*-0.10*	**8**	*-0.54*	**120**	*0.63*	**154**	*-0.65*	**4.591**	*-0.24*	**4.315**	*0.12*
		Riley	**4**	*0.20*	**29**	*0.16*	**10**	*-0.21*	**3**	*-0.47*	**19**	*2.11*	**134**	*1.04*	**159**	*-0.45*	**3.559**	*-1.06*	**4.402**	*0.18*
		Sidney	**3**	*-0.15*	**5**	*-0.58*	**9**	*-0.36*	**3**	*-0.47*	**15**	*1.15*	**65**	*-1.00*	**147**	*-0.93*	**6.552**	*1.33*	**5.840**	*1.17*
		Taylor	**1**	*-0.84*	**29**	*0.16*	**12**	*0.09*	**3**	*-0.47*	**10**	*-0.06*	**68**	*-0.91*	**178**	*0.29*	**5.198**	*0.25*	**4.286**	*0.10*
		Uli	**3**	*-0.15*	**4**	*-0.61*	**9**	*-0.36*	**2**	*-0.85*	**18**	*1.87*	**53**	*-1.35*	**162**	*-0.34*	**4.572**	*-0.25*	**2.910**	*-0.85*
		Val	**13**	*3.33*	**172**	*4.56*	**21**	*1.44*	**8**	*1.40*	**12**	*0.42*	**54**	*-1.32*	**153**	*-0.69*	**4.184**	*-0.56*	**3.221**	*-0.63*

group	Cycle	name	false starts		fillers		self-corrections		repetitions		dropped sentences		bumps		respites		*EVS1*		*EVS2*	
			count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	count	*z*	second	*z*	second	*z*
IB	II	Alex	**0**	*-1.19*	**11**	*-0.40*	**3**	*-1.26*	**4**	*-0.10*	**8**	*-0.54*	**117**	*0.54*	**174**	*0.13*	**4.075**	*-0.65*	**3.299**	*-0.58*
		Blake	**1**	*-0.84*	**11**	*-0.40*	**7**	*-0.66*	**6**	*0.65*	**8**	*-0.54*	**120**	*0.63*	**238**	*2.65*	**8.627**	*2.98*	**8.065**	*2.71*
		Casey	**3**	*-0.15*	**2**	*-0.67*	**8**	*-0.51*	**8**	*1.40*	**12**	*0.42*	**98**	*-0.02*	**152**	*-0.73*	**5.882**	*0.79*	**3.851**	*-0.20*
		Dana	**3**	*-0.15*	**3**	*-0.64*	**11**	*-0.06*	**4**	*-0.10*	**15**	*1.15*	**100**	*0.04*	**202**	*1.23*	**4.001**	*-0.71*	**4.346**	*0.14*
		Erin	**1**	*-0.84*	**13**	*-0.33*	**13**	*0.24*	**5**	*0.28*	**6**	*-1.02*	**73**	*-0.76*	**142**	*-1.12*	**5.920**	*0.82*	**3.995**	*-0.10*
		Frankie	**5**	*0.55*	**1**	*-0.70*	**8**	*-0.51*	**4**	*-0.10*	**7**	*-0.78*	**105**	*0.18*	**183**	*0.49*	**5.921**	*0.82*	**5.369**	*0.85*
		Gale	**3**	*-0.15*	**12**	*-0.36*	**8**	*-0.51*	**5**	*0.28*	**13**	*0.66*	**109**	*0.30*	**165**	*-0.22*	**5.941**	*0.84*	**4.484**	*0.24*
		Harley	**5**	*0.55*	**42**	*0.56*	**17**	*0.84*	**5**	*0.28*	**11**	*0.18*	**125**	*0.77*	**119**	*-2.02*	**11.663**	*5.41*	**12.415**	*5.72*
		Ira	**4**	*0.20*	**20**	*-0.12*	**6**	*-0.81*	**3**	*-0.47*	**13**	*0.66*	**107**	*0.24*	**155**	*-0.61*	**4.933**	*0.04*	**4.054**	*-0.06*
		Jordan	**9**	*1.94*	**8**	*-0.49*	**11**	*-0.06*	**4**	*-0.10*	**11**	*0.18*	**163**	*1.90*	**179**	*0.33*	**5.276**	*0.31*	**4.591**	*0.31*
		Kelly	**6**	*0.90*	**6**	*-0.55*	**36**	*3.69*	**2**	*-0.85*	**9**	*-0.30*	**148**	*1.45*	**211**	*1.59*	**4.192**	*-0.56*	**2.978**	*-0.80*
		Lee	**1**	*-0.84*	**3**	*-0.64*	**10**	*-0.21*	**5**	*0.28*	**11**	*0.18*	**161**	*1.84*	**169**	*-0.06*	**5.564**	*0.54*	**5.224**	*0.75*
XL	II	Morgan	**5**	*0.55*	**43**	*0.59*	**14**	*0.39*	**5**	*0.28*	**5**	*-1.26*	**96**	*-0.08*	**219**	*1.90*	**5.576**	*0.55*	**4.206**	*0.05*
		Noel	**3**	*-0.15*	**41**	*0.53*	**9**	*-0.36*	**5**	*0.28*	**13**	*0.66*	**77**	*-0.64*	**180**	*0.37*	**5.633**	*0.59*	**4.558**	*0.29*
		Oakley	**12**	*2.99*	**13**	*-0.33*	**29**	*2.64*	**3**	*-0.47*	**9**	*-0.30*	**192**	*2.75*	**188**	*0.68*	**5.508**	*0.49*	**5.561**	*0.98*
		Peyton	**1**	*-0.84*	**10**	*-0.43*	**21**	*1.44*	**5**	*0.28*	**11**	*0.18*	**100**	*0.04*	**210**	*1.55*	**4.221**	*-0.53*	**3.743**	*-0.27*
		Quinn	**1**	*-0.84*	**22**	*-0.06*	**12**	*0.09*	**4**	*-0.10*	**6**	*-1.02*	**97**	*-0.05*	**178**	*0.29*	**4.351**	*-0.43*	**3.396**	*-0.51*
		Riley	**4**	*0.20*	**40**	*0.50*	**12**	*0.09*	**3**	*-0.47*	**10**	*-0.06*	**102**	*0.09*	**147**	*-0.93*	**4.000**	*-0.71*	**3.629**	*-0.35*
		Sidney	**1**	*-0.84*	**9**	*-0.46*	**2**	*-1.41*	**2**	*-0.85*	**13**	*0.66*	**70**	*-0.85*	**157**	*-0.53*	**4.391**	*-0.40*	**3.063**	*-0.74*
		Taylor	**0**	*-1.19*	**25**	*0.04*	**10**	*-0.21*	**1**	*-1.22*	**11**	*0.18*	**89**	*-0.29*	**169**	*-0.06*	**5.697**	*0.65*	**4.322**	*0.13*
		Uli	**3**	*-0.15*	**2**	*-0.67*	**6**	*-0.81*	**0**	*-1.60*	**14**	*0.91*	**71**	*-0.82*	**188**	*0.68*	**3.376**	*-1.21*	**2.071**	*-1.43*
		Val	**6**	*0.90*	**137**	*3.48*	**20**	*1.29*	**9**	*1.78*	**12**	*0.42*	**57**	*-1.23*	**179**	*0.33*	**4.137**	*-0.60*	**2.928**	*-0.84*

group	Cycle	name	false starts		fillers		self-corrections		repetitions		dropped sentences		bumps		respites		*EVS1*		*EVS2*	
			count	*Z*	count	*Z*	count	*Z*	count	*Z*	count	*Z*	count	*Z*	count	*Z*	second	*Z*	second	*Z*
IB	III	Alex	**0**	*-1.19*	**9**	*-0.46*	**4**	*-1.11*	**0**	*-1.60*	**5**	*-1.26*	**96**	*-0.08*	**177**	*0.25*	**4.306**	*-0.46*	**3.812**	*-0.23*
		Blake	**0**	*-1.19*	**19**	*-0.15*	**2**	*-1.41*	**1**	*-1.22*	**8**	*-0.54*	**107**	*0.24*	**201**	*1.19*	**3.636**	*-1.00*	**3.063**	*-0.74*
		Casey	**0**	*-1.19*	**105**	*2.50*	**26**	*2.19*	**10**	*2.15*	**9**	*-0.30*	**58**	*-1.21*	**190**	*0.76*	**4.260**	*-0.50*	**3.346**	*-0.55*
		Dana	**1**	*-0.84*	**0**	*-0.73*	**6**	*-0.81*	**1**	*-1.22*	**6**	*-1.02*	**115**	*0.48*	**184**	*0.53*	**3.833**	*-0.84*	**3.384**	*-0.52*
		Erin	**2**	*-0.50*	**5**	*-0.58*	**6**	*-0.81*	**4**	*-0.10*	**7**	*-0.78*	**105**	*0.18*	**136**	*-1.36*	**3.731**	*-0.92*	**3.018**	*-0.77*
		Frankie	**9**	*1.94*	**2**	*-0.67*	**10**	*-0.21*	**5**	*0.28*	**4**	*-1.51*	**112**	*0.39*	**159**	*-0.45*	**5.137**	*0.20*	**4.914**	*0.53*
		Gale	**0**	*-1.19*	**12**	*-0.36*	**8**	*-0.51*	**10**	*2.15*	**13**	*0.66*	**105**	*0.18*	**139**	*-1.24*	**4.888**	*0.00*	**3.157**	*-0.68*
		Harley	**4**	*0.20*	**26**	*0.07*	**17**	*0.84*	**4**	*-0.10*	**11**	*0.18*	**125**	*0.77*	**136**	*-1.36*	**5.852**	*0.77*	**6.255**	*1.46*
		Ira	**4**	*0.20*	**34**	*0.31*	**11**	*-0.06*	**0**	*-1.60*	**8**	*-0.54*	**110**	*0.33*	**150**	*-0.81*	**4.253**	*-0.51*	**3.742**	*-0.27*
		Jordan	**4**	*0.20*	**6**	*-0.55*	**11**	*-0.06*	**2**	*-0.85*	**7**	*-0.78*	**135**	*1.07*	**172**	*0.06*	**4.547**	*-0.27*	**4.120**	*-0.01*
		Kelly	**4**	*0.20*	**4**	*-0.61*	**20**	*1.29*	**3**	*-0.47*	**11**	*0.18*	**54**	*-1.32*	**151**	*-0.77*	**4.141**	*-0.60*	**3.150**	*-0.68*
		Lee	**2**	*-0.50*	**1**	*-0.70*	**7**	*-0.66*	**2**	*-0.85*	**6**	*-1.02*	**124**	*0.75*	**165**	*-0.22*	**3.997**	*-0.71*	**3.349**	*-0.55*
XL	III	Morgan	**5**	*0.55*	**21**	*-0.09*	**6**	*-0.81*	**3**	*-0.47*	**9**	*-0.30*	**78**	*-0.61*	**213**	*1.66*	**5.338**	*0.36*	**4.907**	*0.53*
		Noel	**2**	*-0.50*	**46**	*0.68*	**10**	*-0.21*	**0**	*-1.60*	**7**	*-0.78*	**93**	*-0.17*	**171**	*0.02*	**4.513**	*-0.30*	**3.737**	*-0.28*
		Oakley	**6**	*0.90*	**10**	*-0.43*	**23**	*1.74*	**7**	*1.03*	**5**	*-1.26*	**133**	*1.01*	**197**	*1.04*	**4.491**	*-0.32*	**4.534**	*0.27*
		Peyton	**0**	*-1.19*	**22**	*-0.06*	**19**	*1.14*	**4**	*-0.10*	**1**	*-2.23*	**122**	*0.69*	**210**	*1.55*	**3.124**	*-1.41*	**3.040**	*-0.76*
		Quinn	**3**	*-0.15*	**31**	*0.22*	**10**	*-0.21*	**1**	*-1.22*	**5**	*-1.26*	**149**	*1.48*	**164**	*-0.26*	**3.769**	*-0.89*	**3.734**	*-0.28*
		Riley	**2**	*-0.50*	**29**	*0.16*	**8**	*-0.51*	**5**	*0.28*	**10**	*-0.06*	**91**	*-0.23*	**142**	*-1.12*	**5.033**	*0.12*	**4.953**	*0.56*
		Sidney	**4**	*0.20*	**5**	*-0.58*	**11**	*-0.06*	**1**	*-1.22*	**6**	*-1.02*	**73**	*-0.76*	**164**	*-0.26*	**5.421**	*0.43*	**4.126**	*-0.01*
		Taylor	**4**	*0.20*	**21**	*-0.09*	**13**	*0.24*	**2**	*-0.85*	**6**	*-1.02*	**91**	*-0.23*	**184**	*0.53*	**4.193**	*-0.55*	**3.660**	*-0.33*
		Uli	**3**	*-0.15*	**0**	*-0.73*	**3**	*-1.26*	**1**	*-1.22*	**14**	*0.91*	**76**	*-0.67*	**180**	*0.37*	**4.702**	*-0.15*	**3.365**	*-0.54*
		Val	**10**	*2.29*	**96**	*2.22*	**15**	*0.54*	**8**	*1.40*	**8**	*-0.54*	**53**	*-1.35*	**148**	*-0.89*	**4.431**	*-0.36*	**3.316**	*-0.57*

Table 9. Overall fluency performance by each informant across Cycles.

In the next sections until **§ 3.2.4,** the alpha level is set at 0.05 for inferential statistics.

3.2.1 False starts

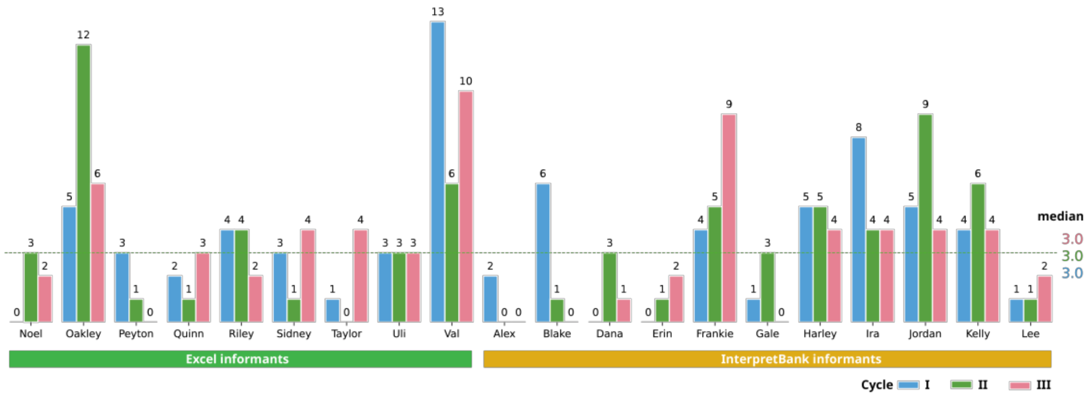


Figure 39. Distribution of false starts across cycles by informants.

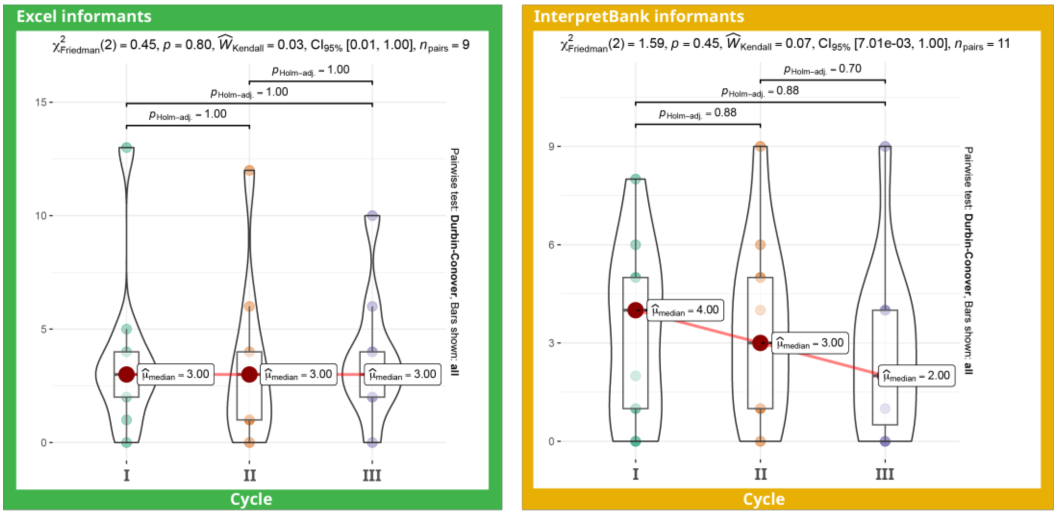


Figure 40. Within-group comparison of *false starts* across cycles.

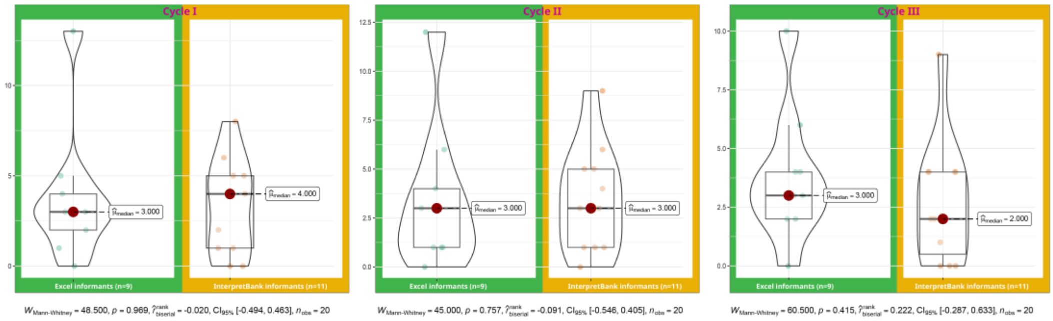


Figure 41. Comparison of false starts across cycles between groups.

In **Figure 39**, Oakley from the Excel group shows a steep rise in *false starts* from Cycle I to Cycle II and a similarly sharp decline from Cycle II to Cycle III, improving from 12 to 6. Val from the same group demonstrates an inverse pattern. Meanwhile, Erin and Frankie from the InterpretBank group exhibit a steady increase in false starts from Cycle I to Cycle III. The median lines indicate that the three cycles have the same median values. The analysis of the data in **Figure 40** using Friedman's ANOVA indicated that there is no statistically significant difference in the frequency of *false starts* across the three cycles within Excel group and InterpretBank group, x_r^2 (2, *N* = 9) = 0.45, p-value = 0.80 for Excel group, and x_r^2 (2, *N* = 11) = 1.59, p-value = 0.45 for InterpretBank group. On·the other hand, the Mann-Whitney U Test in **Figure 41** for Excel group and InterpretBank group indicates that p-value in all three cycles are well above the conventional alpha level of 0.05, suggesting that there is no statistically significant difference in the distribution of f*alse starts* between InterpretBank and Excel groups for any of the cycles.

3.2.2 Self-corrections

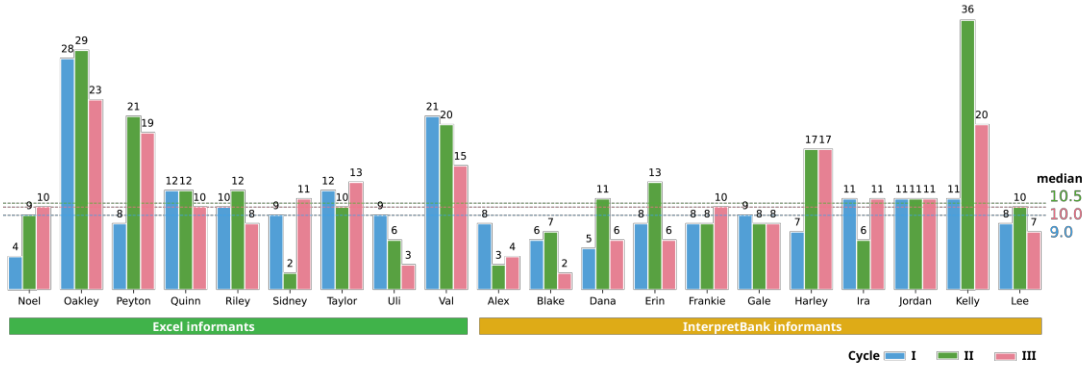


Figure 42. Distribution of *self-corrections* across cycles by informants.

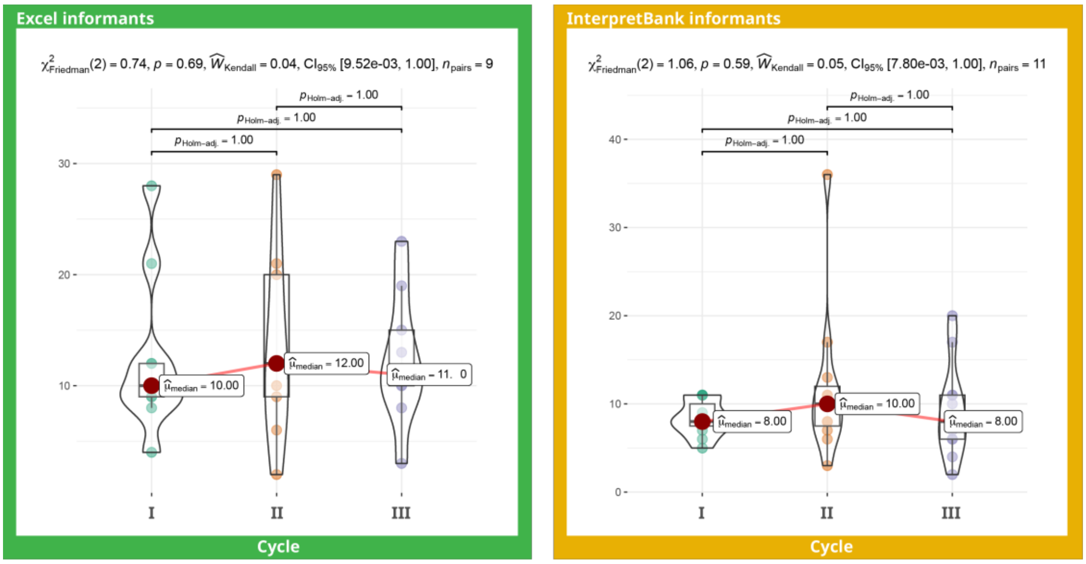


Figure 43. Within-group comparison of *self-corrections* across cycles.

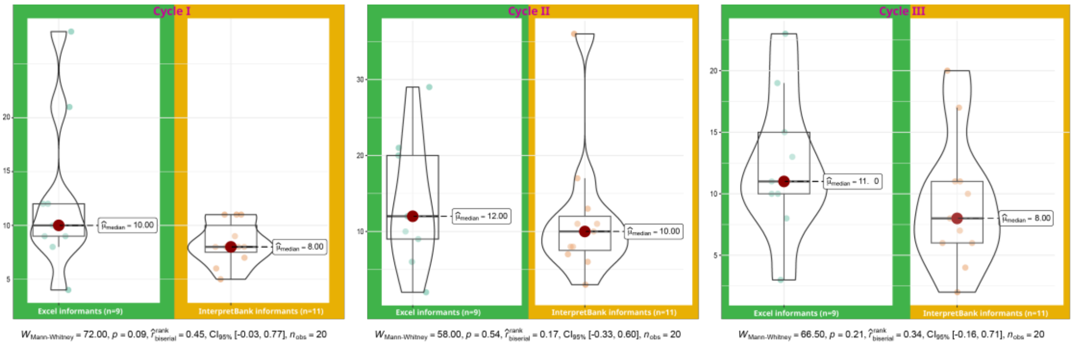


Figure 44. Comparison of *self-corrections* across cycles between groups.

Figure 42 shows the distribution of *self-corrections* across cycles for each informant in the Excel (left) and InterpretBank groups (right). There are significant variations among individuals within each group across cycles. For example, Kelly from the InterpretBank group shows many more *self-corrections* in Cycle II (from 11 in Cycle I to 36) and a less dramatic decrease to 20 in Cycle III. On the other hand, Quinn from the Excel group remained consistent at around 10 *self-corrections* across all cycles. The median also indicates that Cycle II has a higher number of *self-corrections* across informants. Six out of 12 InterpretBank informants experienced an increase in *self-corrections* in Cycle II (right after introducing InterpretBank, although Alex behaved the opposite), whereas only three Excel informants did. One possible explanation might be that using InterpretBank might have had an impact on the *self-corrections*, perhaps due to higher multitasking demands. For the within-group analysis, No significant changes were observed in *self-corrections* within either the Excel (x_r^2 (2, N = 9) = 0.74, p-value = 0.69) or InterpretBank (x_r^2 (2, N = 11) = 1.06, p-value = 0.59) groups over time (**Figure 43**). Mann-Whitney U tests revealed no significant differences in *self-corrections* between the two groups across cycles (see **Figure 44**), with p-values above 0.05 for all cycles.

3.2.3 Fillers

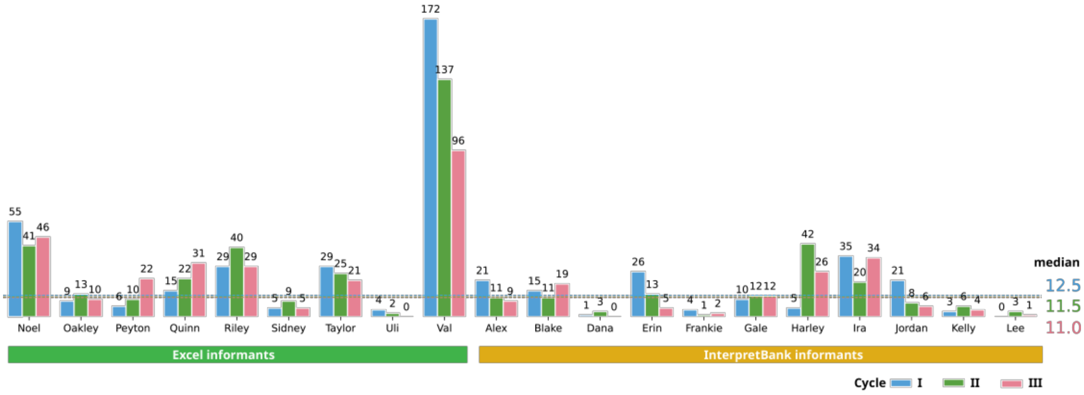


Figure 45. Distribution of fillers across cycles by informant.

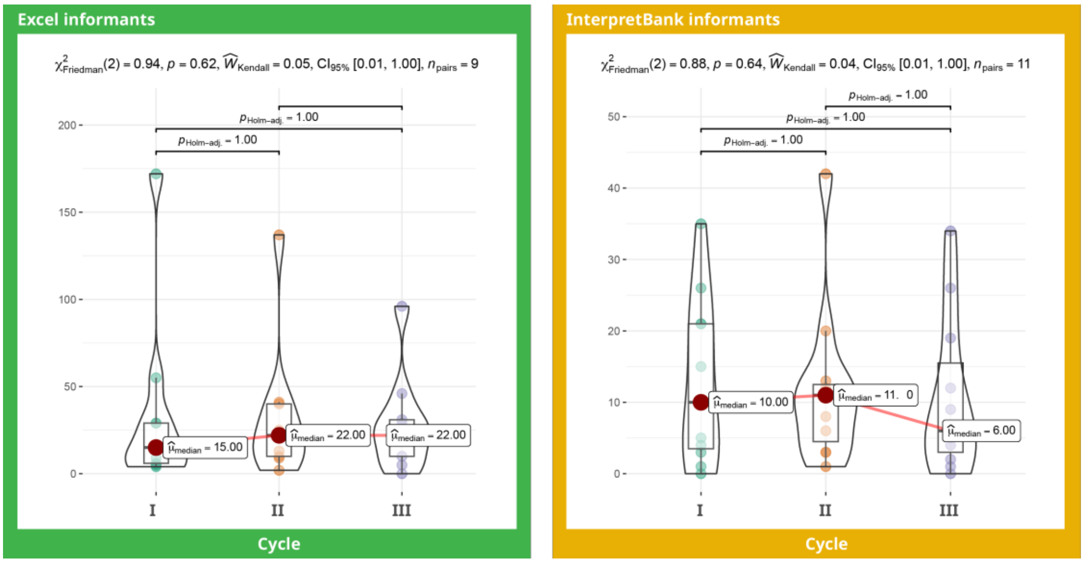


Figure 46. Within-group comparison of *fillers* across cycles.

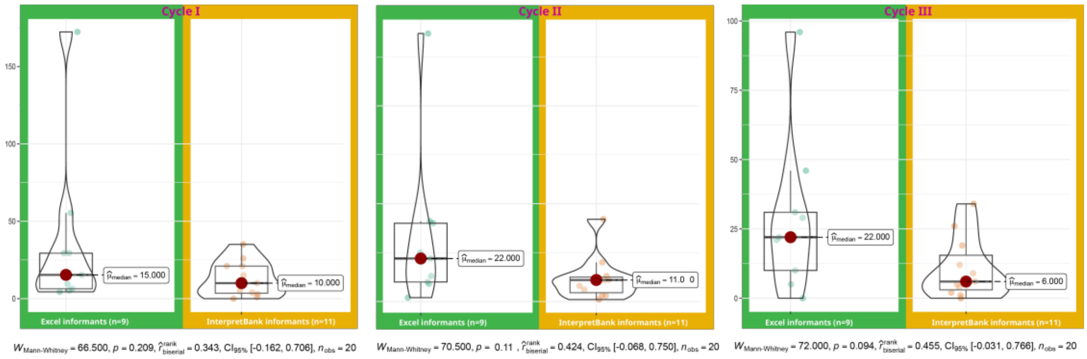


Figure 47. Comparison of *fillers* across cycles between groups.

Figure 45 reveals several outliers within the Excel group, particularly Val, who consistently exhibited a remarkably high frequency of fillers across all cycles compared to other participants in both groups. While some informants, such as Peyton and Quinn from the Excel group, demonstrated a steady increase in filler use from Cycle I to Cycle III, others, including Taylor, Val, and Erin from the Excel group, and Erin and Jordan from the InterpretBank group, exhibited a declining trend. Additionally, several informants showcased large variations over time, including Noel, Oakley, and Sydney in the Excel group, and Alex, Blake, Harley, and Ira in the InterpretBank group. In contrast, the remaining informants (Uli, Dana, Frankie, Kelly, and Lee from both groups) produced very few fillers. Despite these individual differences, the median performance remained relatively consistent across cycles for both groups. The statistical analysis in **Figures 47–48** further supports these observations. The Mann-Whitney U test confirmed that these differences were not statistically significant inter-group across cycles, both have a p-value greater than 0.05. The Friedman test yielded no significant temporal changes within-group across cycles, further supported by a p-value greater than 0.05, x_r^2 (2, N = 9) = 0.94, p-value = 0.62 for Excel group, and x_r^2 (2, N = 11) = 0.88, p-value = 0.64 for InterpretBank group.

3.2.4 Repetitions

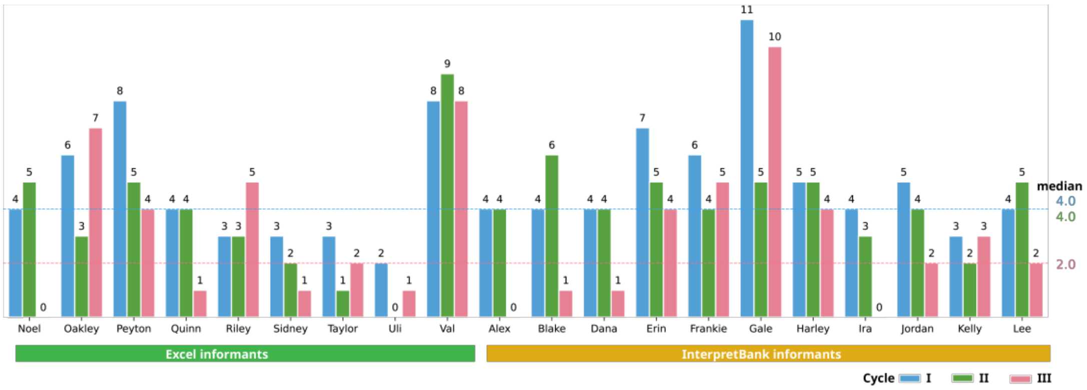


Figure 48. Distribution of *repetitions* across cycles by informants.

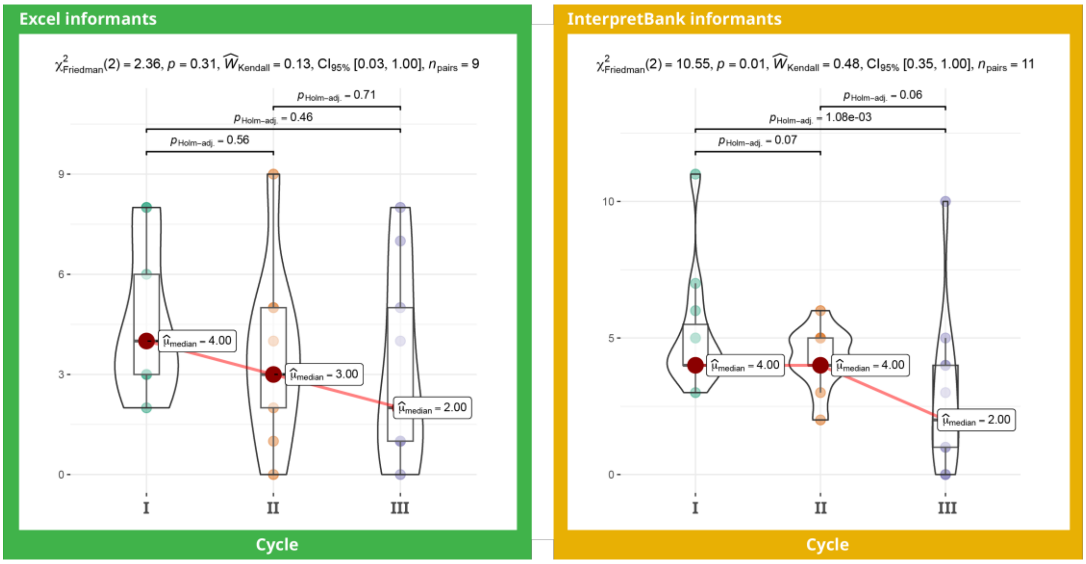


Figure 49. Within-group comparison of *repetitions* across cycles.

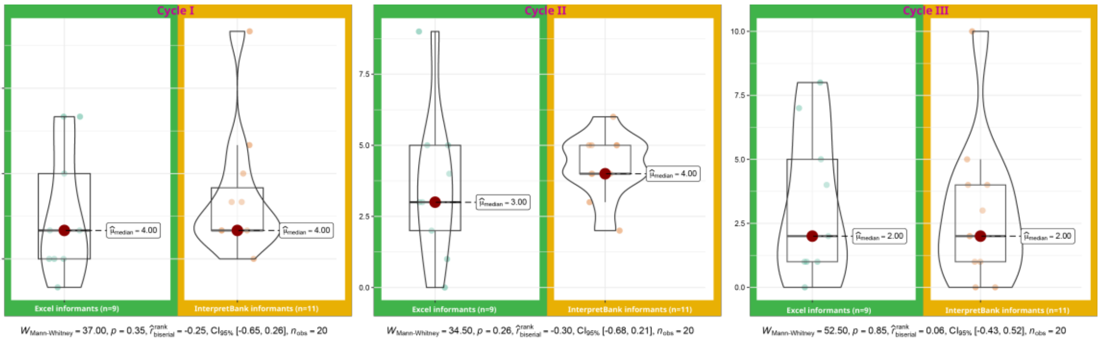


Figure 50. Comparison of *repetitions* across cycles between groups.

In the case of *repetitions,* **Figure 48** points to some noteworthy degree of individual variation within the InterpretBank group. Specifically, the data portrayed Gale as an outlier, with a substantially higher frequency of *repetitions* compared to other members within the same group. This diverges from the pattern observed in the Excel group, where Uli represents a comparable point of variation. Median values for *repetitions* in Cycles II and III are identical and exceed that of Cycle I. The Mann-Whitney U test in **Figure 50**, reveals no significant inter-group differences, both show *p* values above conventional alpha level of 0.05. The Friedman test supported this observation with the within-group comparison in the Excel group across cycles. x_r^2 (2, *N* = 9) = 2.36, p-value = 0.31 for Excel group. Conversely, the InterpretBank group exhibits significant intra-group variation across cycles, as evidenced by the value of chi-square, x_r^2 (2, *N* = 11) = 10.55, p-value < 0.05.

3.2.5 Bumps

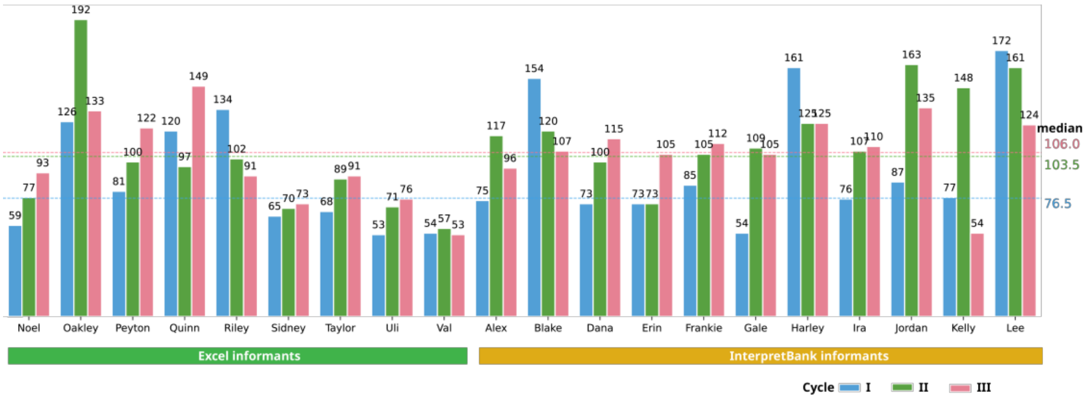


Figure 51. Distribution of *bumps* across cycles by informants.

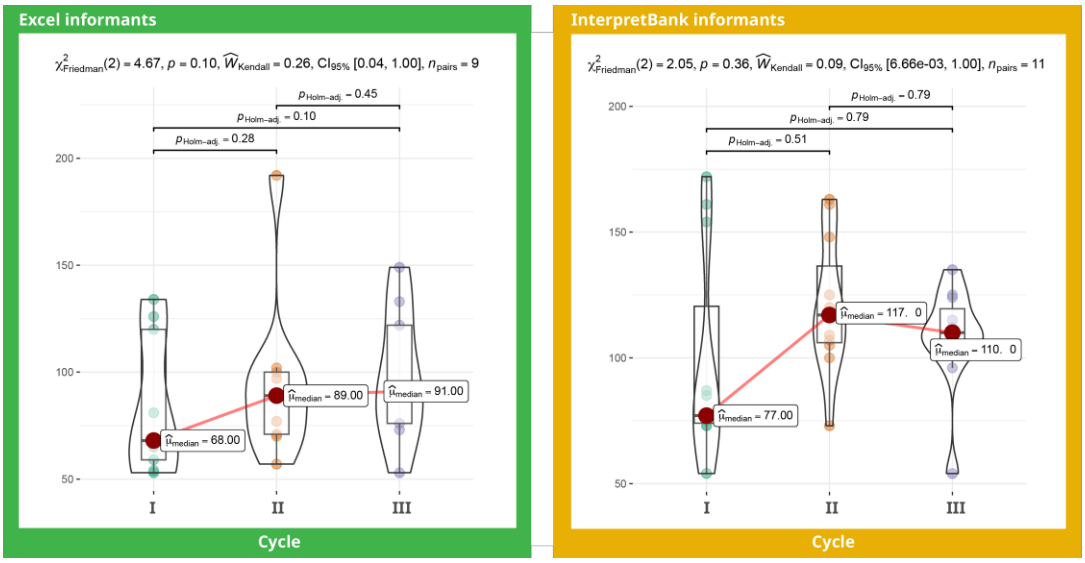


Figure 52. Within-group comparison of *bumps* across cycles.

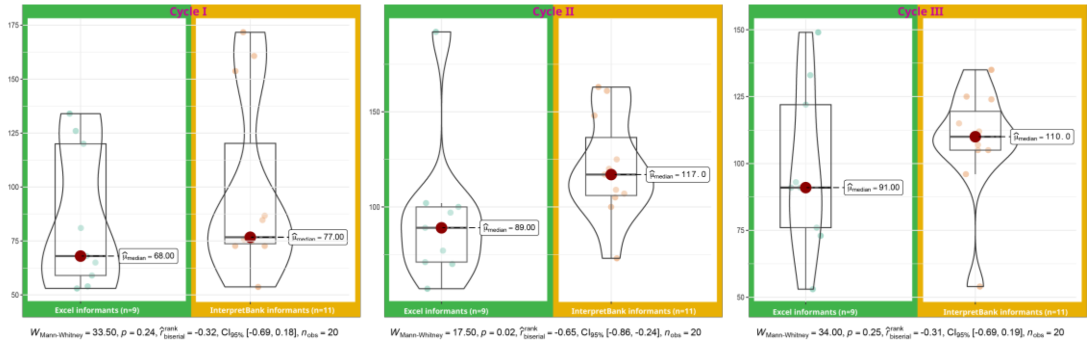


Figure 53. Comparison of *bumps* across cycles between groups.

Figure 51 clearly illustrates that Cycle II exhibited a broader distribution of bump occurrences for both the Excel group and InterpretBank group. However, Cycle III exhibited the highest frequency of bumps, surpassing the other two cycles. Especially in Cycle II, nine informants from the InterpretBank group registered bump occurrences above the median value of 103.5, while only one informant from the Excel group surpassed this threshold. In contrast, Cycle III saw seven InterpretBank informants and three Excel group informants exceeding the median value of 106. The Friedman test in **Figure 52** for the Excel group and InterpretBank group, however, showed no significant changes within-group from Cycle I to Cycle III. The Mann-Whitney U test in **Figure 53** revealed a statistically significant difference between the groups in Cycle II (p-value < 0.05), but not for Cycle I and Cycle III.

3.2.6 Respites

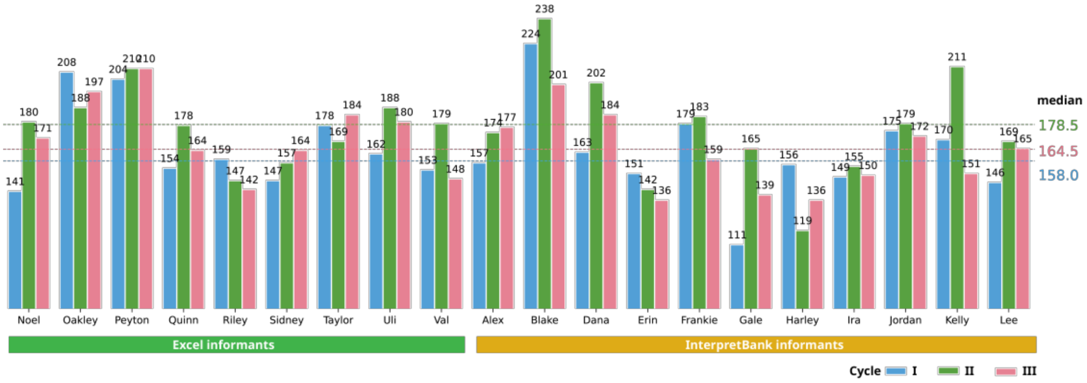


Figure 54. Distribution of *respites* across cycles by informants.

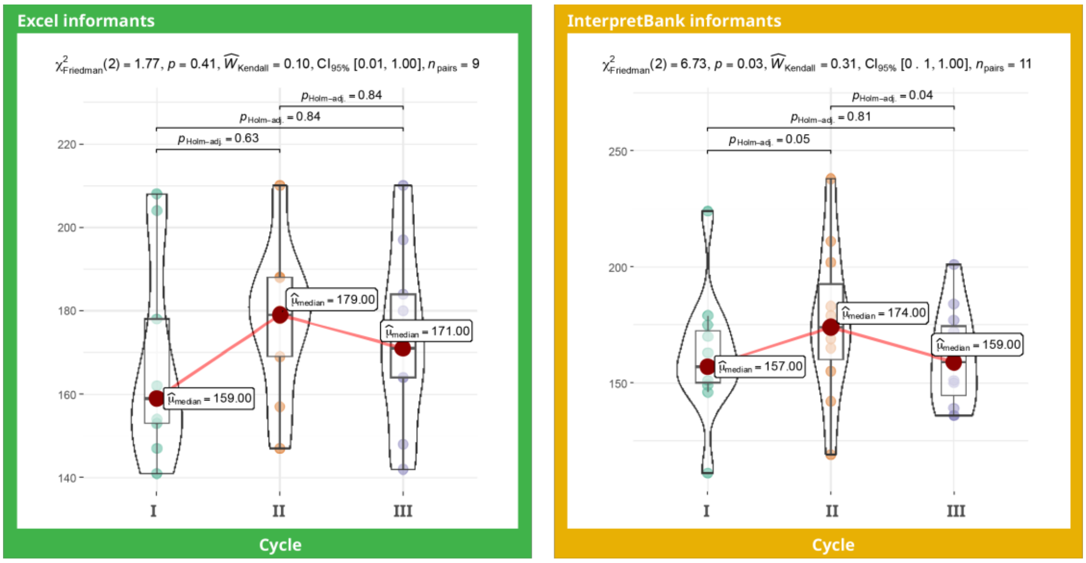


Figure 55. Within-group comparison of *respites* between groups.

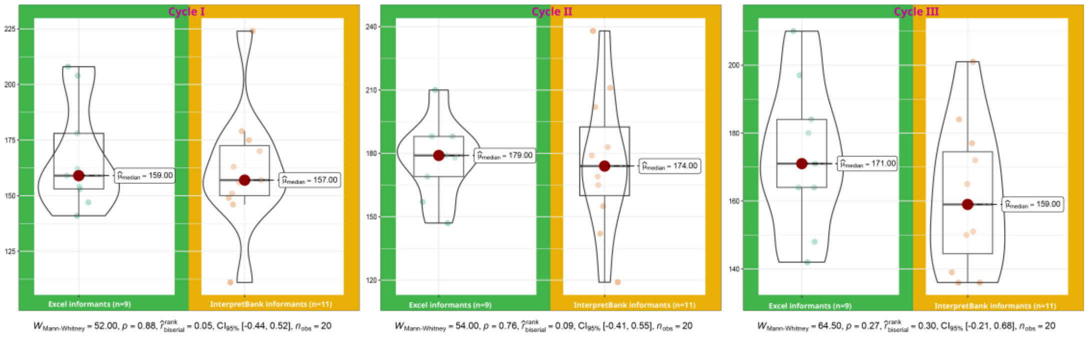


Figure 56. Comparison of *respites* across cycles between groups.

Figure 54 depicts a high number of *respite* production across all cycles for both groups of informants, as evidenced by the increasing median values. To enhance visual clarity, the Y-axis commences at 75 instead of zero. Overall, InterpretBank informants seem to have faced higher cognitive demands, probably related to multitasking. However, differences in respites were not so obvious, and smaller for InterpretBank informants. This is consistent with the hypothesized tendency of interpreters to avoid respites in any circumstances, for they may be noticed and thus impact negatively on the quality assessment of their performance. The within-group analysis of respites, the Friedman Test (see **Figure 55**) for the groups revealed no significant changes across cycles, with a p-value above 0.05. That is, no changes were apparent in group performance over time, except for the InterpretBank group between Cycles II and III, which might be explained by the InterpretBank informants' mastering their control over respites when using the new tool. No significant differences in respite production were noted between groups, as indicated by Mann-Whitney U test results (p-values: 0.88 for Cycle I, 0.79 for Cycle II, and 0.27 for Cycle III). All p-values exceeded the conventional alpha threshold of 0.05.

3.2.7 Chunk-initial EVS (*EVS1*)

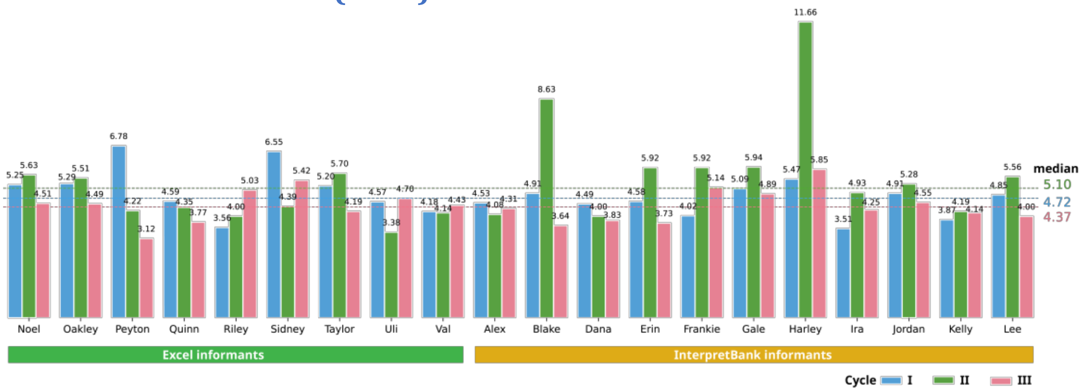


Figure 57. Distribution of *Chunk-initial EVS* across cycles by informants.

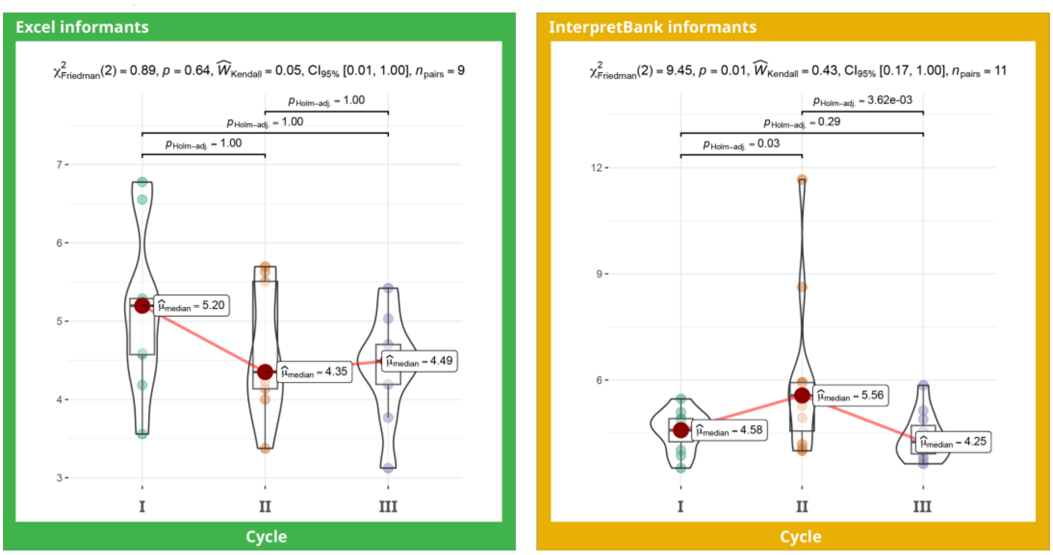


Figure 58. Within-group comparison of *Chunk-initial EVS* across cycles.

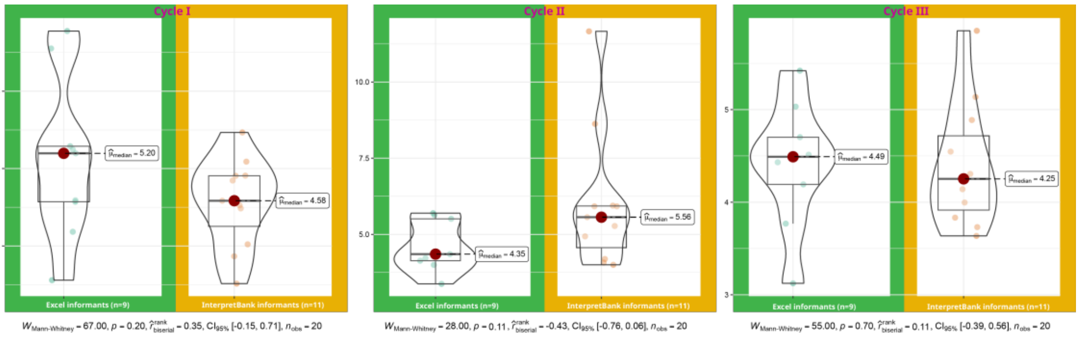


Figure 59. Comparison of *Chunk-initial EVS* across cycles between groups.

Figure 57 illustrates both individual and group-level *Chunk-initial EVS* or *EVS1* variation across the three cycles. There are slight fluctuations with respect to the overall median *EVS1* score. It starts at 4.72 s in Cycle I, increases to 5.10 s in Cycle

II, but *drops* slightly to 4.37 s in Cycle III. Within these overarching trends, individual performance differences stand out. In the Excel group, for instance, there stand out the *EVS1* values for Peyton's in Cycle I (6.78 s) and Taylor's in Cycle II (5.70 s). Sidney stole the limelight in Cycle III, with an *EVS1* of 5.42 s. In the InterpretBank group, Harley emerged as a consistent outlier, with scores of 5.47 s, 11.66 s, and 5.85 s in Cycles I, II, and III. The Mann-Whitney U test in **Figure 59** yielded p-values of 0.20, 0.11, and 0.70 for Cycles I, II, and III, respectively. These p-values were all above the conventional alpha level of 0.05, suggesting that the differences in *EVS1* between the Excel and InterpretBank groups were not statistically significant in any cycle. For within-group changes of *EVS1* over time, the Friedman Test revealed no statistically significant temporal changes in these metrics across cycles within this group, x_r^2 (2, N = 9) = 0.89, p-value = 0.64 for the Excel group. Within the InterpretBank group, however, there were statistically significant changes over the cycles, x_r^2 (2, N = 11) = 9.45, p-value < 0.05.

3.2.8 Chunk-final EVS (*EVS2*)

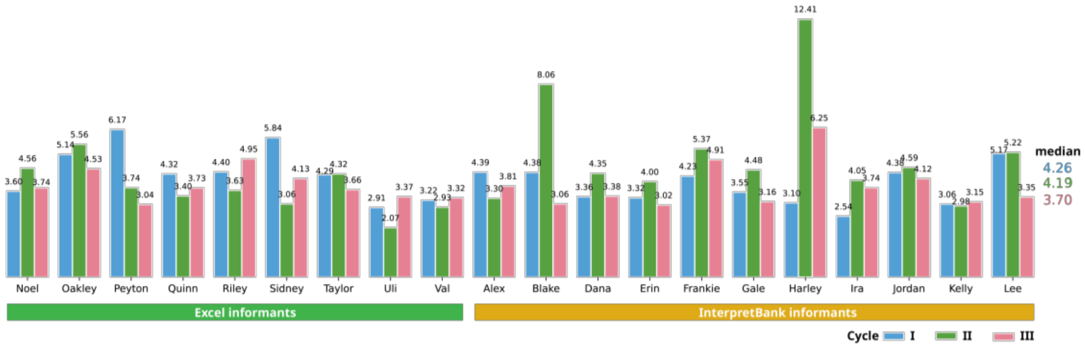


Figure 60. Distribution of *Chunk-final EVS* across cycles by informants.

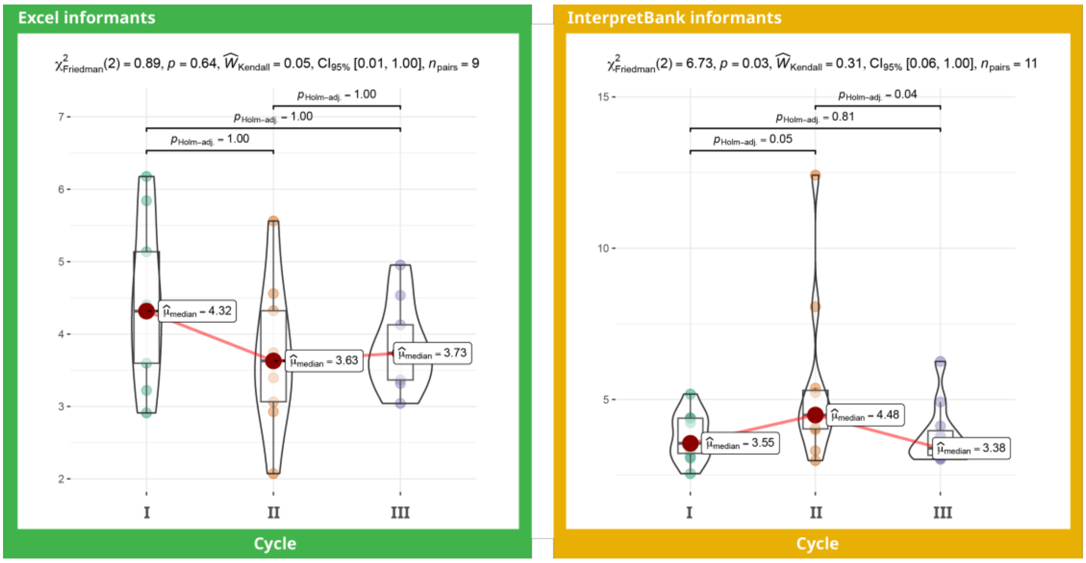


Figure 61. Within-group comparison of *Chunk-final EVS* across cycles.

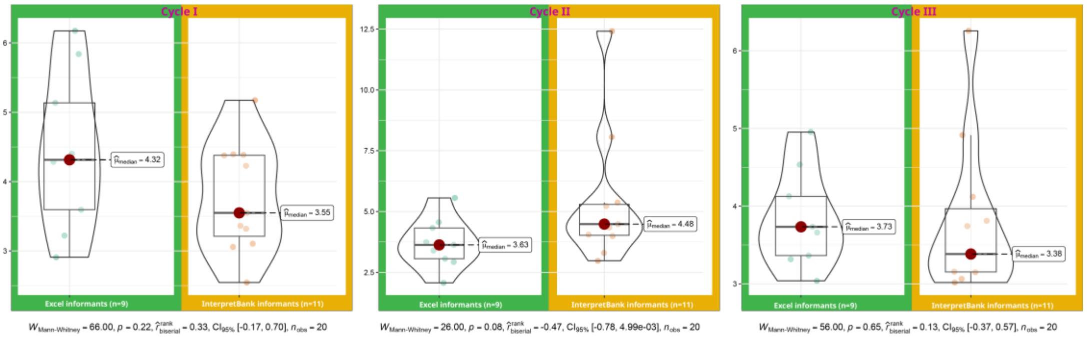


Figure 62. Comparison of *Chunk-final EVS* across cycles between groups.

Notable differences in individual performance can be observed in *Chunk-final EVS* or *EVS2* (**Figure 60)**. In the Excel group, Peyton's *EVS2* value of 6.17 s in Cycle I, Sidney's 5.84 s in Cycle I, and Oakley's 5.56 s in Cycle II stood out from the rest. On the other hand, in the InterpretBank group, many informants (Dana, Erin, Frankie, Gale, Ira, and Jordan) showed a significant increase in *EVS2* from Cycle I to Cycle II and a decrease from Cycle II to Cycle III. Harley, in particular, reached 12.41 s in Cycle II, far exceeding the others. Within groups (**Figure 61)**, the Friedman test for the Excel group showed no statistically significant changes in *EVS2* over the cycles, with a p-value of 0.64. The Friedman test for the InterpretBank group yielded a p-value of 0.03, below the alpha level of 0.05, indicating statistically significant temporal changes in the *EVS2* scores. The Mann-Whitney U test (see **Figure 62**) revealed no statistically significant differences between the *EVS2* values of the two groups across cycles, with p-values of 0.22, 0.08, and 0.65 for Cycles I, II, and III, respectively. An alpha level of 0.05 was used for the tests.

3.2.9 Duration of source speech chunks and EVS

Figures 63 and 64 illustrate the data from Cycle I for the Excel group and the InterpretBank group. The solid lines plot the median values for both the *EVS1* and *EVS2*. The median values show moderate fluctuations between the chunks, suggesting varying degrees of cognitive effort. The grey dots illustrate the lowest and highest measurements of *EVS1* and *EVS2*, and the dashed lines linking them depict the range. This range proved valuable to assess the consistency and variance in performance. The lower diagram shows the duration of each source speech chunk in Cycle I. For instance, in **Figure 63**, chunk 10 had a duration of 6.31 s in the source speech, as shown in the lower diagram. Corresponding to the upper diagram, we observed the duration of *EVS1* for nine informants in the Excel group. The minimum value of *EVS1* was 2.54 s, the maximum was 8.57 s, and the median value was 3.98 s.

In the first cycle, both groups *EVS1* and *EVS2* showed a trend toward fitting in the median values of several chunks at the beginning and end, and in the speech beginning section, they presented a gradual increase, suggesting that the informants adapted to the interpreting task. The maximum and minimum values of the two groups *EVS1* and *EVS2* were generally concentrated in the range of 0-10 seconds. However, it is not ruled out that there are individual chunks with outliers, for instance, the Excel group in **Figure 63** EVS 2 in chunks 19, 78, 82, and the InterpretBank group in **Figure 64**, *EVS2* in chunks 39, 48, both of which showed extreme values.

In Cycle II, the Excel group (**Figure 65**), the median values for *EVS1* and *EVS2* show less variation than in Cycle I. For instance, in Chunk 2, the *EVS1* ranged from 1.26 s to 5.32 s, with a median of 2.79 s, whereas *EVS2* spanned from 2.27 s to 7.53 s, with a median of 3.17 s. The InterpretBank group (**Figure 66**) exhibits noticeable changes in median values in Cycle II, especially for *EVS1*. The range of *EVS1* and *EVS2* between the minimum and maximum values remains stable (mostly concentrated within 0-10 s), reinforcing the notion of consistent performance. The second cycle revealed a continuation of this trend from Cycle I. **Figures 67–68** illustrate a similar pattern observed in two groups during *EVS1* and *EVS2*, paralleling earlier observations in Cycles I and II. The minimum and maximum value range continues to concentrate within the 0-10 s range. The two groups, as before, show an upward trend in the initial chunks for both *EVS1* and *EVS2*.

In the present case, the absolute value of the outliers is smaller than before, and there is a difference in the number of outliers between the two groups. The Excel group has fewer outliers, which appear in *EVS2* chunks 39 and 54. Additionally, there are fewer *EVS2* values that fall within the range of 0 to -5 s in the Excel group. In contrast, the InterpretBank group has more frequent outliers, appearing in *EVS2* chunks such as 33, 38, 44, 57, and 63, among others. Furthermore, multiple chunks of *EVS2* values in the InterpretBank group fall within the range of 0 to −5 s.

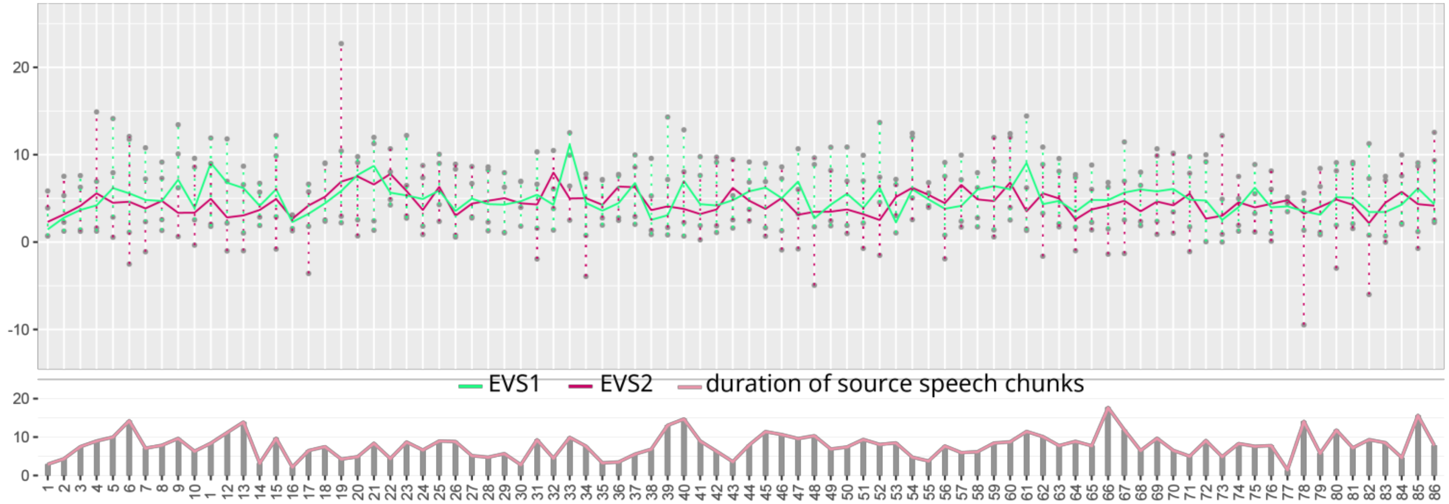


Figure 63. *EVS1*, *EVS2*, and chunks duration in Cycle I (Excel Group).

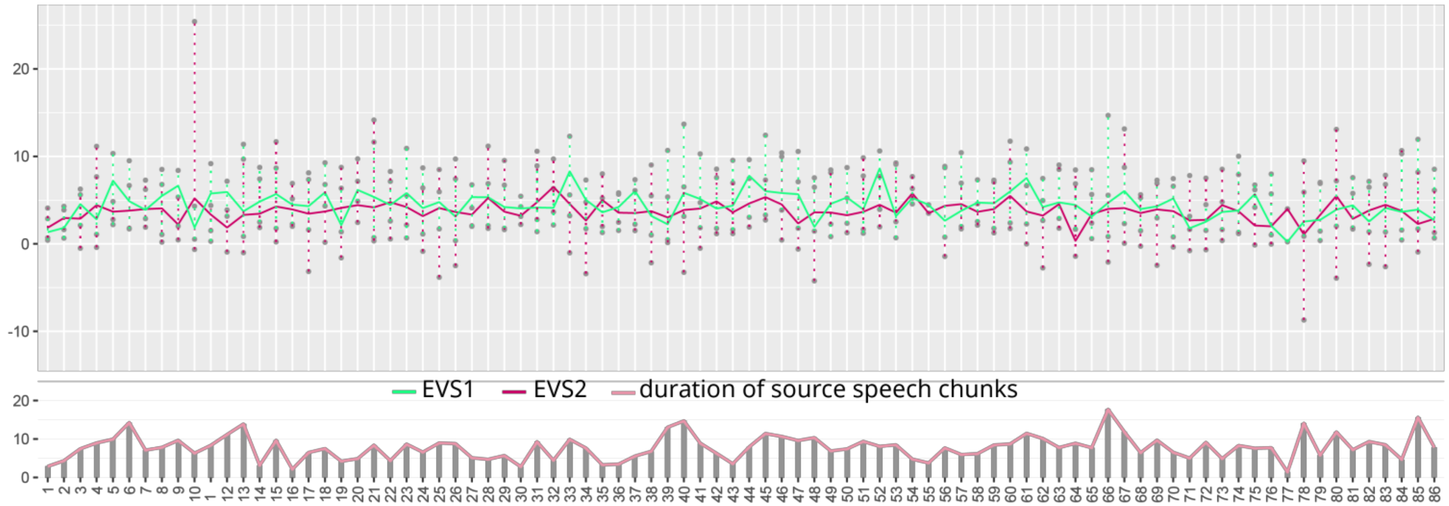


Figure 64. *EVS1*, *EVS2*, and chunks duration in Cycle I (InterpretBank Group).

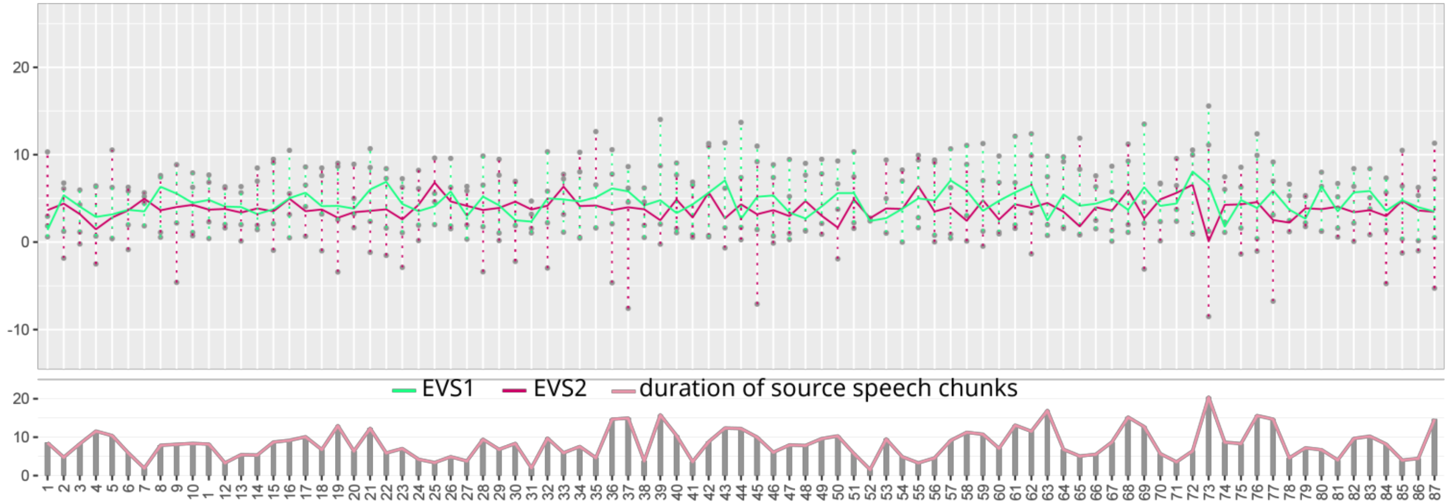


Figure 65. *EVS1*, *EVS2*, and chunks duration in Cycle II (Excel Group).

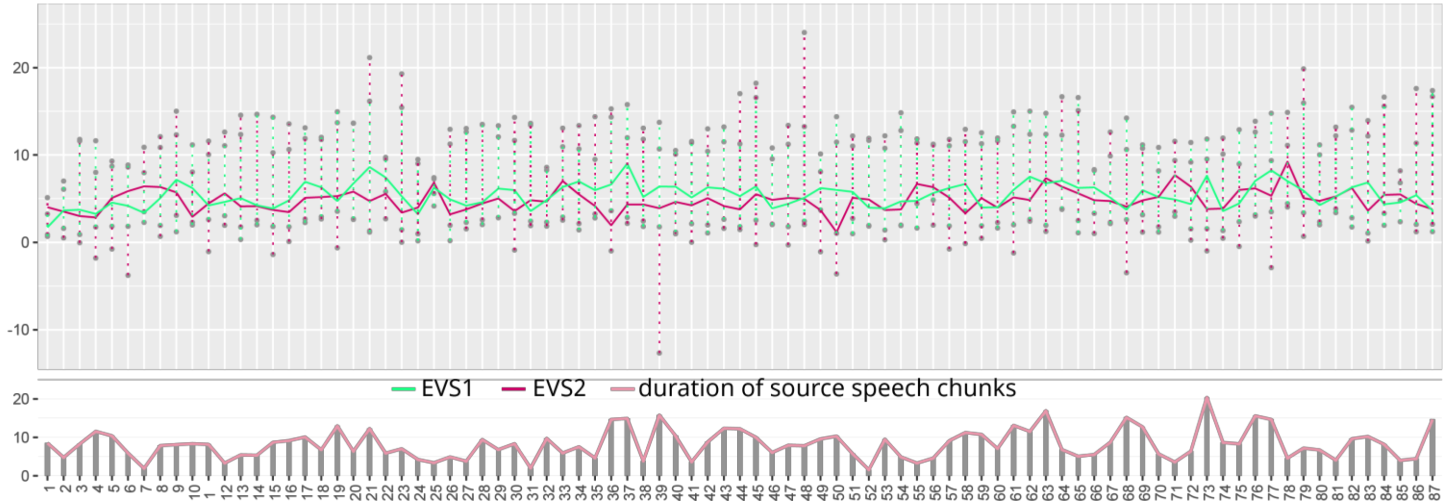


Figure 66. *EVS1*, *EVS2*, and chunks duration in Cycle II (InterpretBank Group).

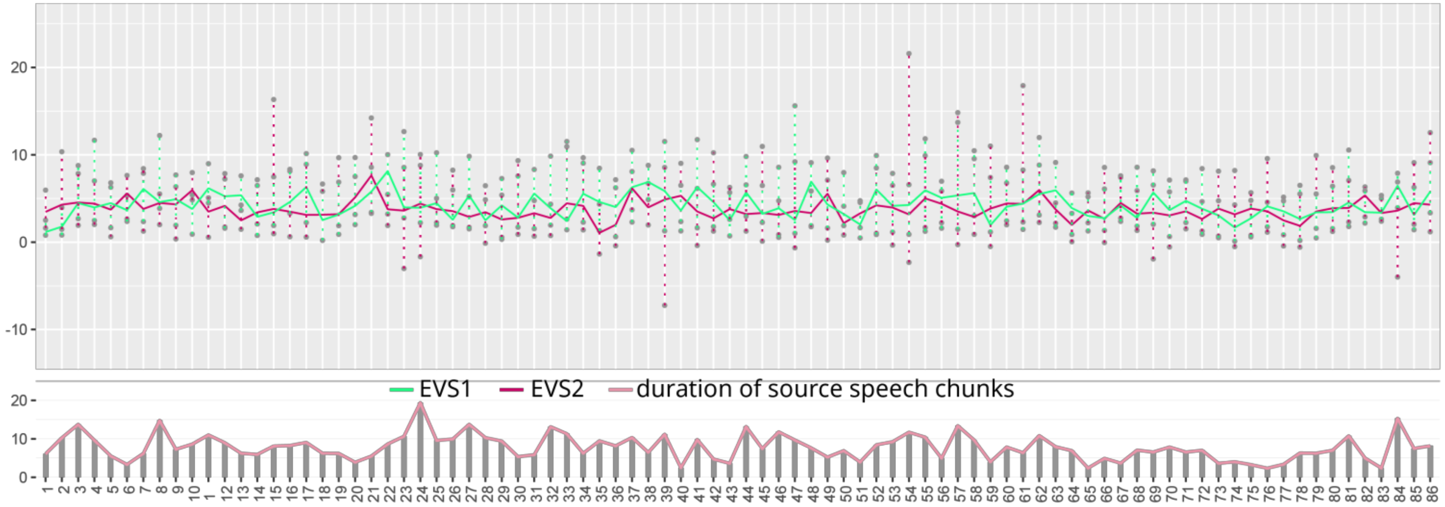


Figure 67. *EVS1*, *EVS2*, and chunks duration in Cycle III (Excel Group).

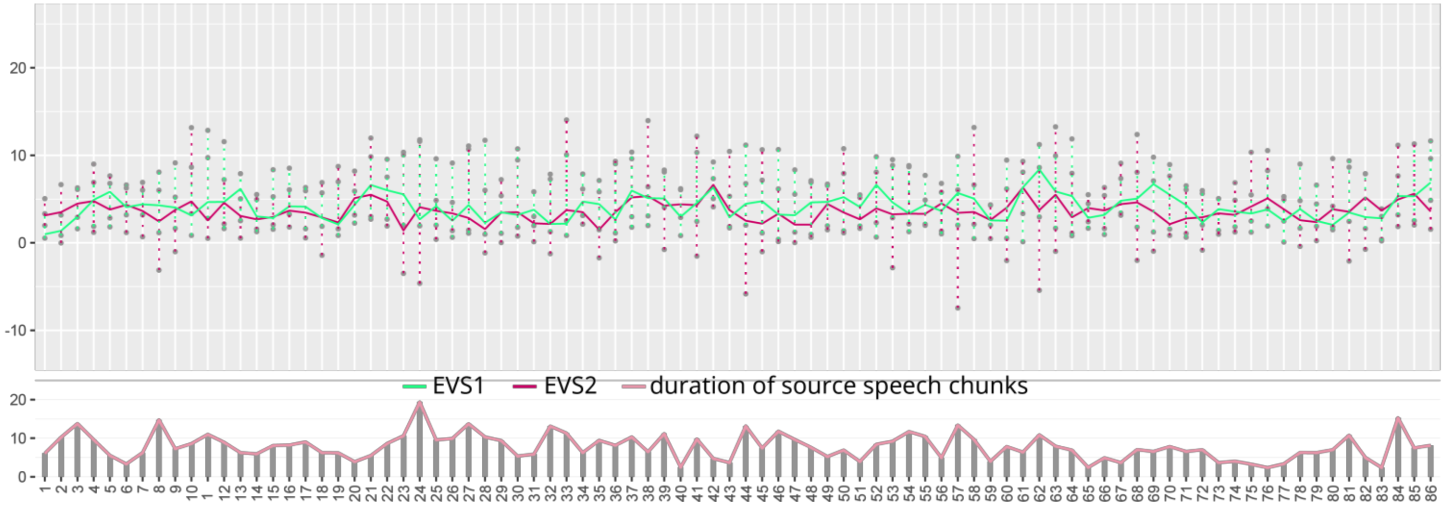


Figure 68. *EVS1*, *EVS2*, and chunks duration in Cycle III (InterpretBank Group).

Cycle	group	*EVS1* vs. source speech chunks	*EVS2* vs. source speech chunks
I	Excel	0.125 (p < 0.001)	-0.087 (p < 0.001)
	InterpretBank	0.102 (p < 0.001)	-0.053 (p = 0.035)
II	Excel	0.111 (p < 0.001)	-0.072 (p = 0.005)
	InterpretBank	0.079 (p = 0.001)	-0.076 (p = 0.002)
III	Excel	0.151 (p < 0.001)	0.005 (p = 0.845)
	InterpretBank	0.070 (p = 0.004)	-0.057 (p = 0.020)

Table 10. Correlation coefficients and p-values among EVS vs. chunks.

Table 10 shows that both groups had non-existent or minimal statistically significant negative correlation between the duration of source speech chunks and individual EVS data (i.e., *EVS1* and *EVS2*) across three Cycles. The correlations are generally weaker in Cycle II compared to Cycles I and III, for both groups. The InterpretBank group tends to have slightly lower correlation coefficients than Excel across most cycles. The strongest positive correlation is between *EVS1* and source speech chunks in Cycle III from the Excel group (0.151, p-value < 0.001). The strongest negative correlation is between *EVS2* and source speech chunks in Cycle I from the Excel group (-0.087, p-value < 0.001). The table shows correlations between the temporal indicators (*EVS1* and *EVS2*) and the duration of source speech chunks, with some variations across cycles and groups.

In summary, the results show that there was a really weak and barely statistically significant negative correlation between the duration of source speech and the *EVS1* and *EVS2* for both groups in the first two cycles. In the third cycle, the Excel group showed almost no correlation, while the InterpretBank group continued to show the same stable relationship. One possible interpretation is that longer source speech chunks are associated with shorter *EVS2*. However, the coefficient is so small that the topic is open to speculation. The absolute values of outliers were also smaller in the third cycle compared to the first two cycles.

3.3 Term accuracy analysis

As in the fluency analysis, group tendency is discussed first. Again, two informants are missing from the analysis, one from each group, because they chose to use the opposite tool in Cycle III. Also, four categories were used to study term accuracy: *correct*, *adequate*, *wrong*, and *skipped terms*. First, the overall term accuracy of all 39 targeted terms across cycles is explored in the two groups. Then we distinguish between *first-time* potential problem triggers and *repeated* terms (see **§ 2.2.2**). The Excel group demonstrated a consistent increase in mean number of *correct* renditions, from 8.3 in Cycle I to 15.3 by Cycle III (see **Table 11**). Variation also rose, from a standard deviation of 2.71 to 7.03. In contrast, the InterpretBank group started with a lower mean of 7.7 in Cycle I and surpassed the Excel group in Cycle III, with a mean value of 18.6. Interestingly, the standard deviation for this group in Cycle III was narrower than that of the Excel group. Still, the data followed an approximately normal distribution. In *adequate renditions*, the Excel group started at a mean of 1 in Cycle I, peaked at 2.1 in Cycle II, and *dropped* to 0.7 in Cycle III. The InterpretBank group started higher, at 1.9 in Cycle I, but also dipped down to 0.5 in Cycle III. This group displayed considerable variation in Cycles I and II. The Shapiro-Wilk test revealed no normal distribution of data (p-value < 0.05) for both groups in Cycle III.

										Shapiro-Wilk	
indicator	**cycle**	**group**	**Mean**	**Median**	**Mode***	**Sum**	**SD**	**Min.**	**Max.**	**W**	**p-value**
correct	**I**	**XL**	8.3	7.5	6	83	2.71	5	12	0.87	0.11
		IB	7.7	7	5	85	3.38	4	16	0.86	0.07
	II	**XL**	11.6	12	13	116	3.69	5	17	0.98	0.93
		IB	14.1	13	10	155	4.35	10	24	0.88	0.09
	III	**XL**	15.3	16.5	10	153	7.03	2	27	0.97	0.92
		IB	18.6	19	15	204	4.41	10	26	0.98	0.95
adequate	**I**	**XL**	1.0	1	1	10	0.67	0	2	0.82	0.02
		IB	1.9	1	12	21	2.12	0	7	0.84	0.03
	II	**XL**	2.1	2	2	21	0.99	0	3	0.83	0.03
		IB	1.6	1	1	18	1.91	0	7	0.67	0.00
	III	**XL**	0.7	1	1	7	0.68	0	2	0.80	0.02
		IB	0.5	0	0	5	0.93	0	3	0.57	0.00
wrong	**I**	**XL**	2.5	3	3	25	1.51	0	5	0.95	0.64
		IB	2.7	2	1	30	2.72	0	10	0.77	0.00
	II	**XL**	2.8	2.5	1	28	1.99	0	6	0.94	0.53
		IB	3.6	4	4	40	1.63	0	6	0.92	0.35
	III	**XL**	3.9	3.5	3	39	3.21	0	11	0.87	0.10
		IB	3.0	2	2	33	2.61	0	8	0.90	0.17
skipped	**I**	**XL**	27.2	27.5	25	272	3.49	22	33	0.97	0.92
		IB	26.6	30	30	293	5.94	15	33	0.85	0.05
	II	**XL**	22.5	22	20	225	3.6	18	29	0.94	0.55
		IB	19.6	20	17	216	3.64	12	24	0.91	0.25
	III	**XL**	19.1	17.5	17	191	5.47	11	29	0.96	0.82
		IB	17.0	17	17	187	4.54	9	27	0.93	0.42

* More than one mode exists, only the first is reported.

Table 11. Accuracy indicators for two groups (Shapiro-Wilk test included).

In *wrong renditions*, both groups had a relatively stable mean—around 2.5 to 3.9 for the Excel group and 2.7 to 3 for the InterpretBank group. However, the SD for the Excel group in Cycle III was notably higher. The Shapiro-Wilk tests indicated that the data for both groups across all cycles was approximately normally distributed, except for the InterpretBank group in Cycle I (p-value < 0.05). Regarding *skipped terms*, both groups had started at a high mean value in Cycle I, with 27.2 for the Excel group and 26.636 for the InterpretBank group. However, both groups saw their mean numbers of *skipped terms* decline by Cycle III, with the Excel group reaching 19.1 and the InterpretBank group reaching 17. The SD for both groups remained fairly consistent across cycles, and the Shapiro-Wilk test suggested that the data was approximately normally distributed for both groups.

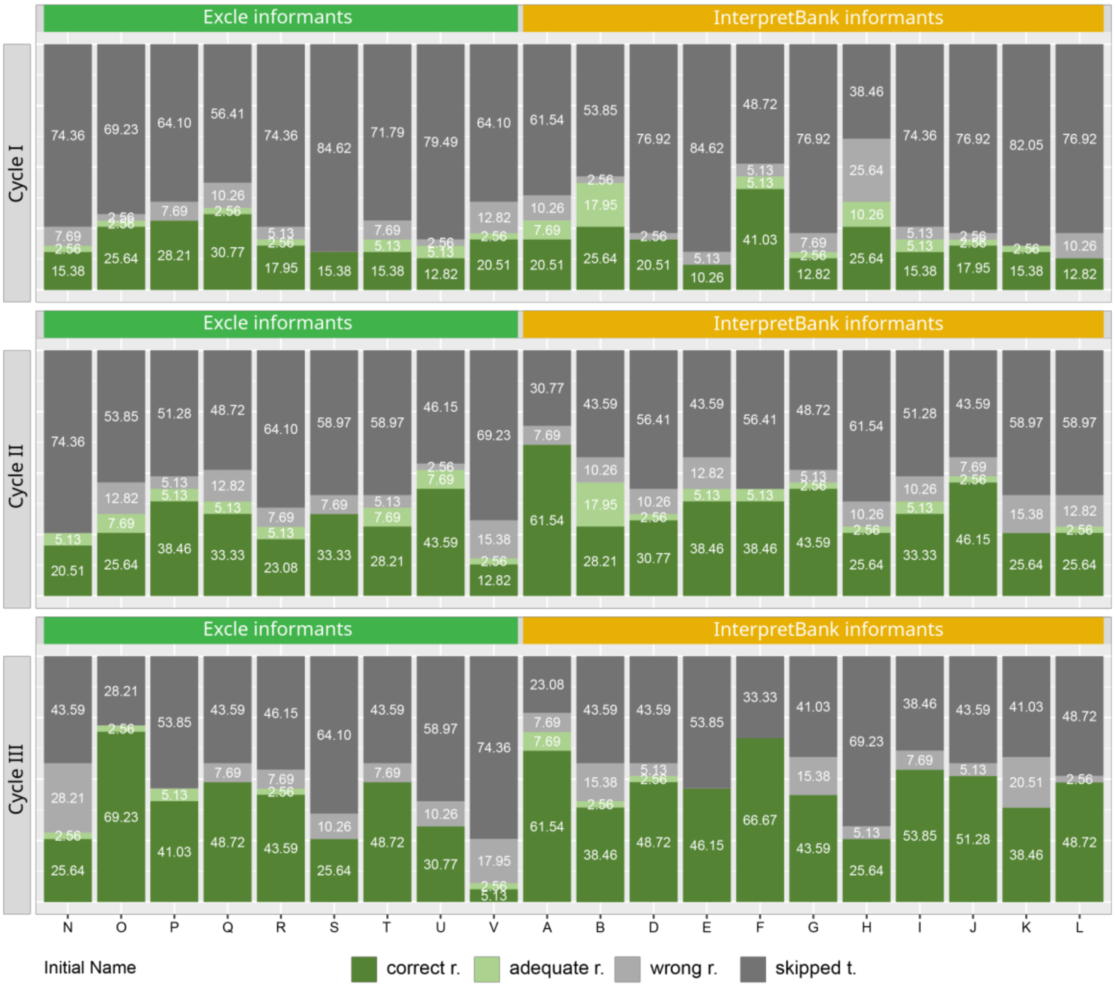


Figure 69. Percentage of accuracy indicators across cycles by two groups.

So, the study assessed term accuracy comparing two groups over three cycles. In terms of correct renditions, the Excel group showed steady improvement, while the InterpretBank group began lower but eventually surpassed Excel. The Excel group's adequate renditions peaked in Cycle II and then decreased, whereas the InterpretBank group's adequate renditions declined consistently. Both groups maintained stable means in wrong renditions, but the Excel group's standard deviation increased by Cycle III. Initially, high skipped terms in both groups decreased by Cycle III. The data was approximately normally distributed for both groups in most categories and cycles, but not for all categories. Besides, the Excel group and InterpretBank group also exhibited some unique features between categories. In terms of quantity, for both groups, the number of *skipped terms* far exceeded other categories, followed by *correct renditions*. The Excel group had a higher number of *skipped terms* than InterpretBank, while it had a lower number of *correct renditions* than InterpretBank. Both groups showed a smaller number of adequate and wrong renditions. We will come back to this point in the discussion.

Starting with the first plot in **Figure 69**, which focuses on Cycle I, it is evident that the range of accuracy varies significantly among the informants. Noel (N), from the Excel group, had a 74.36% rate of *skipped terms* with targeted potential problem triggers; 15.38% of her renditions were *correct,* and 7.69% were *wrong*. The *adequate* category only reached a 2.56% rate. In Cycle II, there is a general shift in accuracy. For instance, Gale (G) improves in her correct renditions compared to Cycle I. One of the possible interpretations is that the interventions or treatments applied between the cycles might have been effective for this individual. In Cycle III, some informants, such as Blake (B), Peyton (P), and Quinn (Q) kept a consistent level of accuracy across all cycles, while others fluctuated in their accuracy.

In summary, the proportion of *correct* renditions in both the Excel and InterpretBank groups, for most participants, showed varying degrees of increase from Cycle I to Cycle III. However, there were two notable exceptions: the proportion of *correct renditions* in the Excel group exhibited a decreasing trend over the cycles for Val (V), and Harley (H) in the InterpretBank group remained constant. Compared to Cycle I as the baseline, both Cycle II and Cycle III showed a decrease in the percentage of *skipped terms*, with most informants in the two groups experiencing a reduction, except for Val (V). Since adequate renditions are considered acceptable, their percentage decreased from Cycle II to Cycle III, particularly in the InterpretBank group, where only Alex (A), Blake (B), and Dana (D) had adequate renditions. In comparison, the Excel group had five participants with adequate renditions (Noel (N), Oakley (O), Peyton (P), Riley (R), and Val (V))

3.3.1 Correct renditions

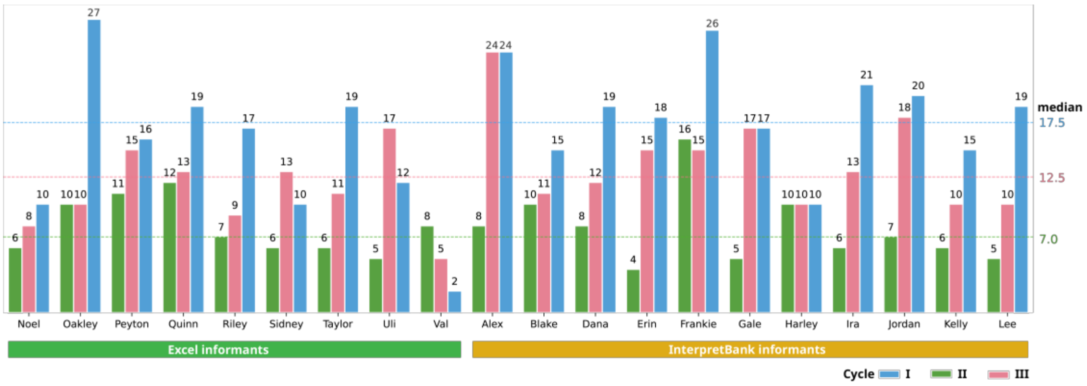


Figure 70. Distribution of *correct* renditions across cycles by informants.

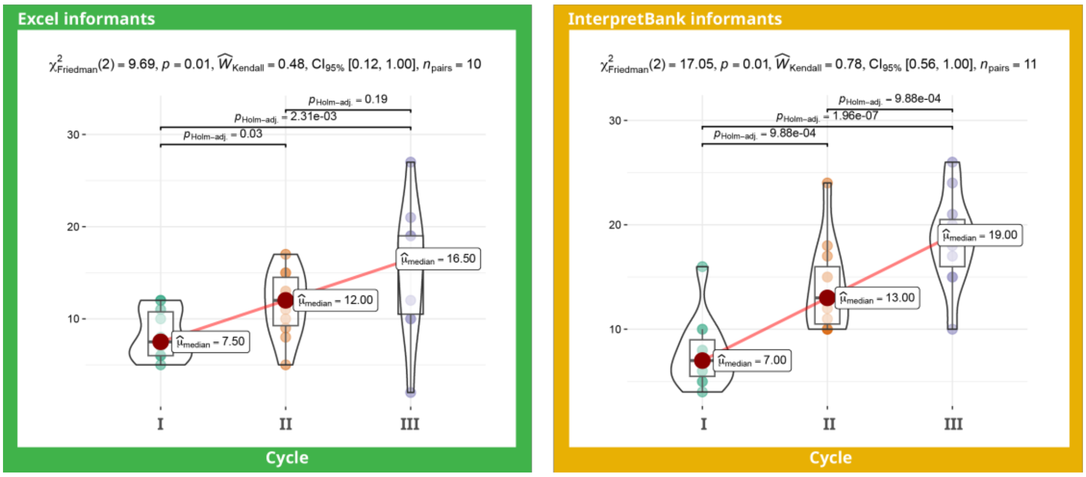


Figure 71. Within-group comparison of *correct* renditions across cycles.

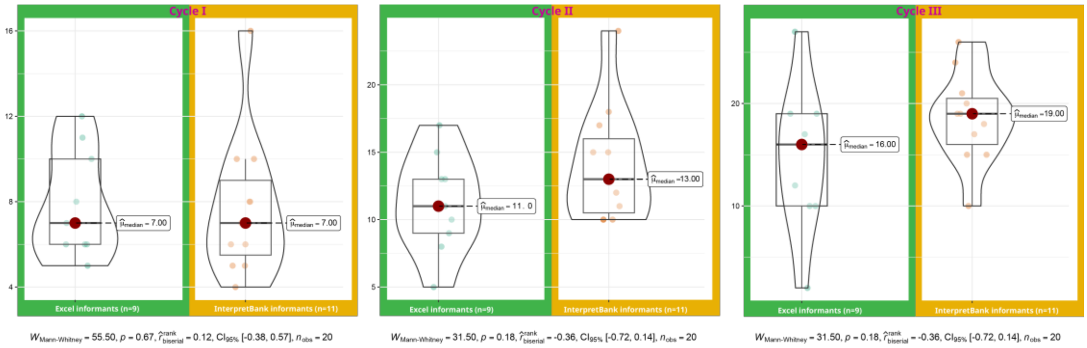


Figure 72. Comparison of *correct* renditions across cycles between groups.

In *correct renditions* (**Figure 70**) some informants show noticeable trends in both groups from Cycle I to Cycle III. For instance, in the Excel group, Noel, Peyton, Quinn, and Taylor exhibit increasing trends, while more informants from the InterpretBank group, such as Blake, Dana, Erin, Gale, Ira, Jordan, Kelly, and Lee, also show upward trends. However, Uli shows a downward trend from Cycle I to Cycle III, being the only informant in both groups to exhibit a decreasing trend. Additionally, some informants in both groups show fluctuations across the cycles, such as Sidney in the Excel group and Frankie in the InterpretBank group.

The Mann-Whitney U test shows p-values of 0.67 for Cycle 1, 0.18 for Cycle 2, and 0.18 for Cycle 3, all above 0.05, indicating no significant differences between the groups. Copy. On the other hand, the Friedman test within each group indicated significant changes in *correct* renditions across cycles, with p-value = 0.01 for both the Excel group and the InterpretBank group. That is, the choice of tool does not seem to significantly affect the number of *correct renditions*, but there are significant changes within each group across cycles.

3.3.2 Adequate renditions

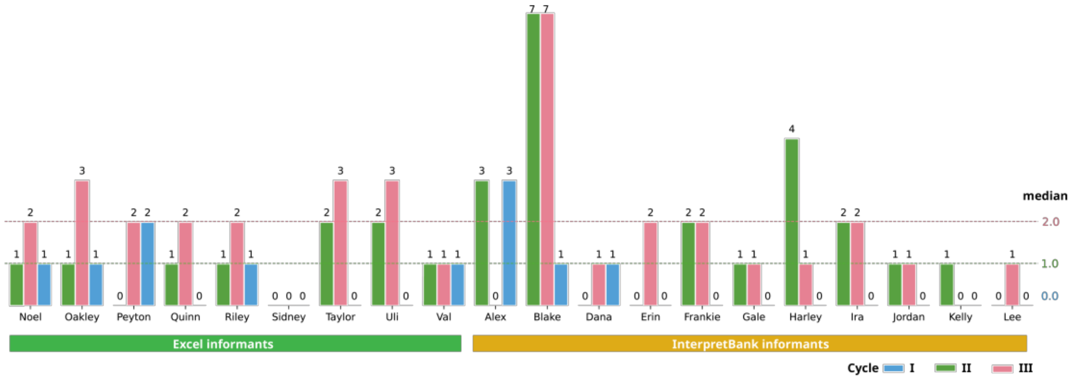


Figure 73. Distribution of *adequate* renditions across cycles by informants.

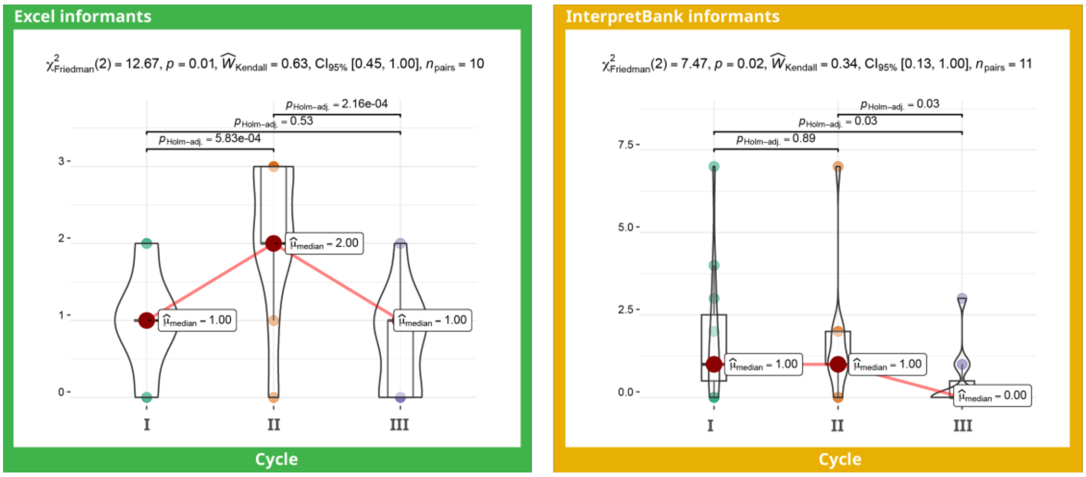


Figure 74. Within-group comparison of *adequate* renditions across cycles.

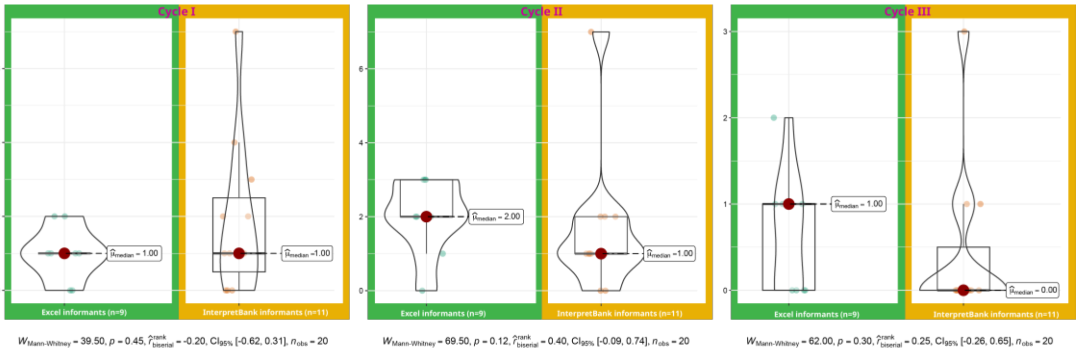


Figure 75. Comparison of *adequate* renditions across cycles between groups.

Although both the InterpretBank and Excel groups had informants with relatively few *adequate* renditions (see **Figure 73**), the only informant with a relatively higher number of *adequate renditions* was Blake in the InterpretBank group. The informants in the Excel group had *adequate* renditions ranging from 1 to 3 instances, with Sidney being an exception, and having no *adequate* renditions in all three cycles. In comparison, the InterpretBank group had a range of 1 to 2 instances, with more informants having only one *adequate* rendition. The median for Cycle I lay near the #1, suggesting a lower frequency of *adequate* rendition. The Friedman test (see **Figure 74**) within each group found significant changes across cycles, with p-value = 0.01 for the Excel group and p-value = 0.024 for the InterpretBank group. This again suggests that the cycle, rather than the choice of tool, played a more critical role in the number of *adequate* renditions. However, the Mann-Whitney U test (p-values: 0.45 in Cycle I, 0.12 in Cycle II, 0.30 in Cycle III) revealed no statistically significant differences between the two groups across the cycles, as shown in **Figure 75**.

3.3.3 Wrong renditions

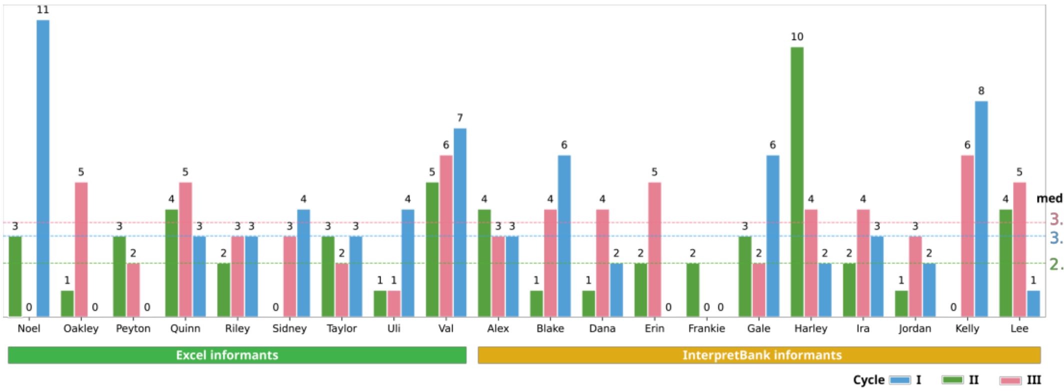


Figure 76. Distribution of *wrong* renditions across cycles by informants.

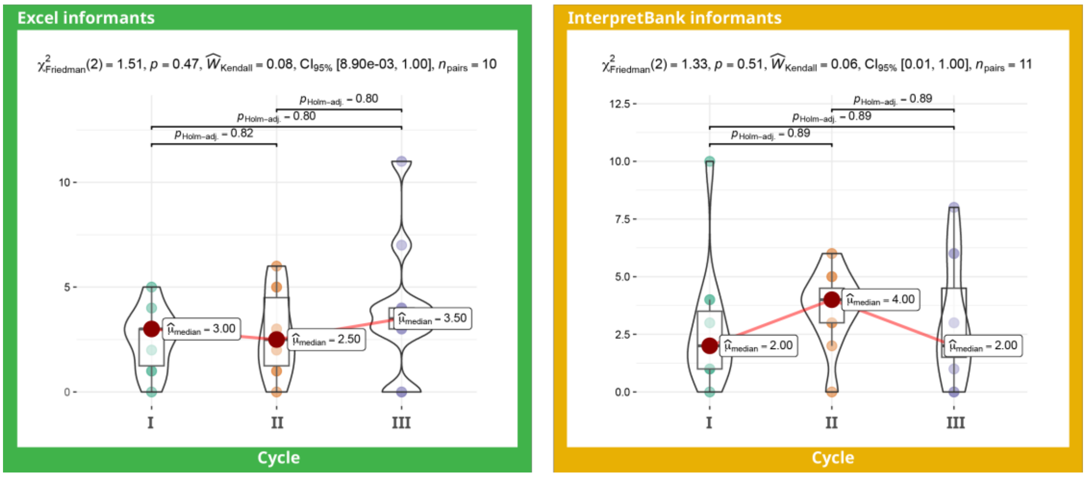


Figure 77. Within-group comparison of *wrong* renditions across cycles.

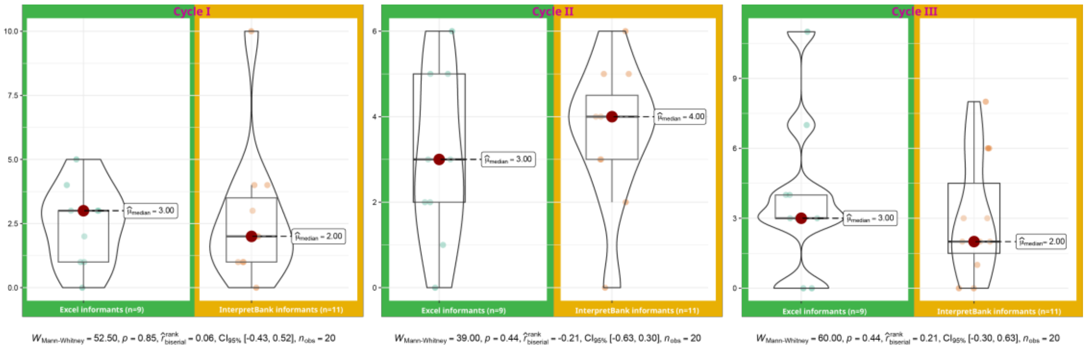


Figure 78. Comparison of *wrong* renditions across cycles between groups.

Wrong renditions are shown in **Figure 76.** Alex, from the InterpretBank group, had 4 wrong instances in Cycle I. Noel, from the Excel group, recorded 3 wrong instances in Cycle I, not far behind Alex. The median values for each cycle were also fairly low, hinting at a general tendency to avoid them. The Mann-Whitney U test in **Figure 78** concurred, indicating no significant differences between the groups with a p-value above conventional alpha 0.05: 0.85 in Cycle I, 0.44 in Cycle II, and 0.44 in Cycle III. This lack of difference was also apparent in the Friedman test within each group, showing no significant changes across cycles with p-values of 0.47 and 0.51 for the Excel and the InterpretBank groups respectively.

3.3.4 Skipped terms

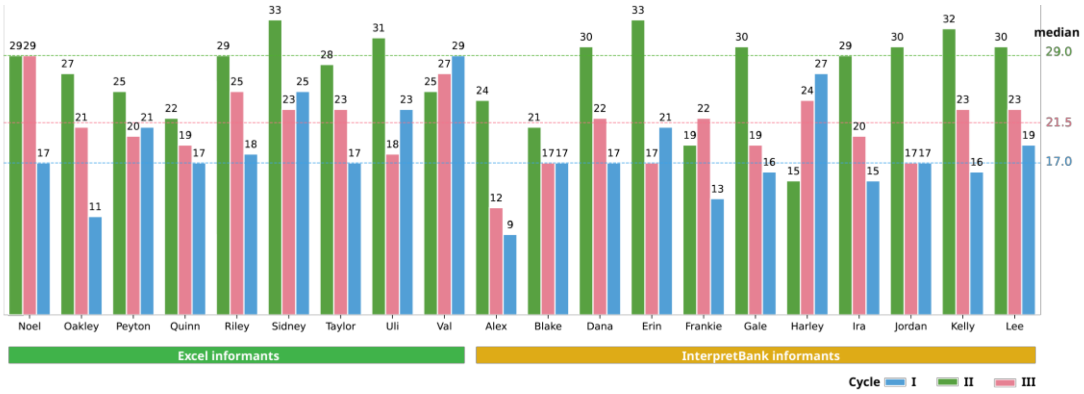


Figure 79. Distribution of *skipped terms* across cycles by informants.

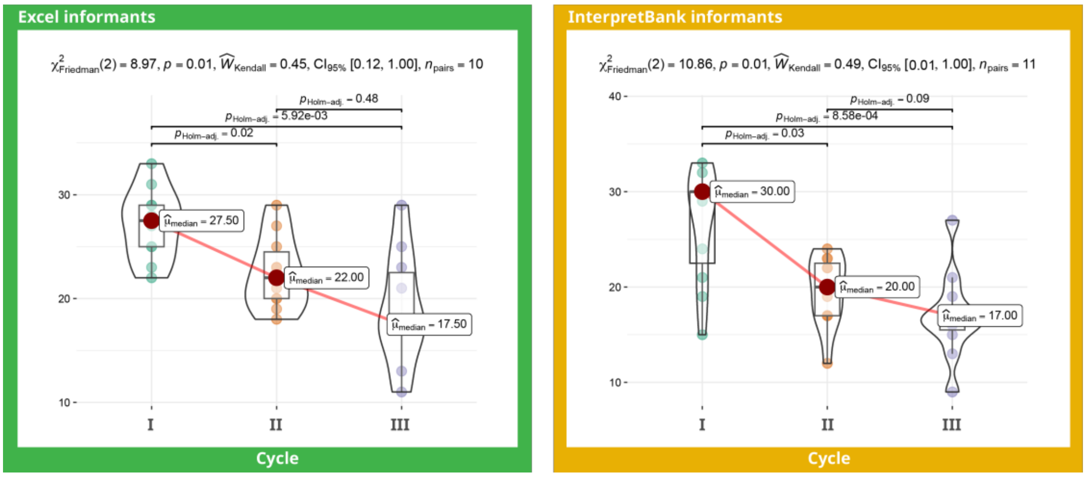


Figure 80. Within-group comparison of *skipped terms* across cycles.

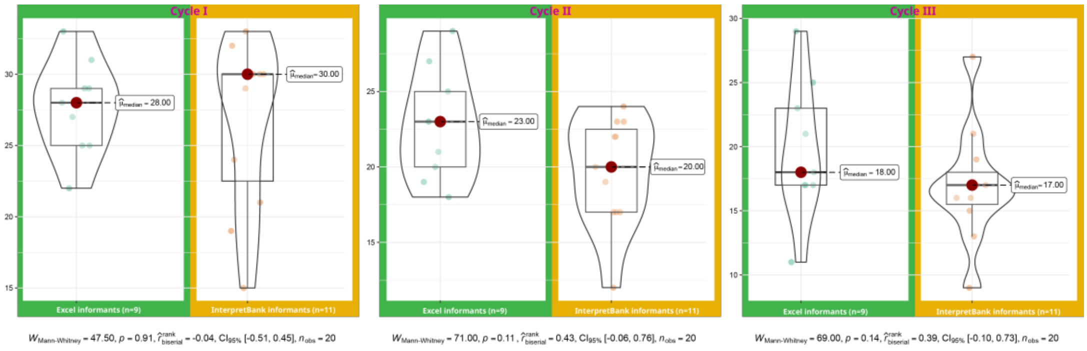


Figure 81. Comparison of *skipped terms* across cycles between groups.

From **Figure 79**, the median values show that although each cycle has a relatively high number of *skipped terms*, there is a decreasing trend from Cycle I to Cycle III. This observation is also supported by most informants' situations. The number of *skipped terms* in Cycle I is often the highest among the three cycles. However, there are exceptions, such as Val in the Excel group, which shows an increasing trend across the cycles. In the InterpretBank group, Harley shows an increasing trend across the cycles, while Frankie exhibits a fluctuating pattern with an increase followed by a decrease across the cycles. The Mann-Whitney U test in **Figure 81** supported the finding of no statistically significant difference between the two groups, with p-values of 0.91 in Cycle I, 0.11 in Cycle II, and 0.14 in Cycle III. Interestingly, the within-group Friedman tests indicated significant changes within each group across cycles, with a p-value of 0.01 for both the Excel group and the InterpretBank group.

In summary, while InterpretBank informants demonstrated more correct renditions and fewer skipped terms from Cycle II to Cycle III compared to those using Excel, there were no significant differences between the groups in terms of correct, adequate, and wrong terms, nor in terms of skipped terms. Some informants showed increasing trends in *correct renditions*, while others fluctuated or decreased across cycles. Adequate renditions were generally low in both groups, with minimal differences between them. Similarly, the occurrence of wrong terms and skipped terms did not significantly differ between the groups, but variations were observed within each group over time.

3.3.5 Repeated potential problem triggers

As a reminder, the category of *first-time terms* refers to potential problem triggers appearing in the source speech for the first time. Three out of those first-time terms show up twice later on in the discourse. *Rep1* refers to the first repetition and *rep2,* to the second repetition of those three selected terms. The first instance of *repeated* terms was not labeled *first-time terms* but rather *1st2Rep,* to remind the reader that they are first appearances of terms that would later appear again.

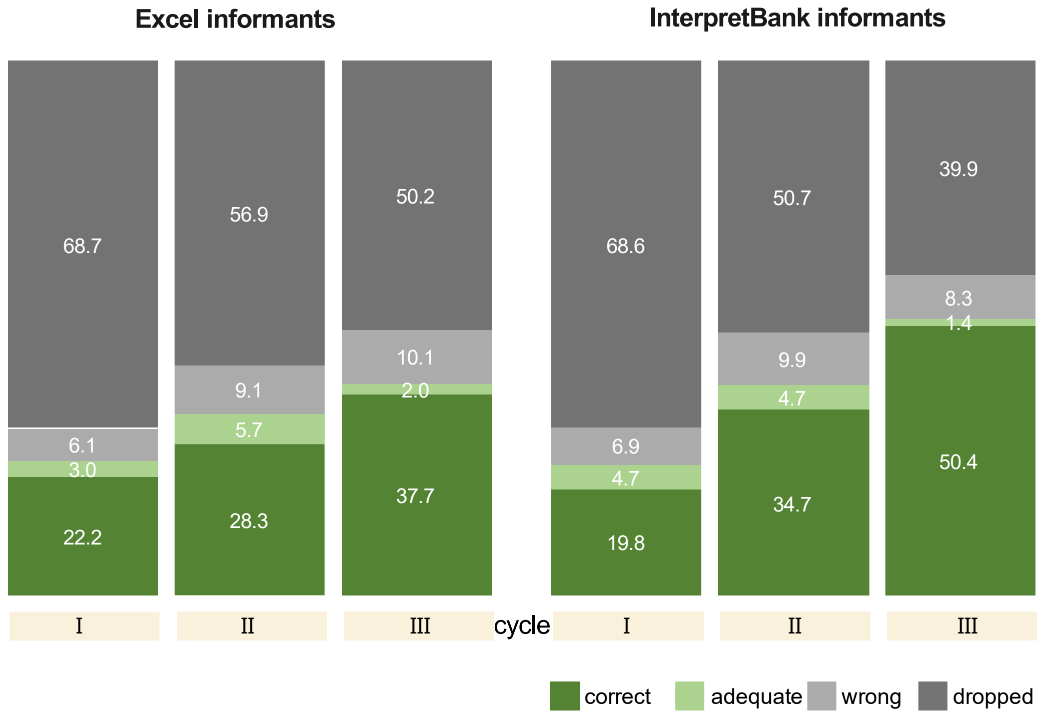


Figure 82. Stacked percentage bar plot of first-time terms by cycles and groups.

Figure 82 focuses on the accuracy of *first-time terms*. The InterpretBank group displayed a significant performance improvement over the cycles. In Cycle I, only about 19.8% of the terms were correctly interpreted, a relatively low baseline for the group. In Cycle II, the percentage of correctly rendered terms increased to 34.7%, an almost 15% increase. There was also a decrease in the number of *skipped terms*, hinting at an overall betterment of their performance. The InterpretBank group achieved an impressive 50.4% of the terms interpreted accurately in Cycle III, confirming the upward trend. The Excel group showed a more modest yet steady improvement as well. In Cycle I, around 22.2% of the terms were *correctly* rendered. This figure slightly increased to 28.3% in Cycle II and 37.7% in Cycle III, a more gradual yet also steady improvement. These findings suggest that

InterpretBank could be an effective support in helping interpreters achieve better accuracy with unfamiliar terms during interpretation.

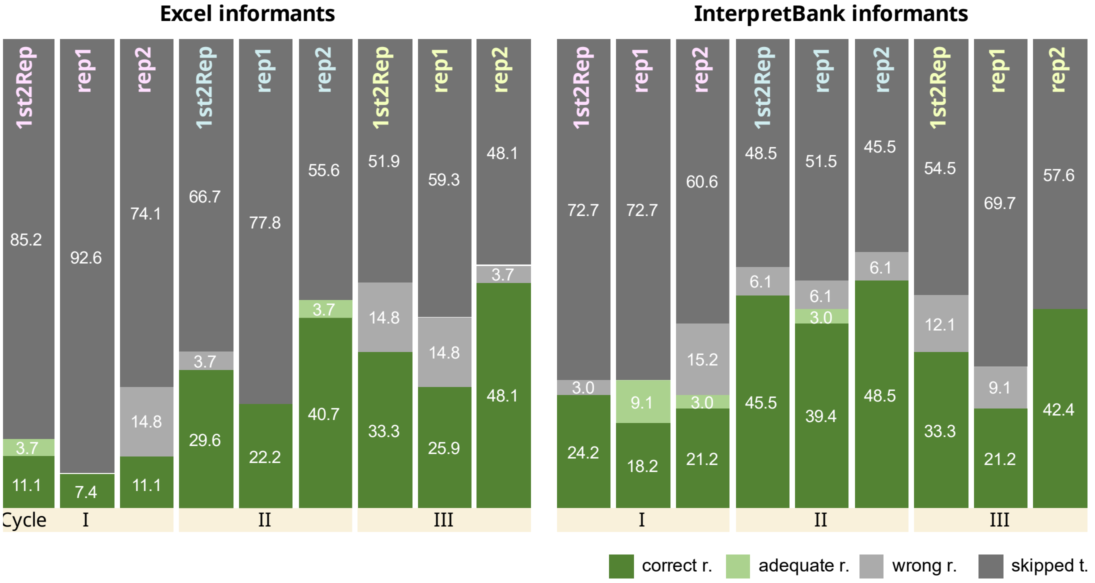


Figure 83. Stacked percentage bar plot of *repeated* terms by cycles and groups.

In **Figure 83** the Excel group experienced a progressive uptick in the correct renditions within the *1st2Rep* category. Starting at a modest 11.11% in Cycle I, the group improved to 29.63% in Cycle II and reached 33.33% in Cycle III. In the *rep1* term type, the group experienced a steep ascent from a low 7.40% in Cycle I to 22.22% in Cycle II, before leveling off at 25.93% in Cycle III. This suggested a learning curve that gained momentum but then started to plateau. For *rep2,* the group made a dramatic improvement from Cycle I's 11.11% to Cycle II's 40.74%, finally achieving a peak of 48.15% correct interpretations in Cycle III. This indicated sustained progress across the cycles.

On the other hand, the InterpretBank group's percentage of correct interpretations in the *1st2Rep* category displays important fluctuations. The group began at 24.24% in Cycle I, peaked at 45.45% in Cycle II, and then moderately adjusted to 33.33% in Cycle III. The *rep1* category experienced a peak in Cycle II with 39.39% correct interpretations, a significant leap from Cycle I's 18.18%. However, the group then experienced a decline, settling at 21.21% in Cycle III, indicating room for further improvement or refinement. In the *rep2* category, the InterpretBank group showed a positive trend, starting at 21.21% in Cycle I, peaking at 48.48% in Cycle II, and maintaining a strong performance of 42.42% in Cycle III. This suggested effective learning or adaptation across the cycles. In brief, the Excel group generally showed an increase in correct renditions across all term types. The InterpretBank group, while also showing improvements, had a more varied performance, with some categories peaking in Cycle II, with first use of InterpretBank, and then lowered in Cycle III.

Cross-comparing **Figure 82** and **Figure 83**, both groups improved in *repeated* terms (i.e., *rep1* and *rep2*) of terms across the cycles. The InterpretBank group displayed robust performance in the *rep1* term category in Cycle II but experienced a decline in Cycle III. The Excel group, in contrast, demonstrated a steady enhancement in *rep1* and *rep2* throughout the cycles. In *rep2,* both groups reached their peak in Cycle III, with nearly identical percentages of accurate output. For *1st2Rep terms,* both groups improved throughout the cycles, but by Cycle III, the InterpretBank group demonstrated a greater percentage of accurate interpretations.

The previous focus of exploration is on how informants perform with *first-time terms* and *repeated terms*. Let us devote particular attention to the *1st2Rep* category. to investigate whether *1st2Rep* terms are rendered well in *rep1* and *rep2*. **Figure 83** illustrates that, generally, both the Excel and InterpretBank groups display variation in the correct rendition of repeated terms across the three categories, with *rep1*'s accuracy consistently trailing behind *1st2Rep* and *rep2* over time. Specifically, for the Excel group, the overall correct rate for repeated terms shows an upward trend. In Cycle I, the correct rate for all three categories is relatively low, but in Cycles II and III, the correct rate for repeated terms significantly increases, although *rep1*'s correct rate remains lower than *1st2Rep* and *rep2*. This suggests that *1st2Rep* terms are effectively rendered in *rep2* in cycles II and III, as the accuracy for *rep2* surpasses that of *1st2Rep*. In contrast, the InterpretBank group displayed fluctuations in rendering *1st2Rep* terms during *rep1* and *rep2*. Specifically, the performance for *1st2Rep* terms slightly declined in *rep1*, while it improved in *rep2* across cycles. When examining both groups, the Excel group shows a persistent uptick in the accuracy of *1st2Rep* from *rep1* to *rep2*, with marked improvements in cycles II and III compared to Cycle I. The InterpretBank group experiences an increase in the correct rates of all three categories in Cycle II but a decline in Cycle III. *1st2Rep* in *rep1* and *rep2* performs better in Cycle II compared to Cycle III. Moreover, in both Cycle I and Cycle II, the InterpretBank group's performance in the three categories of *1st2Rep*, *rep1*, and *rep2* consistently exceeded that of the Excel group.

3.4 Search behavior of the InterpretBank group in Cycles II and III

We recognize the importance of analyzing search behaviors in the Excel group. However, due to the study's primary focus on InterpretBank informants' search behavior, we have to limit our exploration of Excel group search patterns. Therefore, this section presents the analysis of search workflows using InterpretBank, spanning from Cycle II to Cycle III. The focus is on examining search events to determine the number of correct outputs contributed by InterpretBank. This analysis aims to check the effectiveness of InterpretBank in aiding the search process during these cycles.

3.4.1 Search workflows with InterpretBank

Figure 84 presents an Alluvial Diagram illustrating the workflow of the InterpretBank search during Cycle II. For instance, the first axis showed that there were 242 searches in total (recorded and captured by Pynput and TechSmith Capture) by 11 InterpretBank informants (from Jordan to Harley, ordered from most to least), including 87 searches for *non-target terms* (terms in the master glossary but not first-time terms, see details below), and 84 searches for *first-time terms*. Of the 87 searches for *non-target terms*, 71 detected no *typing correction*. Of the 84 searches for first-time terms, 74 detected no *typing correction*. Together with other searches detected in the second axis, there were 207 searches with no *typing correction*. Of these 207 searches, 130 had correct *character strings* when typing. Correct *character strings* also included wrong *typing corrections*, totaling 153 searches which led to *expected search results*. Of these 153 searches, 134 had *correct renditions*.

This diagram, grounded in our observations, methodically represents each step required to complete a search in InterpretBank, initiated upon encountering a term and deciding to retrieve a term. In this diagram, elements (informants and search) are assigned to parallel vertical axes. From left to right, they represent the workflow's chronological order. Each vertical axis stands for a step during the search. Each vertical axis consists of several blocks, each block representing a set of searches. Different axes are connected by color streams connected between blocks, indicating the sequence relationship. Values are represented with blocks on each axis, standing for the number of searches. The blocks on each axis are organized in descending order based on their values, signifying the frequency of the action instances. The height of a block represents the size of the cluster, and the height of a stream field represents the number of searches contained in both blocks connected by the stream field. White dots with number insides in the streams refer to the number of the left-connected block's represented number of searches. Successful research can provide the whole picture of the workflow of the search with InterpretBank. However, to make the research object more specific, we did not count searches for terms unrelated to the master glossary, as they led to unsuccessful or incomplete searches in this study. Instead, we marked them with light grey color streams. These searches are also meaningful but require more space to study in future research.

In the first axis, each informant's name is listed from top to bottom in descending order of the number of searches they conducted with InterpretBank. For example, Jordan conducted 47 searches. These counts do not indicate whether the searches were successful or not. The second axis represents the type of term that was searched for. We identified and classified these terms informants heard (*input*) based on keylogging data and screen recordings. There are five main types of possible target terms based on whether the term is in the master glossary or not: *first-time terms, rep1, rep2*, *non-target terms*, *terms not in the master glossary*, and *unrecognized terms*. *First-time terms*, *rep1,* and *rep2* terms were defined in **§ 3.3.5**.

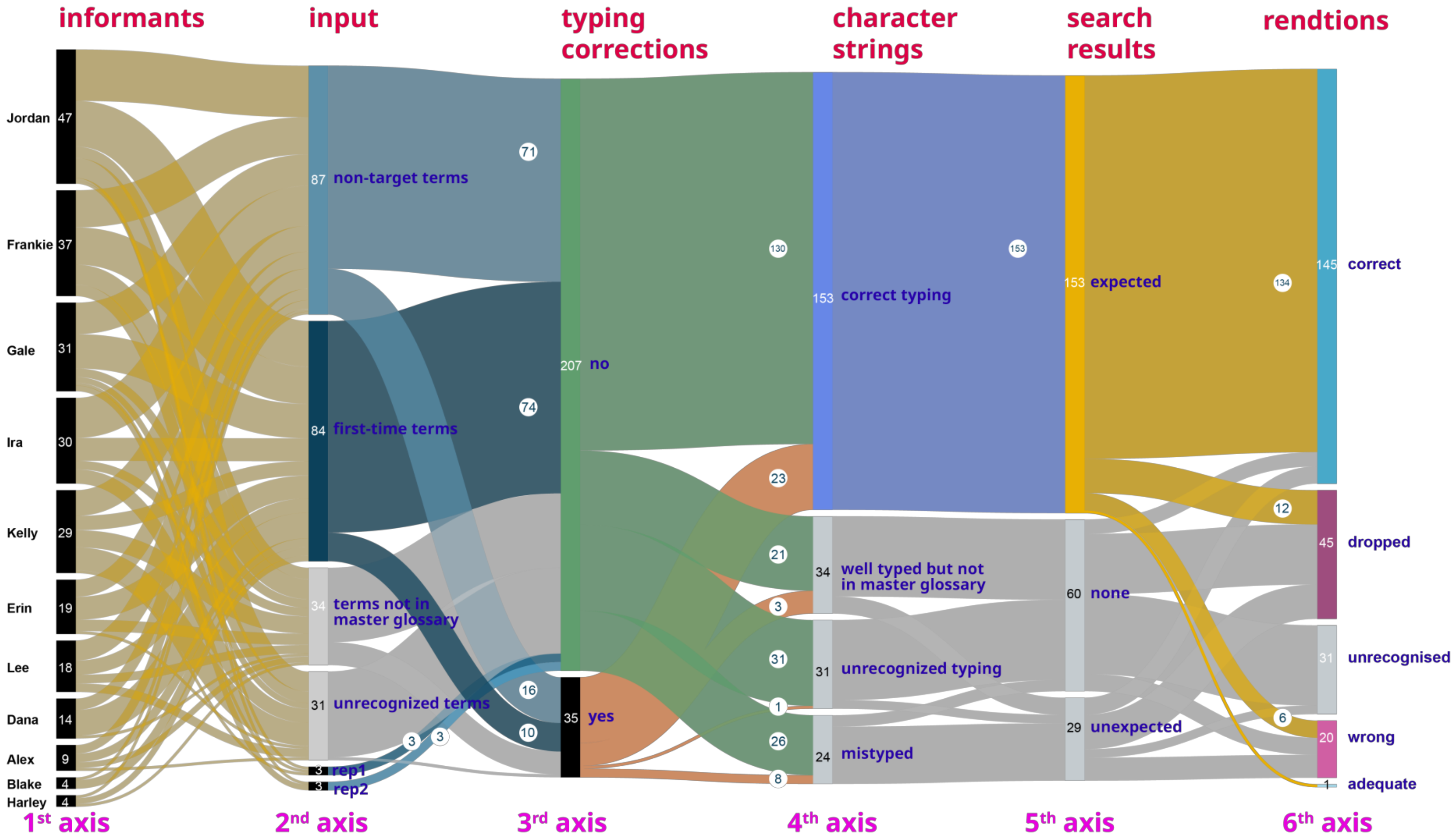


Figure 84. Workflow of search with InterpretBank in Cycle II.

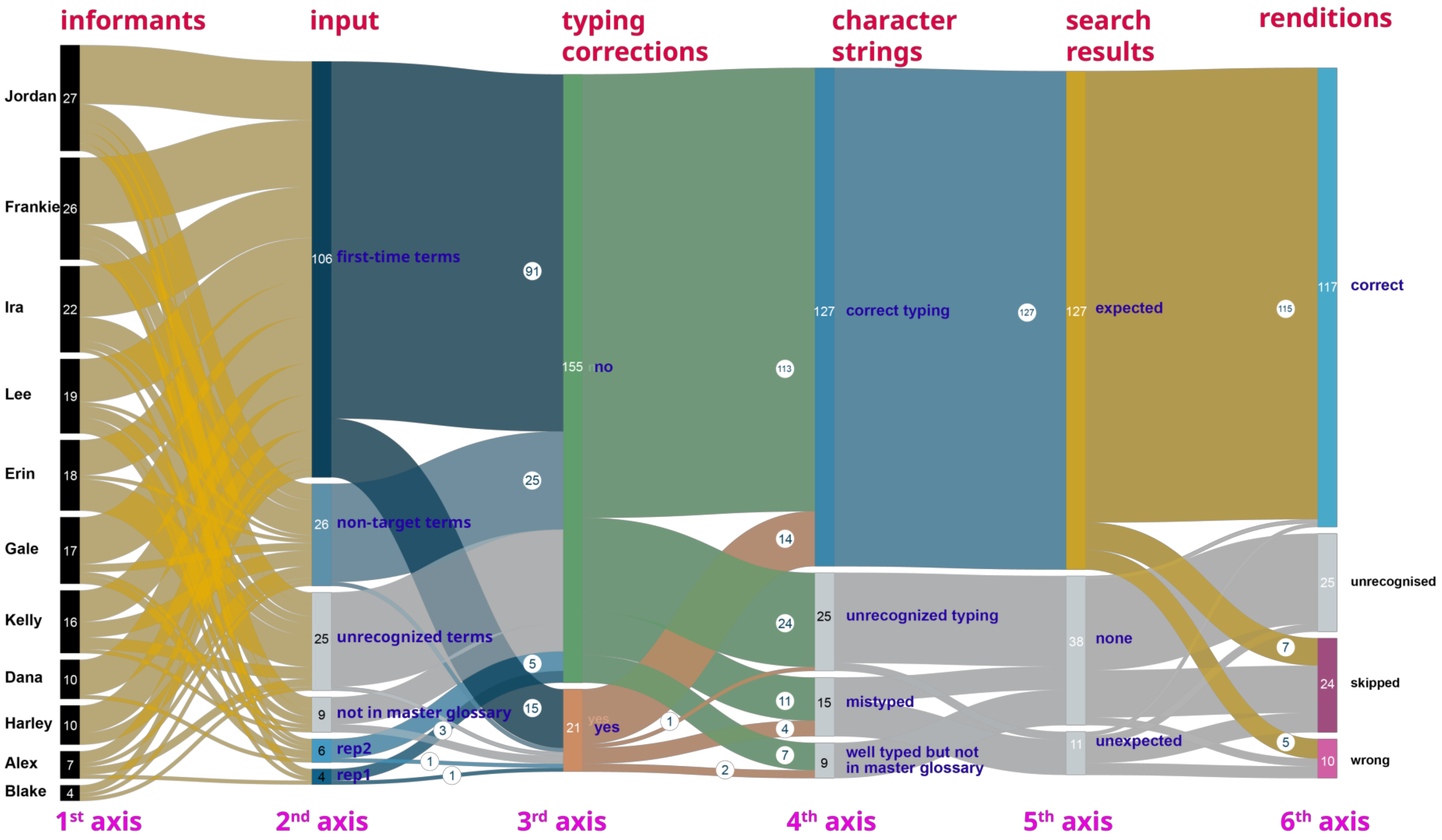


Figure 85. Workflow of search with InterpretBank in Cycle III.

Non-target terms encompass all terms in the master glossary except for *first-time terms*, *rep1*, and *rep2*. These latter three categories, being the primary focus of this study, are regarded as target terms warranting separated analysis, while the remaining terms in the master glossary are noted as *non-target terms*. Although *non-target terms* are of importance, their significance within the scope of this research is lower, irrespective of their frequency of occurrence. *Unrecognized terms* are target terms that could not be identified by us. *Terms not in the master glossary* are terms that the informants searched for that are not in the master glossary at all. Based on the second axis, 87 searches were conducted by 11 informants for *non-target terms*, followed by 84 searches for first-time terms. These two categories account for most searches.

The third axis represents the first stage of the search, specifically focusing on typing behavior. During this stage, informants may experience typos, delete typed letters using the *backspace* key, and then continue typing with new letters. This process is referred to as *typing correction*. 207 searches showed no typing correction in the 3rd axis, while 35 searches involved typing correction. We can link 207 searches without typing correction to the 2nd axis. Among the 207 searches without typing correction, 71 were from *non-target terms*, and 74 were from *first-time terms*. The remaining searches were related to terms not in the master glossary and unrecognized terms. Since these searches were not for terms unrelated to the master glossary, they were not counted in the flow. From the 3rd axis, we can observe that most searches do not involve any typing correction. The 35 searches with typing correction mainly come from 16 searches belonging to *terms not in the master glossary* and 10 from *first-time terms*. Therefore, the third axis provides insight into the most common searches and their relation to typing correction.

The fourth axis represents the *character string* that survived in the search bar of InterpretBank. No results for the wrong *character string* survived in the search bar. We were of course interested in learning whether regular or else fuzzy searches were performed. However, in the overall context of this project, this was a secondary issue and had to be sacrificed and wait for future research. In this context, the 4th axis contains four categories: correct typing, which indicates that the terms are the correct *final character string* and have been recognized by InterpretBank; *well typed but not in the master glossary* which means that we can recognize the terms typed by the informants, but these terms are not from the master glossary and are unrelated to the current study; *unrecognized typing* where we cannot recognize the terms typed by the informants, such as a single letter *a* or two letters *hy*; and *mistyped*, where we can recognize the term, but it has been typed incorrect character string. Except for the first category, the other three categories result in incomplete searches that are not related to this study and are marked with light grey. As a result, 153 searches are correctly inputted, and they are searching for terms in the master glossary.

The fifth axis represents the search results, where 153 searches with *correct input* for terms in the master glossary led to the *expected translation* provided by InterpretBank. However, there are two other categories: *none* and *unexpected*. *None* refers to no results from InterpretBank possibilities because the target terms in the 4th axis were not in the master glossary, possibly due to incomplete typing, or incorrect typing. The *unexpected* category, on the other hand, arises when the informants' inputs lead to unpredicted results from InterpretBank. For example, an informant might hear the term *chronic stress* and intend to search for it. However, if the informant types *cor* instead of the correct initial letters *chro*, this could mistakenly map to a term like *cortisol* that is present in the master glossary. As a result, InterpretBank returns results for *cortisol* instead of the intended *chronic stress*, leading to unexpected outcomes in the search process.

In the last axis or 6th axis, the *rendering quality* is shown. Along with the term accuracy categories, the *output* is assessed in five categories: *correct*, *skipped*, *wrong*, *adequate*, and additional category with *unrecognized*. *Unrecognized* refers to our inability to determine which words in the output are related to the search based on the *output* transcript. As a result, 145 renditions are correct, with 134 *correct renditions* from 153 expected results. There are 45 *skipped terms* from 12 expected results, meaning *expected* results from InterpretBank but informants did not use them in the output. There are 20 wrong renditions from 6 expected results, meaning the correct results but informants still interpreted them incorrectly.

So, the study reveals that out of the expected 153 search results from InterpretBank, 134 led to a correct rendition, accounting for 87.58% effectiveness. When this is compared against the total number of searches contributed by informants, which stands at 242 in the 1st axis, searches resulting in correct hits are determined to be 55.37%, which means 134 of 242 searches with InterpretBank led to correct renditions. Further investigation into this 55.37% rate is planned but not reported here. **Table 12** (see below) details the individual contributions of each informant to the 134 successful searches, alongside the results from Cycle III.

Figure 85 illustrates the Cycle III workflow with InterpretBank. On the first axis, each informant's search count with InterpretBank is listed, totaling 176 searches—fewer than the 242 in Cycle II. Differing from Cycle II, the second axis highlights a change: *first-time terms* now exceed *non-target terms* in search frequency, occupying the foremost position, followed by the latter. Similar to Cycle II, repeated terms *(rep1, rep2)* continue to generate some search demand. Consistent with Cycle II, most searches, 155, required no typing corrections, while 21 did. 91 out of 155 searches with *typing correction* for *first-time terms* and 25 out of 155 for *non-target terms* involved no corrections. Among these 155 searches, 113 led correct character strings, indicating successful recognition by InterpretBank. After typing corrections, 14 searches resulted in correct character strings. Thus, 127 searches (113+14) in the correct character string were accepted by InterpretBank, leading to the predicted outcomes. The remaining searches, either incomplete or irrelevant to this study, are marked in light grey as in Cycle II. Out of 127 searches, there were 115 correct renditions, 7 skipped terms, and 5 incorrect renditions. As previously mentioned, in Cycle II, 242 searches resulted in 134 correct renditions, constituting a *success search hit rate* of 55.37%. However, in Cycle III, the total number of searches decreased to 176, yielding 115 correct renditions and a *success* *search hit rate* of 65.34%. Under this context, **Table 12** provides a description of each informant's search counts and *success search hit* counts across the two cycles, highlighting the efficient use of InterpretBank.

From **Table 12**, it is observed that except for Blake, Harley, and Lee, all other informants' search counts in Cycle III decreased compared to Cycle II. Blake maintained the same number of searches across both cycles, Lee increased by one, and Harley showed a more significant increase of six searches. Overall, most informants reduced their usage of InterpretBank. Conversely, the *success search hit rate* of most informants improved to varying degrees from Cycle II to Cycle III. The exceptions are Blake, Erin, and Gale; however, the remaining eight informants all showed improvements in their *success search hit rates*. However, in Cycle II, 7 informants surpassed the overall *success search hit rate* of 55.37%. In Cycle III, this number reduced, with only 4 informants exceeding the overall *success search hit rate* of 65.34%.

informant	Cycle II	searches	hits*	rate**	Cycle III	searches	hits	rate
Alex		9	5	55.56		7	4	57.14
Blake		4	4	100.00		4	2	50.00
Dana		14	11	78.57		10	8	80.00
Erin		19	11	57.89		18	8	44.44
Frankie		37	17	45.95		26	19	73.08
Gale		31	22	70.97		17	11	64.71
Harley		4	3	75.00		10	8	80.00
Ira		30	20	66.67		22	18	81.82
Jordan		47	23	48.94		27	17	62.96
Kelly		29	12	41.38		16	10	62.50
Lee		18	6	33.33		19	12	63.16

*hits: *success search hit*

** rate: *success search hit rate*

Table 12. Count and percentage of correct searches and correct renditions.

3.4.2 Ear-key span and eye-voice span

As a reminder, *ear-key span,* or *E2K,* refers to the time span between the moment when the sound wave in the audio of a source speech corresponding to a targeted chunk (like a potential problem trigger) finishes to the instant when a listener initiates her first keyboard event related to that targeted chunk. *Eye-voice span,* or *I2V,* refers to the time delay between the moment targeted information is displayed (e.g., InterpretBank retrieval) on the screen to the point when the informant articulates the corresponding output (like a rendition in the target language, here Chinese).

In Cycle II, the average E2K value was 1.925 s, which is slightly higher than the median of 1.805 s, suggesting a fairly symmetrical data distribution (**Table 13**). The standard deviation was 1.392, with the data range spanning from -1.498 s to 8.701 s. The Shapiro-Wilk test for normality indicated that the data are not normally distributed. For Cycle III, the average E2K *dropped* to 1.639 s, with a median of 1.083 s, and the outstanding mode at −0.797 s, although multiple modes exist. The standard deviation rose to 2.227 s. The Shapiro-Wilk test resulted in a W-value of 0.865 (p-value < 0.05), confirming a non-normal distribution.

								Shapiro-Wilk	
	cycle	***mean***	***median***	***mode****	***SD***	***minimum***	***maximum***	***W***	***p***
E2K	II	1.925	1.805	1.078[a]	1.392	−1.498	8.701	0.948	7.37e0-7
	III	1.639	1.083	−0.797 [a]	2.227	−1.272	8.858	0.865	2.03E-10
I2V	II	2.309	2.019	1.214 [a]	1.785	−2.6	8.732	0.954	3.21e0-5
	III	1.503	1.689	1.515 [a]	2.299	−7.778	7.473	0.821	3.86E-11

* More than one mode exists, only the first is reported

Table 13. Statistics and Shapiro-Wilk test for E2K and I2V in Cycles II and III.

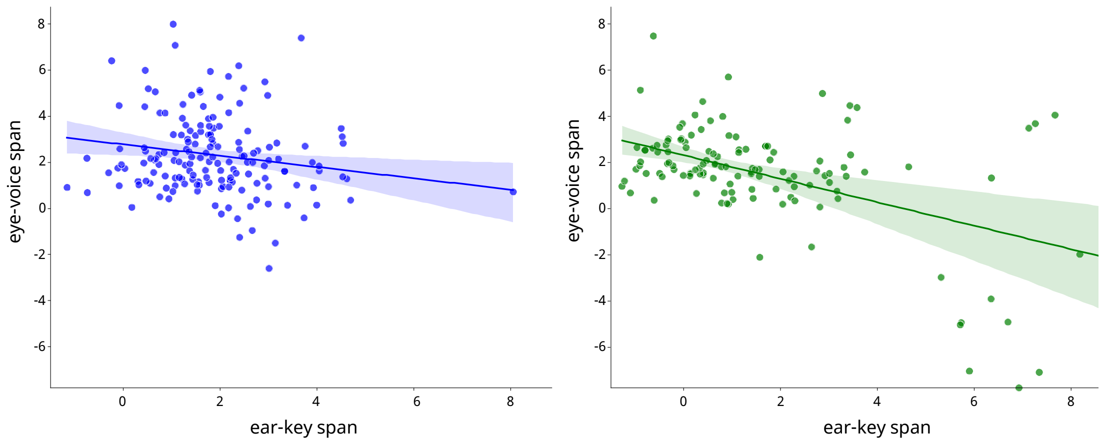


Figure 86. Scatter plots of E2K and I2V across Cycles II and III.

Back to Cycle II, the average I2V was 2.309 s; the median was 2.019 s; and the outstanding mode, was 1.214 s, with the caveat that multiple modes exist. The standard deviation was 1.785 s, and values ranged from −2.600 s to 8.732 s. The Shapiro-Wilk test (W = 0.954, p-value < 0.05) strongly suggested a non-normal distribution. In Cycle III, the average I2V was lower at 1.503 s, with a median of 1.689 s and a mode of 1.515 s. The standard deviation here was 2.299 s, pointing to a broader spread of data, between −7.778 s to 7.473 s. The Shapiro-Wilk test for this cycle pointed to a W-value of 0.821 with p lower than 0.05, reiterating that the data are not normally distributed.

Overall, the statistical metrics and the Shapiro-Wilk tests confirmed non-normal distributions and varying degrees of variation for both E2K and I2V in both cycles. Both the mean and median for E2K and I2V exhibited a downward trend from Cycle II to Cycle III. The significant p-value*s* from the Shapiro-Wilk tests underscored the need for non-parametric statistical methods (e.g., Kendall's Tau-b) in further analyses. The scatter plots in **Figure 86** and Kendall's Tau-b statistical tests offer complementary perspectives on the relationship between E2K and I2V in cycles II and III.

In Cycle II (blue scatter plot), the trend line hints at a moderately negative correlation between E2K and I2V. The data points were dispersed broadly, hinting at a varied range of performances among participants. The expanding 95% confidence interval supports this observation, pointing to a high degree of variation and unpredictability in the results. Kendall's Tau-b was −0.103 with a p-value of 0.048. These figures show a statistically significant, weak inverse correlation between the two variables. In other words, When E2K experiences growth, I2V tends to decrease, although the trend is not very meaningful.

The green scatter plot also shows a negative correlation between E2K and I2V in Cycle III, but the slope of the line of best fit is steeper compared to the blue scatter plot. The data points in the green graph are more tightly clustered around the line of best fit, especially at higher values of I2V. The expanding 95% confidence interval mirrored the variation seen in Cycle II. Kendall's Tau-b was −0.228, with a p-value of less than 0.01, implying a statistically significant, but very weak inverse correlation between the two variables. So, there is still a weak inverse relationship between E2K and I2V in Cycle III, meaning that as E2K increases, I2V tends to decrease. While the scatter plots for both cycles suggest a lack of a strong correlation, Kendall's Tau-b values and p-value*s* confirm a statistically significant inverse correlation—weak in Cycle II and somewhat stronger (though still weak to moderate) in Cycle III. That is, when E2K tends to grow, I2V tends to shrink, but the tendency is quite modest.

3.4.3 Problem triggers with InterpretBank search

The potential problem triggers in each source speech were labeled sequentially from T01 to T39, based on their order of appearance in the source speech soundtrack. Repetitions were coded with their ordinal number followed by the number of the slot when they first appeared. For instance, the term T10 in Cycle II was repeated in slot 21, so the repetition is labeled *T21_10*. When it showed up again, in slot 39, it was labeled *T39_10*. To further clarify their position within their own internal sequence, terms that would later be repeated, or *1st2Rep* terms, display an *A* after the number. For instance, T10 becomes *T10A,* indicating that the term appeared in slot 10 for the first time and (unawares for the informants) it will be repeated later in the source speech. Its first repetition adds a *B* after the number and the second repetition adds a *C* (*rep1* and *rep2 i*n **§ 3.3.5**). So, the first mention and the two repetitions of one and the same term forming an internal sequence within the 39 potential problem triggers in Cycle II are labeled *T10A, T21_10B,* and *T39_10C* (**Figure 87**).

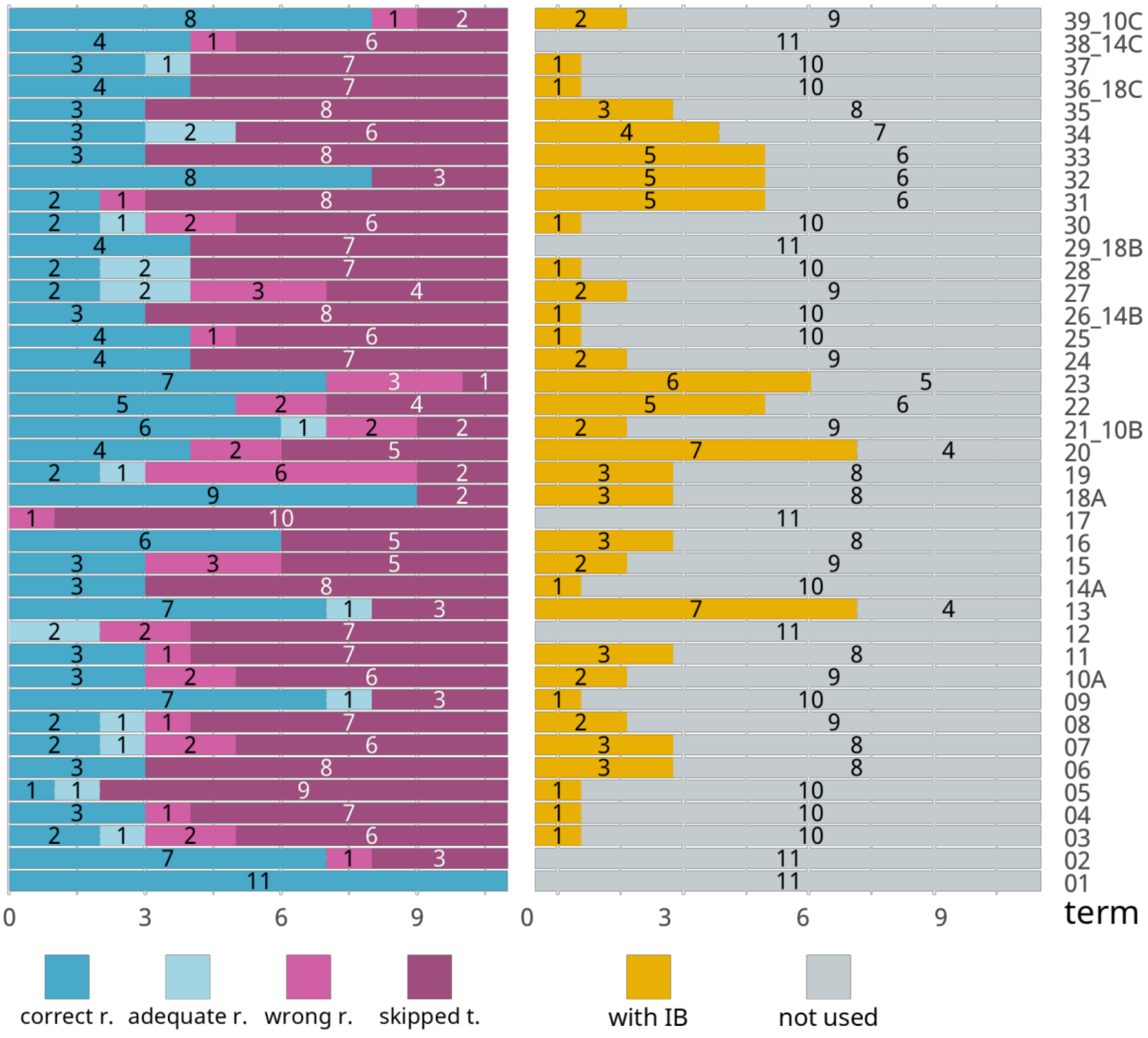


Figure 87. Accuracy of problem triggers and InterpretBank usage for Cycle II.

In **Figure 87**, the left plot illustrates the accuracy levels for the potential problem triggers by InterpretBank informants during Cycle II. Each term had a unique distribution of *correct*, *adequate*, *wrong renditions,* and *skipped terms*. For instance, T01 had an overwhelming count of 11 *correct renditions* by 11 informants. In contrast, T02 presented correct renditions by seven informants, *skipped* by three informants*,* and 1 *wrong* rendition, a mixed performance. Other terms, like T38_14C, displayed a distribution of 4 *correct*, 1 *wrong*, and no counts in the other categories, and T39_10C had 8 *correct*, 2 *skipped*, and 1 *wrong* rendition. T17 was the worst one, with 10 *skipped terms* and one wrong rendition. The right plot in **Figure 87** displays how the use of InterpretBank varied for the terms in the left plot. For example, seven informants used InterpretBank for T20 whereas only four did not use it, Interestingly, in T01, T02, T12, T17, T29_18B, and T38_14C (six terms) nobody used InterpretBank and in T03, T04, T05, T09, T14A, T25, T26_14B, T28, T30, T36_18C and T37 (eleven terms) only one informant used it. On the opposite pole, T13 and T20 had seven informants use InterpretBank, and six used it for T23.

When examining the plots side by side in **Figure 87**, we observe the dependency of InterpretBank and the accuracy of SI output. Each term in the left and right plots represents the number of informants who are *correct*, *adequate*, *wrong, and skipped*, rendered, used InterpretBank, and not used InterpretBank. Using the median number of 6 informants as a threshold, we categorize accuracy: terms with correct or adequate renditions by ≥6 informants are classified as *accurate renditions*, while those with *wrong* renditions or *skipped terms* by ≥6 informants are *inaccurate renditions*. Similarly, *InterpretBank used* and *not used* categories depend on whether ≥6 informants used InterpretBank or not. The outcome reveals that only terms T13 and T23 fall into the *accurate renditions and InterpretBank used* category (i.e., we call it *+A+U* for short). For comprehensive data, please refer to **Tables 14–15**. There is only one term, T20, in the *inaccurate renditions and InterpretBank used* (i.e., *-A+U*). In contrast, eight terms (T01, T02, T09, T16, T18A, T21_10B, T32, T39_10C) are classified as *accurate renditions but InterpretBank not used* (i.e., +A-U). However, there is a significant jump in the *inaccurate renditions but InterpretBank not used* (i.e., -A-U), encompassing 28 terms, highlighting a predominant trend within this category, including T03-T08, T10-T12, T14-T15, T17, T19, T22, T24-T31, and T33-T38.

Figure 88 presents the accuracy levels for the InterpretBank group in Cycle III (left plot), and the, InterpretBank use patterns for the 39 potential problem triggers (right plot). Again, T01 was the only one to be overwhelmingly interpreted correctly, a nearly perfect performance among the participants. T31, in contrast, was *skipped* by all informants, including the four who looked for it. This might imply that both T01 and T31 and its position might have an influence on accuracy.

The informants did not use InterpretBank at all when they tried to render T02, T03, T05, T06, and T22 (five terms) and only one did for T04, T08, T16, T18A-T20, T23_18B, T27_18C, T35_97C and T37 (ten terms). The term T35_97C stands out in that, as a second repetition, it would be expected to cause little to no trouble. However, half of the informants skipped the term where it appeared. Nothing seems to suggest that there might be particular difficulties in the sentence where it appeared:

> [Excerpt from the source speech text in Cycle III]:
> *And this feels good when it is caused by events that you're looking forward to, and production of these neurochemicals is inhibited by events you're looking forward to that don't work out.*
> *This is called* ***reward prediction error****.*
> *And as I mentioned, these amino acid sensors in our gut are detecting how many amino acids are there but they're also detecting which amino acids.*

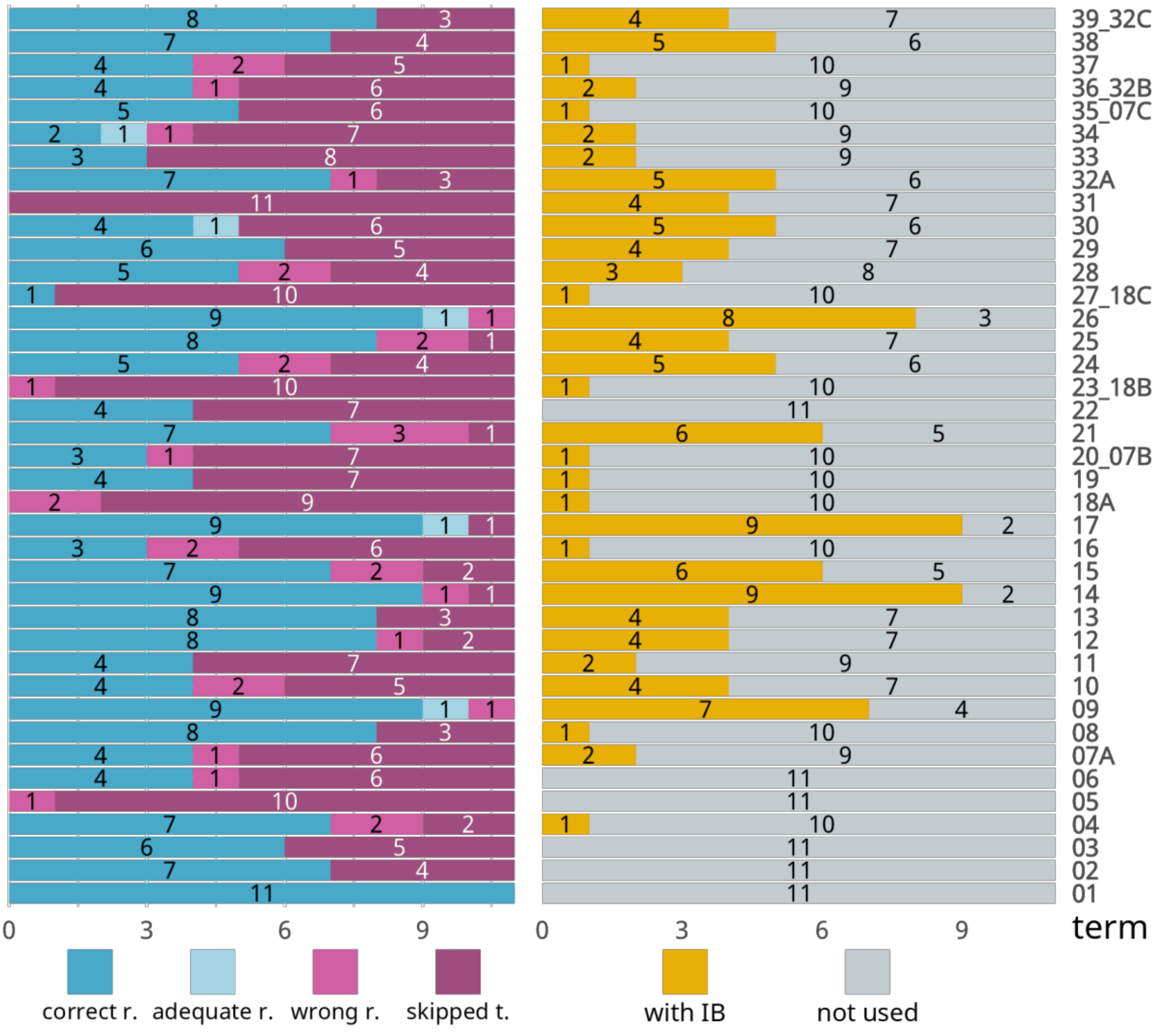


Figure 88. Accuracy of problem triggers and InterpretBank usage for Cycle III.

In Cycle III (see **Figure 88**), we observed notable changes in the dependency on InterpretBank and the accuracy of SI renditions compared to Cycle II. The category *+A+U* saw an increase, now encompassing four terms: T09, T14, T15, T17, and T21, indicating a slight improvement. There were no terms in the *-A+U*. The *+A-U* category expanded to include 12 terms: T01-T04, T08, T12, T13, T25, T29, T32A, T38, and T39_32C. Even with less frequent use of InterpretBank, several terms still achieved high accuracy in interpretation. The *-A-U* decreased to 21 terms, including T05-T07A, T10, T11, T16, T18A-T20_07B, T22-T24, T27_18C, T28, T30, T31, and T33-T37. Despite the decrease, this category still represents a substantial proportion of the 39 terms.

Tables 14–15 illustrate the accuracy of 39 potential problem trigger renditions and their usage of InterpretBank by the 11 informants, and features related to term length. We summed up the *correct* and *adequate* renditions as *accurate* counts, and the *wrong* renditions and *skipped* terms as inaccurate counts. The usage was categorized as either used InterpretBank (i.e., *with IB*) or not used InterpretBank (i.e., *no used*). We observed the informants' performance in terms of term accuracy (*accurate* renditions, *inaccurate* renditions) and InterpretBank usage (*with IB*, *not used*). We set a threshold of six, meaning any category (*accurate* counts*, inaccurate* counts*, with used, not used InterpretBank*) with a count greater than six (i.e., more than six formants) is considered. For instance, the term 07A *reward prediction error* is a trigram that was selected as a repeated term. It has 7 counts (7 informants) of inaccurate renditions and 9 counts (9 informants) of not using InterpretBank, so it is classified as -A-U, which means inaccurate and not used with InterpretBank.

Cycle II terms		accurate renditions			inaccurate renditions			with IB	not used	+A+U	-A+U	+A-U	-A-U
		sum	correct	adequate	sum	skipped	wrong						
01	hormones	11		11	0			0	11			uni.	
02	immune system	7		7	4	3	1	0	11			bi.	
03	cortisol	3	1	2	8	6	2	1	10				uni.
04	epinephrine	3		3	8	7	1	1	10				uni.
05	estrogen	2	1	1	9	9		1	10				uni.
06	cholesterol	3		3	8	8		3	8				uni.
07	dietary cholesterol	3	1	2	8	6	2	3	8				bi.
08	stress hormone	3	1	2	8	7	1	2	9				bi.
09	adrenaline	8	1	7	3	3		1	10			uni.	
10A	neuroplasticity	3		3	8	6	2	2	9				uni.
11	corticotropin releasing hormone	3		3	8	7	1	3	8				tri.
12	pituitary	2	2		9	7	2	0	11				uni.
13	insomnia	8	1	7	3	3		7	4	uni.			
14A	blood vessels	3		3	8	8		1	10				bi.
15	Arteries	3		3	8	5	3	2	9				uni.
16	stress response	6		6	5	5		3	8			bi.	
17	net effect	0			11	10	1	0	11				bi.
18A	sympathetic chain ganglia	9		9	2	2		3	8			tri.	
19	chronic cortisol elevation	3	1	2	8	2	6	3	8				tri.
20	non-sleep deep rest	4		4	7	5	2	7	4		tri.		
21_10B	neuroplasticity	7	1	6	4	2	2	2	9			uni.	
22	stress threshold	5		5	6	4	2	5	6				bi.
23	high-intensity interval training	7		7	4	1	3	6	5	tri.			
24	abdominal fat accumulation	4		4	7	7		2	9				tri.
25	immune response	4		4	7	6	1	1	10				bi.

Cycle II terms		accurate renditions			inaccurate renditions			with IB	not used	+A+U	-A+U	+A-U	-A-U
		sum	correct	adequate	sum	skipped	wrong						
26_14B	blood vessel	3		3	8	8		1	10				bi.
27	neural energy	4	2	2	7	4	3	2	9				bi.
28	chronic stress	4	2	2	7	7		1	10				bi.
29_18B	sympathetic chain ganglia	4		4	7	7		0	11				tri.
30	negative feedback loop	3	1	2	8	6	2	1	10				tri.
31	melanocytes	2		2	9	8	1	5	6				uni.
32	sympathetic nervous system	8		8	3	3		5	6			tri.	
33	hair stem cells	3		3	8	8		5	6				tri.
34	melanocyte stem cells	5	2	3	6	6		4	7				tri.
35	low-density lipoprotein cholesterol	3		3	8	8		3	8				tri.
36_18C	sympathetic chain ganglia	4		4	7	7		1	10				tri.
37	psychological stress	4	1	3	7	7		1	10				bi.
38_14C	blood vessel	4		4	7	6	1	0	11				bi.
39_10C	neuroplasticity	8		8	3	2	1	2	9			uni.	

categories based on accuracy counts and InterpretBank usage (threshold ≥ 6)
+A+U: **accurate** renditions and **IB** (InterpretBank) used
-A+U: **inaccurate** renditions and **IB** used
+A-U: **accurate** renditions but IB **not used**
-A-U: **inaccurate** renditions but IB **not used**

Table 14. N-gram analysis of problem triggers using InterpretBank in Cycle II.

Cycle III term		accurate renditions			inaccurate renditions			with IB	not used	+A+U	-A+U	+A-U	-A-U
		sum	correct	adequate	sum	skipped	wrong						
01	emotions	11	11	0	0	0	0	0	11			uni.	
02	carbohydrates	7	7	0	4	4	0	0	11			uni.	
03	micronutrients	6	6	0	5	5	0	0	11			uni.	
04	vagus nerve	7	7	0	4	2	2	1	10			bi.	
05	10th cranial nerve	0	0	0	11	10	1	0	11				tri.
06	neurons	4	4	0	7	6	1	0	11				uni.
07A	reward prediction error	4	4	0	7	6	1	2	9				tri.
08	heart rate	8	8	0	3	3	0	1	10			bi.	
09	polyvagal theory	10	9	1	1	0	1	7	4	bi.			
10	dorsal vagus	4	4	0	7	5	2	4	7				bi.
11	spinal cord	4	4	0	7	7	0	2	9				bi.
12	hypothalamus	8	8	0	3	2	1	4	7			uni.	
13	lateral hypothalamus	8	8	0	3	3	0	4	7			bi.	
14	locus coeruleus	9	9	0	2	1	1	9	2	bi.			
15	amino acid	7	7	0	4	2	2	6	5	bi.			
16	neurochemicals	3	3	0	8	6	2	1	10				uni.
17	intestines	10	9	1	1	1	0	9	2	uni.			
18A	L-tyrosine	0	0	0	11	9	2	1	10				uni.
19	plant-based foods	4	4	0	7	7	0	1	10				bi.
20_07B	reward prediction error	3	3	0	8	7	1	1	10				tri.
21	raphae nucleus	7	7	0	4	1	3	6	5	bi.			
22	antidepressants	4	4	0	7	7	0	0	11				uni.
23_18B	L-tyrosine	0	0	0	11	10	1	1	10				uni.
24	gut brain axis	5	5	0	6	4	2	5	6				tri.
25	blood brain barrier	8	8	0	3	1	2	4	7			tri.	

Cycle III term		accurate renditions			inaccurate renditions			with IB	not used	+A+U	-A+U	+A-U	-A-U
		sum	correct	adequate	sum	skipped	wrong						
26	long-chain fatty acids	10	9	1	1	0	1	8	3	tri.			
27_18C	L-tyrosine	1	1	0	10	10	0	1	10				uni.
28	fatty acid ratio	5	5	0	6	4	2	3	8				tri.
29	heart rate variability	6	6	0	5	5	0	4	7			tri.	
30	autonomic nervous system	5	4	1	6	6	0	5	6				tri.
31	respiratory sinus arrhythmia	0	0	0	11	11	0	4	7				tri.
32A	gut microbiome	7	7	0	4	3	1	5	6			bi.	
33	prebiotics	3	3	0	8	8	0	2	9				uni.
34	central nervous system	3	2	1	8	7	1	2	9				tri.
35_07C	reward prediction error	5	5	0	6	6	0	1	10				tri.
36_32B	gut microbiome	4	4	0	7	6	1	2	9				bi.
37	neurotransmitters	4	4	0	7	5	2	1	10				uni.
38	circadian type fasting	7	7	0	4	4	0	5	6			tri.	
39_32C	gut microbiome	8	8	0	3	3	0	4	7			bi.	

categories based on accuracy counts and InterpretBank usage (threshold ≥ 6)
+A+U: **accurate** terms and used **with IB** (InterpretBank)
-A+U: **inaccurate** terms but used **with IB**
+A-U: **accurate** terms but **not used** with IB
-A-U: **inaccurate** terms and **not used** with IB

Table 15. N-gram analysis of problem triggers using InterpretBank in Cycle III.

Cycle	N-gram	+A+U	-A+U	+A-U	-A-U
II	unigrams	1	0	4	8
	bigrams	0	0	2	11
	trigrams	1	1	2	9
III	unigrams	1	0	4	8
	bigrams	4	0	5	4
	trigrams	1	0	3	9

Table 16. N-gram counts by accuracy and InterpretBank usage.

Based on **Tables 14** and **15,** we summarized the N-gram counts in the four categories (see **Table 16**): *+A+U*, *-A+U*, *+A-U*, and *-A-U*. The *-A-U* category had relatively prominent counts. For instance, we found that 8 out of 39 unigrams and 9 out of 39 trigrams in both Cycle II and Cycle III belonged to the *-A-U* category. The bigram counts in the *-A-U* category varied greatly, from 11 in Cycle II to 4 in Cycle III. We only focused on term length (unigram, bigram, and trigram), because other factors, such as frequency and transparency, were in this case less likely to yield clear results.

Cycle	repeated terms	words		chunks		sentences	
		1st2Rep-rep1	***rep1-rep2***	***1st2Rep-rep1***	***rep1-rep2***	***1st2Rep-rep1***	***rep1-rep2***
I	*melatonin*	200	235	11	14	11	14
	light-dark cycle	188	282	10	15	10	15
	resting blood glucose	253	436	15	21	15	21
II	*neuroplasticity*	543	798	29	40	24	42
	blood vessels	668	474	34	22	32	22
	sympathetic chain ganglia	798	307	42	14	40	14
III	*reward prediction error*	568	696	24	38	24	37
	L-tyrosine	197	180	10	7	9	7
	gut microbiome	136	115	10	7	10	7

Table 17. Distances between *repeated terms and1st2Reps* .

Next, we report the detailed data regarding the frequency of *repeated* terms and their search frequency in InterpretBank **Table 17** lists the following distance measures for each Cycle: *words distance*—the number of words between the start of the first word after *1st2Rep* and the end of the last word before *rep1*, as well as the number of words between the start of the first word after *rep1* and the end of the last word before *rep2*; *chunks distance*—all chunks in the source speech texts were labeled with numbers in corresponding cycles (see **Appendix H**), counting the difference between the chunk number containing *1st2Rep* and the chunk number containing *rep1*, as well as the difference between the chunk number containing *rep1* and the chunk number containing *rep2*; *sentence distance*—all sentences in the source speech texts were labeled with numbers in corresponding cycles (see **Appendix H**), counting the difference between the sentence number containing *1st2Rep* and the sentence number containing *rep1*, as well as the difference between the sentence number containing *rep1* and the sentence number containing *rep2*.

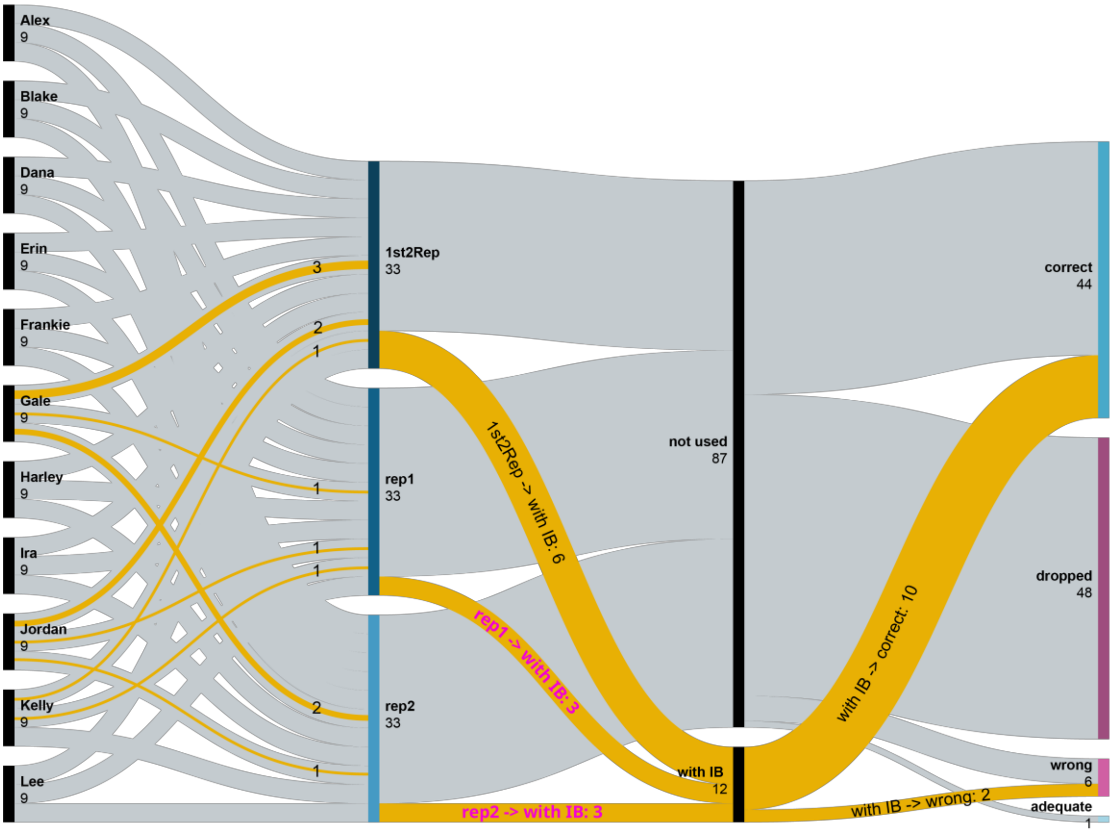


Figure 89. Search for *repeated terms* in problem triggers in Cycle II.

The Sankey diagram in **Figure 89** depicts the use of InterpretBank when handling specific term types with more detail—*1st2Rep*, *rep1*, and *rep2*. Term appearing for the first time were portrayed in **Figure 89** as well to enable comparisons. The diagram has four axes: The first axis lists all informants in the InterpretBank group, indicating a total of 9 potential problem triggers for each informant comprising 3 instances each of *1st2Rep*, *rep1*, and *rep2*, amounting to 66 repeated terms + 33 first-time terms across all informants in Cycle II. The second axis categorizes these terms into *1st2Rep*, *rep1*, and *rep2*. The numbers associated with each category reflect the count of terms within each, calculated as 3 terms multiplied by 11 informants. The third axis details the usage of InterpretBank, including options *not used* and *with IB*. The 4th or final axis relates to the accuracy of SI output. Each block on this axis is numbered to indicate term counts. Specifically, we use streams to connect each block between axes. The yellow stream denotes searches with InterpretBank. For instance, among the 11 informants, only three searched for *1st2Reps*, including 3 term searches from Gale, 2 from Jordan, and 1 from Kelly, totaling 6 searches with InterpretBank. The search count decreases for *rep1*, with only one search each from

Gale, Jordan, and Kelly. Following this, there are 3 searches for *rep2*, conducted by 2 informants: 2 from Gale and 1 from Jordan. This brings the total to 12 searches (6 from *1st2Reps,* 6 from *repeated* terms) with InterpretBank, leading to 10 terms rendered correctly *(5 from 1st2Reps, 5 from repeated terms)* and 2 incorrect searches *(1 from 1st2Rep, 1 from rep1).* Therefore, 5 out of the 66 repeated terms were supported by InterpretBank for accurate renditions.

However, the relatively low frequency of its use—only 5 instances across both repetitions—indicates that most participants preferred not to use InterpretBank for these terms. Hence, perhaps there is a component of *confirmatory* searching. That is, perhaps the informational needs of informants for these terms were not so pressing in repetitions.

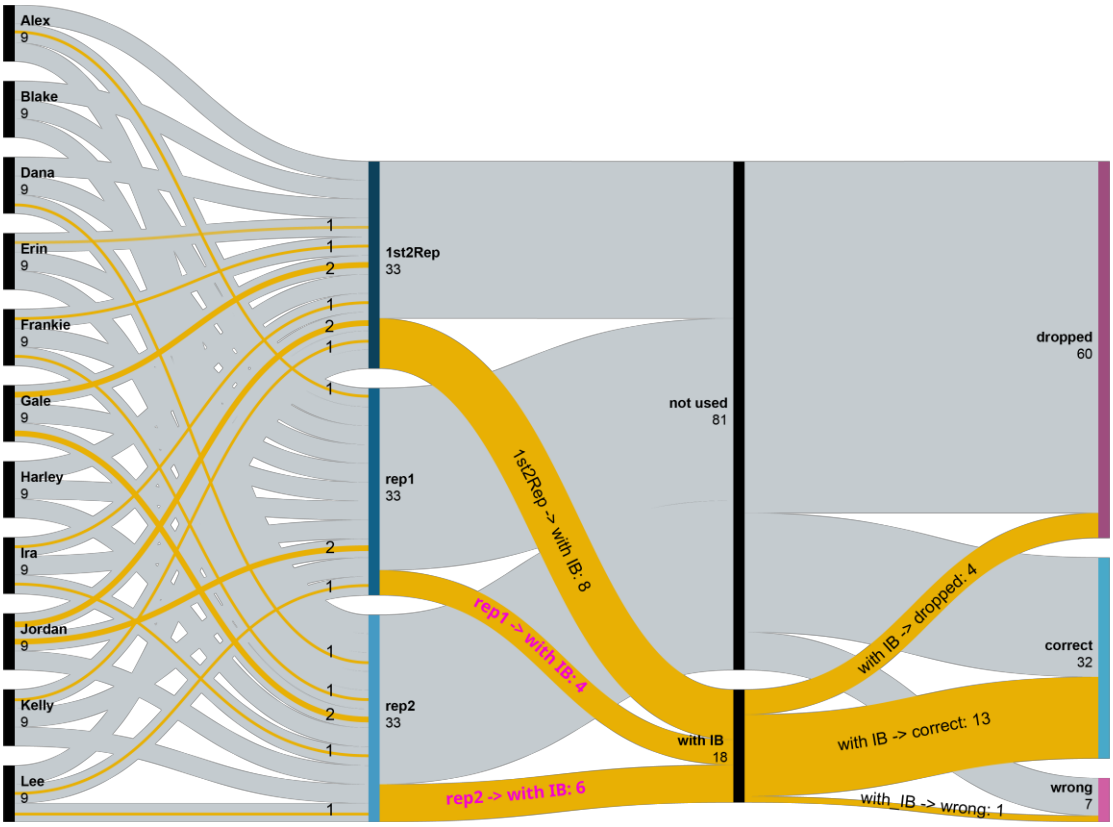


Figure 90. Search for *repeated terms* in problem triggers in Cycle III.

The Sankey diagram in **Figure 90** depicts how informants from the InterpretBank group handled terms related to repetitions—*1st2Rep, rep1*, and *rep2* in Cycle III. Similar to Cycle II, each of the 11 informants handled 9 repeated terms, amounting to 66 repeated terms + 33 first-time terms across all informants in Cycle II.

For *1st2Rep* terms in Cycle III, the number of informants engaging with these terms with InterpretBank increased to 6 (including Erin, Frankie, Gale, Ira, Jordan, and Kelly), with the term search count rising to 8, an increase from the 6 term searches by 3 informants in Cycle II. When *1st2Rep* terms are repeated for the first time, as in *rep1*), we observed 4 searches in total for *rep1* from three informants using InterpretBank (2 searches from Jordan, one from Alex, and one from Lee).

However, when *1st2Rep* terms are repeated for the second time, marked as *rep2*, the number of informants adopting InterpretBank rebounded to 6, including one term search each from Dana, Frankie, Ira, and Lee, and 2 term searches from Gale, making up 6 term searches for *rep2*.

Altogether, for the repeated terms (*1st2Rep, rep1*, and *rep2*), there were 18 term searches *(8 from 1st2Reps, 10 from repeated terms),* including 13 leading to *correct* renditions *(5 from 1st2Reps, 8 from repeated terms),* 4 resulting in *skipped* terms *(2 from 1st2Reps, 2 from repeated terms),* and 1 *wrong rendition from 1 1st2Rep*. This shows a difference from Cycle II, where only 5 out of the 66 *repeated* terms received adequate support from InterpretBank for *correct* renditions. In comparison, Cycle III exhibited a light improvement with 8 out of 66 repeated terms being accurately rendered with the support of InterpretBank.

3.4.4 Search duration and dropped chunks

To facilitate an understanding of search duration and *dropped chunks* and enable a consistent comparison, these events in informants' recordings from Cycles II and III were aligned onto respective universal timelines. Based on the method described in **§ 2.7.1.2**, we aligned all informants' SI output soundtracks into a universal timeline, virtually starting the soundtrack recording at the same time.

Figure 91 consists of two diagrams each cycle (i.e., Cycles II and III), with the left diagram depicting two events that occurred in the universal time: the distribution *search action* and *dropped chunks*. The right-hand side shows the counts of *search action* counts and *dropped chunk* counts for each informant. The left side features a universal timeline of 803.804 s in the cycle.

The X-axis represents time, indicating the duration of the soundtrack in seconds, while the Y-axis lists the informants by their initials, for instance, Alex (A). Blue spans on this diagram represent *dropped chunks*. Following the alignment of SI rendition transcripts with chunks of the source speech script, we marked the informants' dropped source speech sentences, with yellow spans indicating the duration of these *dropped chunks* in the universal time. Similarly, yellow spans denote the duration of searches conducted with InterpretBank. Each span on the universal timeline varies in length, representing different durations, and is colored according to the type of event.

Pink dots are used to indicate overlap points, where *search actions* coincide with *dropped chunks*. This overlap is determined by comparing the onset and offset timestamps of both the *search action* and the *dropped chunks*. An overlap can occur at any moment during an event, for example, a *search action* might completely cover the duration of a *dropped chunk*, or it might partially overlap with the beginning of a *dropped chunk* event. The exploration of the position of these overlaps may shed light on individual differences, the relationship between *searches* and *dropped chunks,* and the dynamics of searching as impacted by the unfolding of the task (possibly due to changes in cognitive efforts and demands) but only the third option falls within the feasible scope of this study; As for the rest, suffice it for now to state that we used pink dots to signify where two events overlap so that the reader can have a glimpse at the potential of this approach.

The right-hand side in **Figure 91** presents the counts of *search actions* and *dropped chunks* occurrences for each informant, each labeled with a number in the corresponding bar. The blue bars correspond to *dropped chunks*, echoing the spans on the left diagram, while the yellow bars represent *search actions*.

For the universal timeline that extends to 803.804 s in Cycle II, we divided the timeline into three parts: *beginning* (0–200 s), *middle* (200–600 s), and *toward the end* (600–800 s). In Cycle II, the session began with 13 overlaps within 200 s, witnessed 12 in the *middle* spread within 400 s, and 13 overlaps *toward the end* within 200 s. That is, the searches were quite evenly distributed. In Cycle III, with a universal timeline of 788.577 s, as shown in **Figure 91**, the same segmentation strategy would reveal differences compared to Cycle II. The *beginning* of Cycle III had only one instance, while the *middle* had a moderate increase to 7 overlaps within 400 s, and there were 6 instances *toward the end*, also a decrease in density.

Both figures collectively illustrate that searches were not balanced between Cycles II and III and within Cycle III. The informants made more than twice the number of searches in Cycle II than in Cycle III and nearly none at the beginning of Cycle III. This suggests that customary indicators of source text difficulty and speech delivery difficulty (see **Table 1** and **Table 2**) may not correctly represent the actual difficulties interpreters face. In order to illustrate differences in individual performance, let us mention that Alex (A), Blake, and Dana in Cycle II and Erin (E), Frankie, and Gale in Cycle III presented unique patterns. For instance, Alex *dropped* 8 sentences and performed 9 *searches,* with overlaps at timestamps 50.43 s, 205.86 s, and 635.98 s. Erin, on the other hand, exhibited *dropped* 7 sentences and performed 18 *searches,* with overlaps at 413.49 s, 647.10 s, and 698.23 s. These variations did not affect only the ratio of *dropped chunks* vs *searches* but also their frequency and the duration of their overlaps (see **Figures 84-85** for results on correct renditions out of searches).

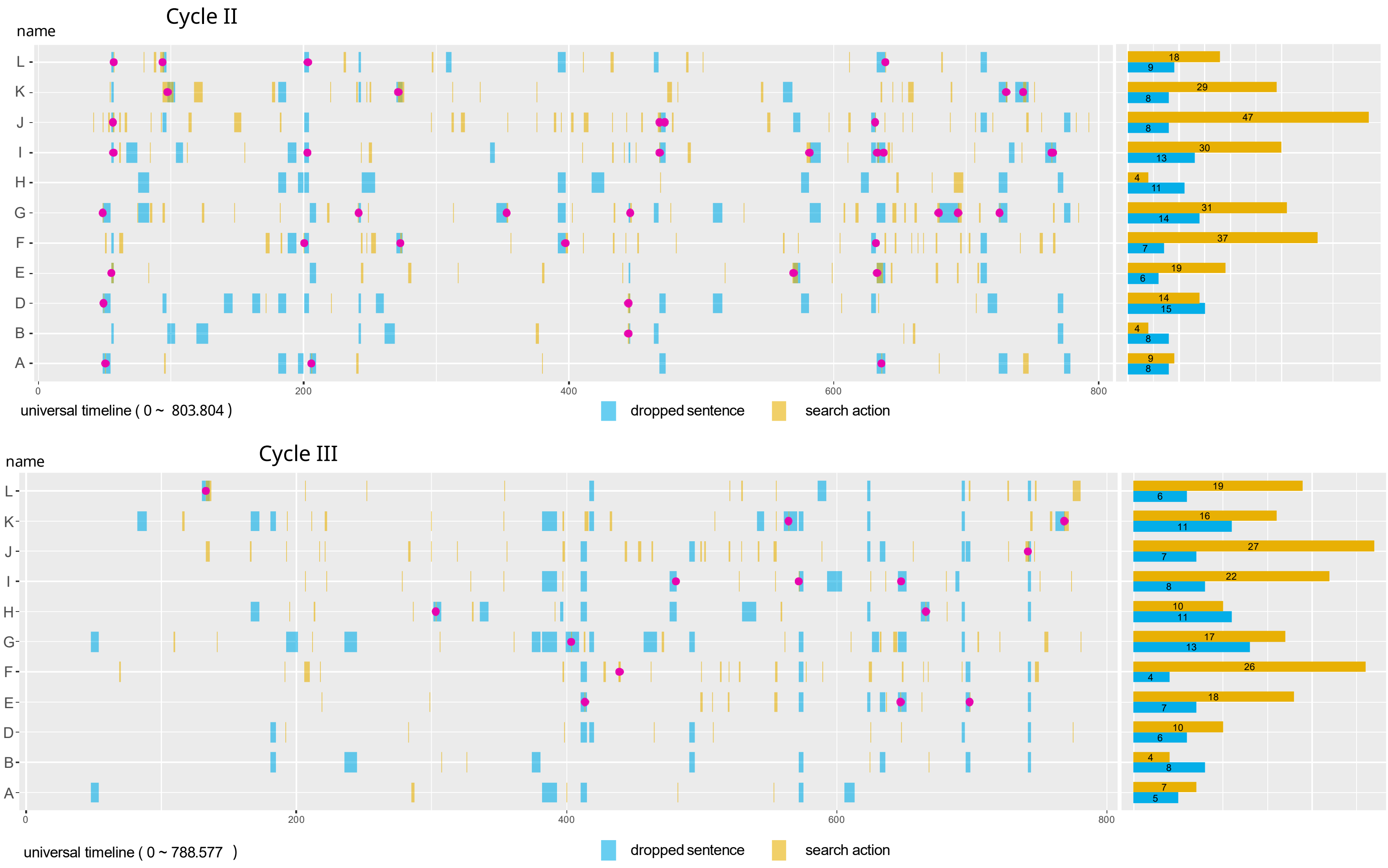


Figure 91. Event counts and durations with overlaps for Cycles II and III.

3.5 Survey analysis

After Cycles II and III, the InterpretBank group received two identical surveys. The same brief survey was administered to the Excel group, in this case only after Cycle III.

3.5.1 Surveys of the InterpretBank informants

The surveys for the InterpretBank group (see **Appendix E**) covered three areas: overall opinion, glossary tasks, and booth tasks. We focus on an intra-subject analysis, primarily on individual changes.

3.5.1.1 *Overall opinion* **Figure 92** displays the informants' self-assessment of their own interpreting performance, whose changes between Cycles II and III were varied. For instance, Dana Erin and Lee thought their performance improved in Cycle III, compared to their opinion in Cycle II. Their upswing was always of 2/10, which for Dana and Erin changed to a positive assessment that coincided with other informants in upper values. In contrast, Alex and Harley thought they performed worse in Cycle III than in Cycle II, although such a fall in performance was only 1/10. Interestingly, about half (5) of the informants thought that their performance was the same, even when it was at different levels, from low (Blake, 4/10) to high (Ira and Kelly, 6/10).

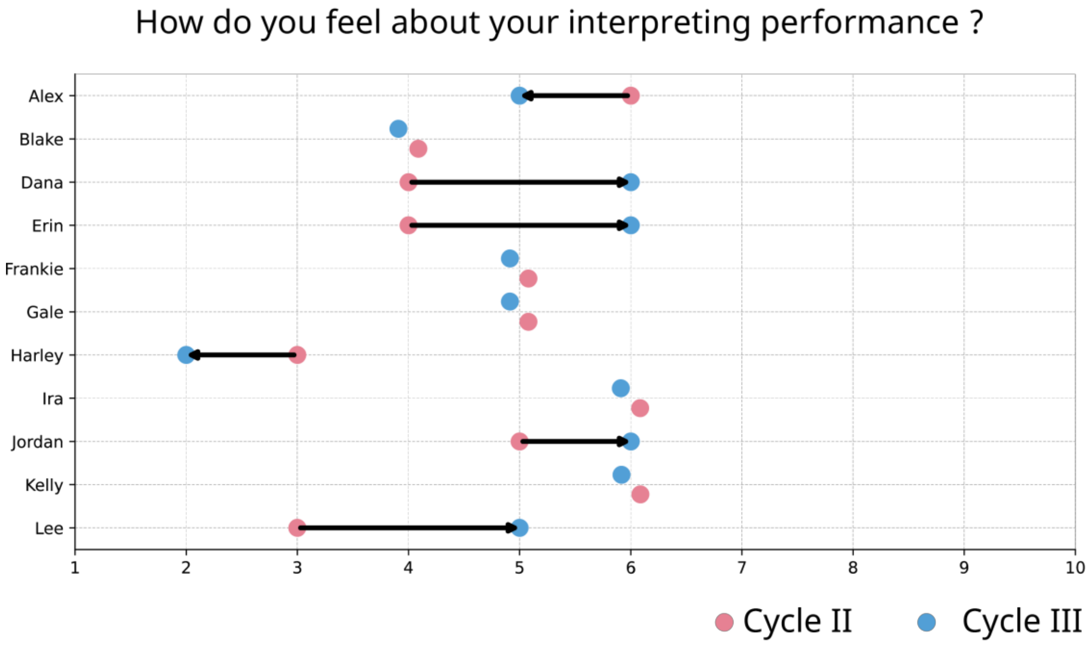


Figure 92. Self-assessment of interpreting performance.

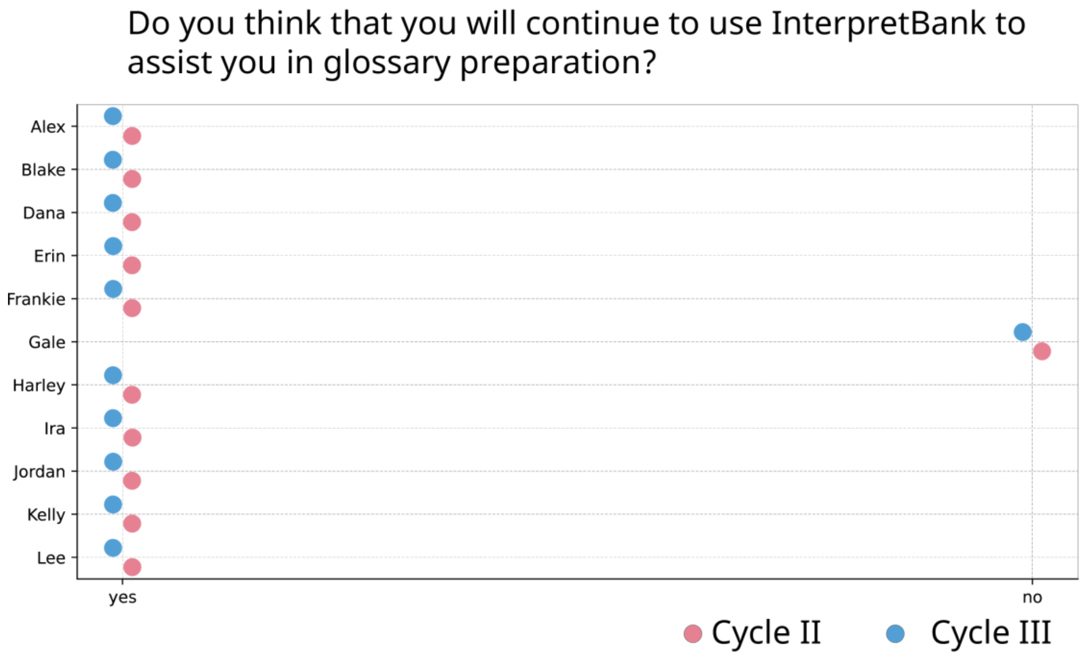


Figure 93. Intent to continue using InterpretBank for glossary preparation.

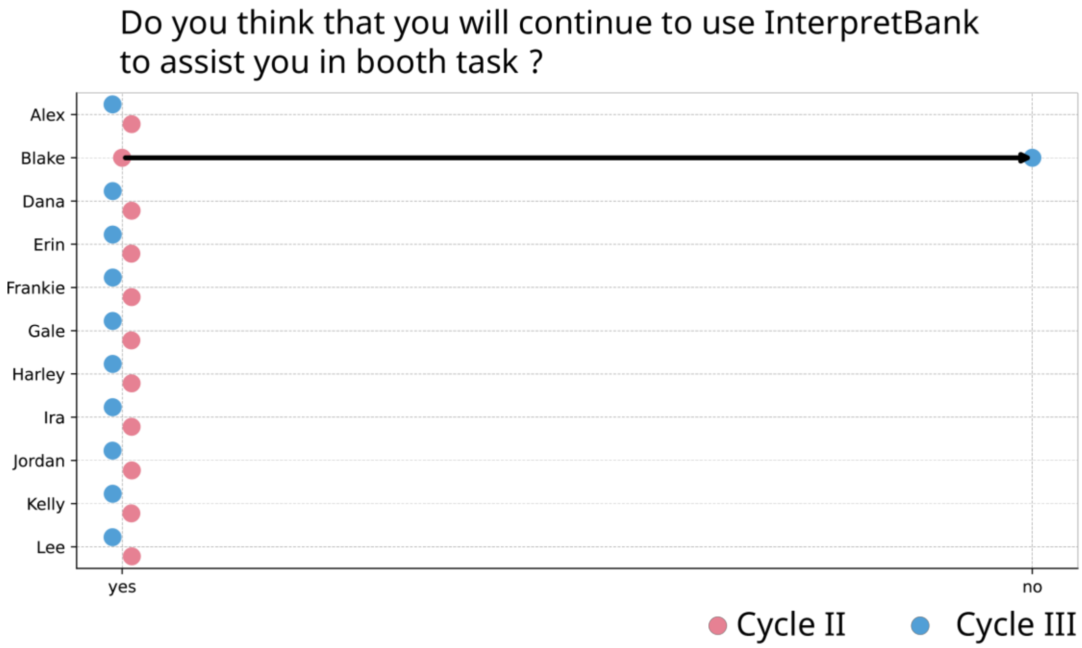


Figure 94. Intent to continue using InterpretBank for booth tasks.

Half of the informants (5) did not seem to feel their performance changed from Cycle II to Cycle III. The rest were evenly distributed between those who thought there was a slight change (3, one degree) or a more noticeable change (3, two degrees). Among those who thought there was a change, two-thirds felt they had improved (4) and the other third (2) thought they had actually performed worse. The opinion of two informants was always below the midpoint, and no informant was always above the midpoint (5-6).

Regarding the use of InterpretBank to prepare the glossaries (**Figure 93**) all informants but Gale declared they would keep using InterpretBank for this. As for using InterpretBank for booth tasks (**Figure 94**) also records a high degree of consistency in informants' responses. The question was worded quite openly so as not to lead informants, so it might have been interpreted as referring to either Cycle III or after the experiment. However, the only informant to drop InterpretBank use in Cycle III declared her intention to keep using it (both to prepare glossaries and at the booth), so misinterpretations seem unlikely. All informants but Blake chose the option *yes* when they were asked in Cycle III

> *Do you think that you will continue to use InterpretBank to assist you in booth tasks?*

Blake, however, changed from *yes* to *no*. Considered together, **Figures 92–94** show contending opinions on the effects of InterpretBank on their performance, with most informants thinking it did not change their quality. Yet the informants shared a largely consistent and positive outlook on the future use of InterpretBank for both glossary preparation and booth tasks.

3.5.1.2 *Glossary task*

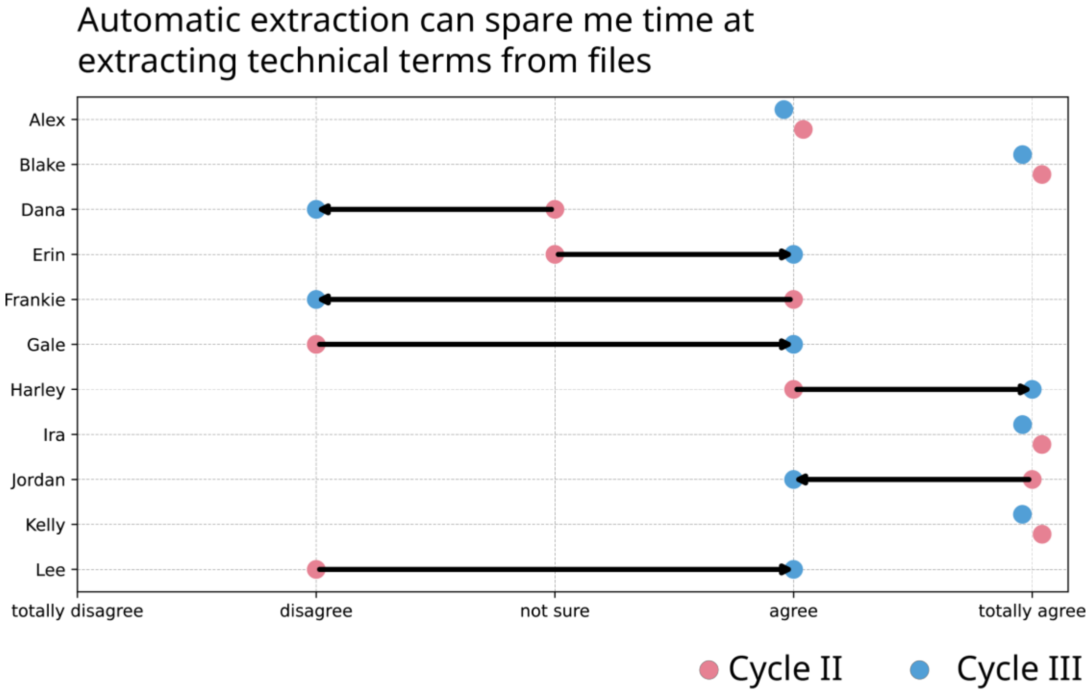


Figure 95. Attitudes toward automatic extraction of technical terms.

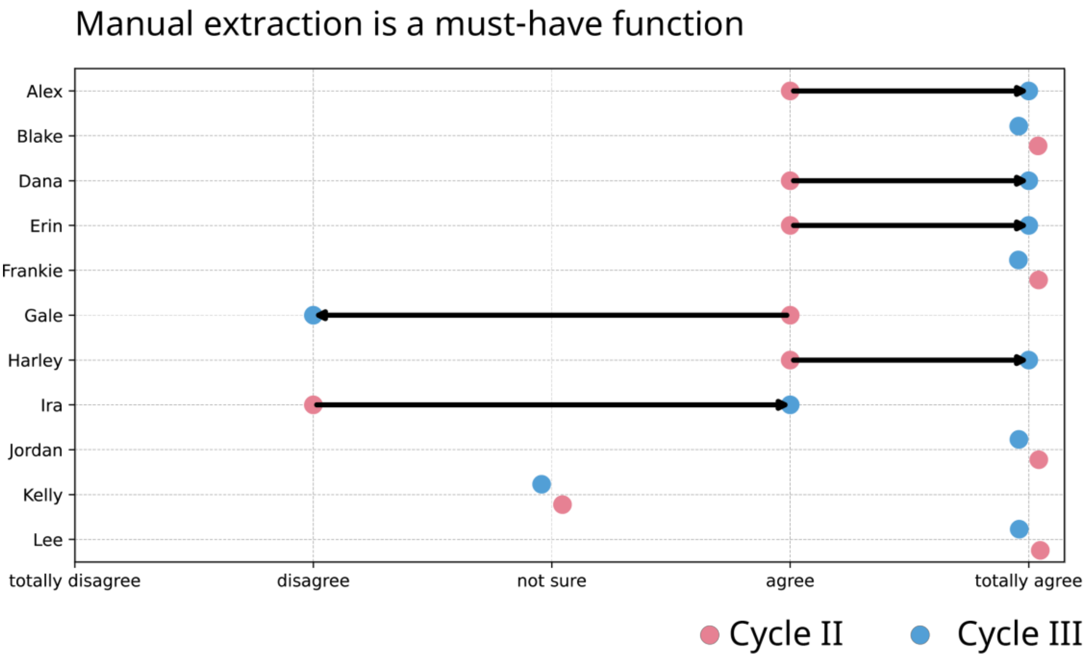


Figure 96. Views on the need for manual extraction.

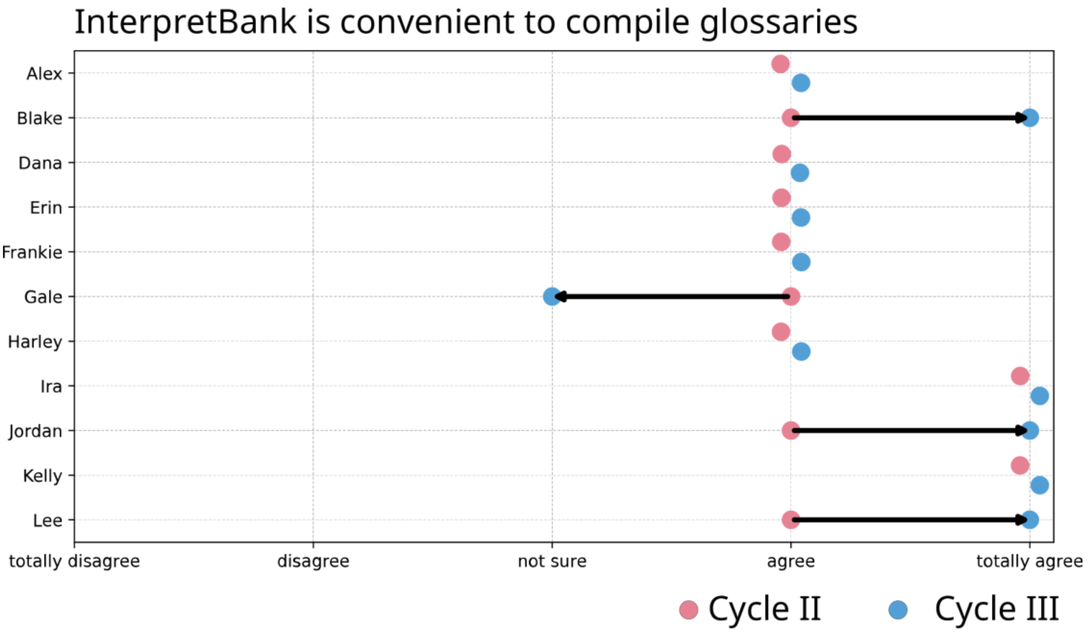


Figure 97. InterpretBank convenience for compiling glossaries.

Figure 95 shows different opinions on how useful InterpretBank's feature is for automatic term extraction. Alex consistently agreed that it saves time and Blake was even more enthusiastic. Three informants were more skeptical about the time-saving benefits of automatic extraction in Cycle III, compared to Cycle II: Dana went from *not sure* to *disagree;* Frankie, from *agree* to *disagree,* and Jordan, from *totally agree* to just *agree.* On the other hand, Erin, Gale, and Lee were more convinced in Cycle III. The last two, notably, changed their minds. Four informants (e.g., Alex, were positive or very positive and did not change their minds between cycles. For the need for a feature for manual extraction (**Figure 96**) in InterpretBank, only Gale disagreed in Cycle III, and Kelly remained unsure. The others *totally agreed* from the start or were convinced after Cycle III.

Figure 97 discusses whether the informants agree with the statement, that *InterpretBank is convenient for compiling glossaries.* Out of 11 informants, 7 maintained their stance throughout. Among them, Alex, Dana, Erin, and Frankie expressed *agree*, while Ira and Kelly held a *totally agree* position. Meanwhile, Black's attitude shifted from *agree* in Cycle II to *totally agree* in Cycle III. However, Gale's opinion changed from *agree* in Cycle II to *not sure* in Cycle III. By Cycle III, 10 out of 11 informants (*agree* and *totally agree*) supported the statement, with only one informant remaining *not sure*.

3.5.1.3 *Booth task* **Figure 98** examines informants' perspectives on the statement, *I use Booth mode when I notice a technical term.* Booth mode, a key function of InterpretBank, offers term retrieval during interpreting tasks. Over half of the informants (6/11) changed their stance from Cycle II to Cycle III regarding this core function. Indicating a positive shift toward greater reliance on InterpretBank, five informants altered their responses: Harley and Lee shifted from *sometimes* to *always*, and Alex moved from *mostly* to *always*. Similarly, Frankie and Ira transitioned from *sometimes* to *mostly*. Conversely, Erin's attitude decreased in enthusiasm, going from *always* to *mostly* between Cycle II and Cycle III. Dana and Jordan maintained a consistent *sometimes* position, indicating neutrality, while Gale's response remained steady *mostly* from Cycle II to Cycle III.

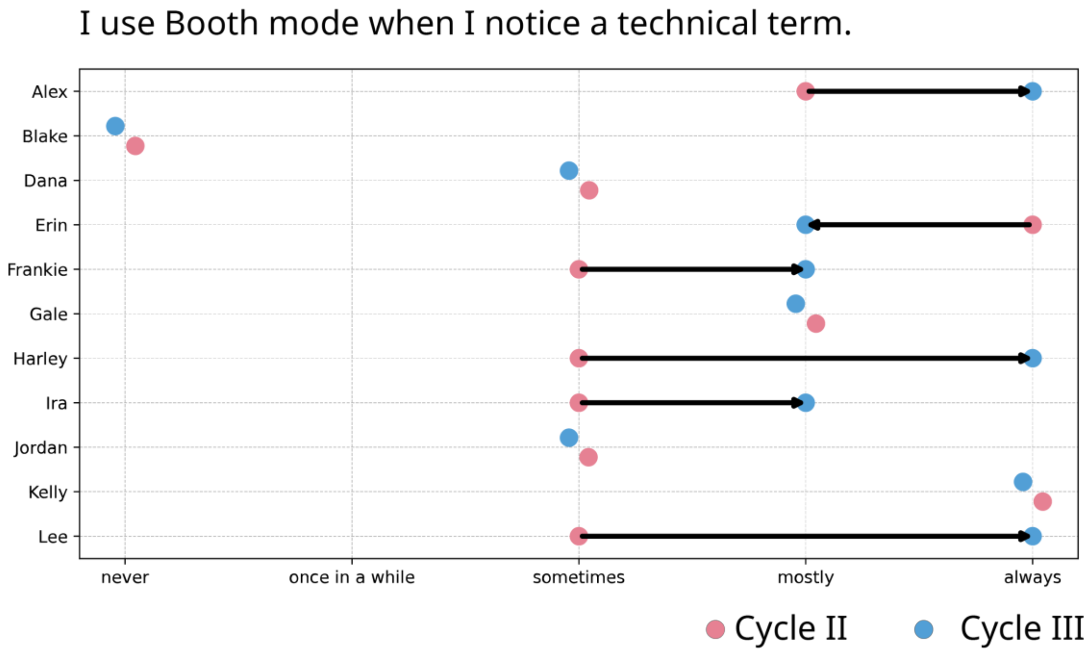


Figure 98. Frequency of using booth mode for technical terms.

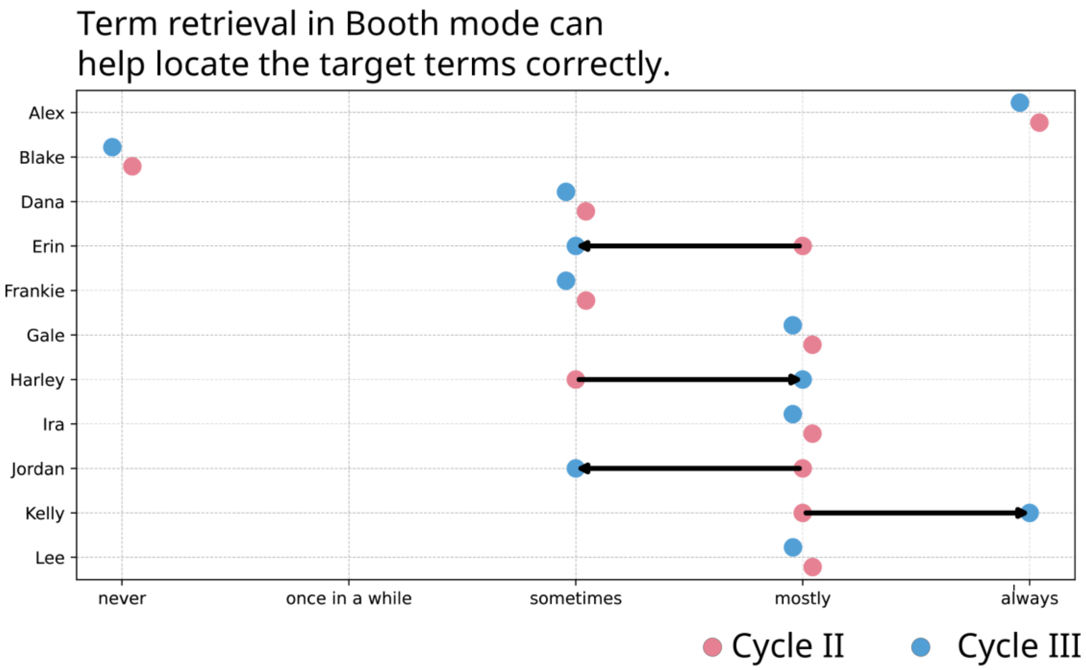


Figure 99. Confidence in booth mode's term retrieval capabilities.

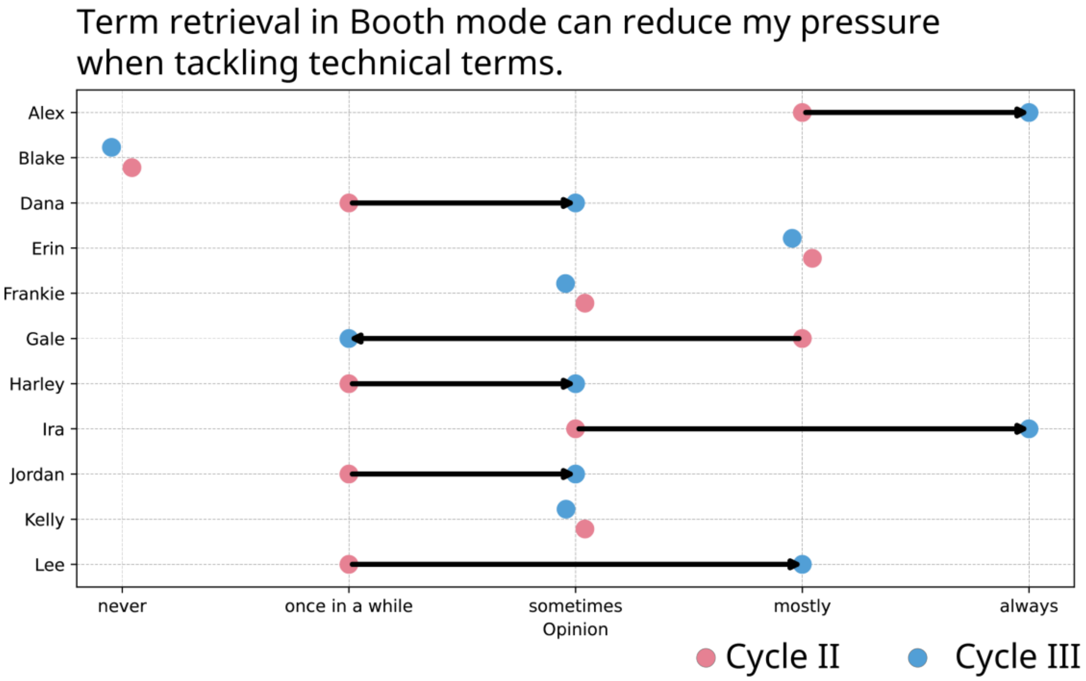


Figure 100. Impact of Booth mode on stress levels when handling technical terms.

Regarding the accuracy of term retrieval in the Booth mode (see **Figure 99**), opinions differed again. Alex consistently thought that the Booth mode could help locate the terms correctly, and Blake was consistent in her opinion that it never does. For instance, Erin and Jordan scaled down from *mostly* in Cycle II to *sometimes* in Cycle III, hinting at diminished confidence in the feature. In contrast, Kelly increased her confidence in it and shifted from *mostly* to *always*, joining Alex. At the group level, Cycle II saw *mostly* as the prevalent choice, comprising 54% of the responses. In Cycle III, the distribution balanced out, with *mostly* and *sometimes* each accounting for 36%. That is, although there was a tendency to trust the Booth mode, it became somewhat more tempered.

As for reducing pressure (again, as self-reported), **Figure 100** shows that Blake consistently thought the Booth mode does *never* reduce pressure, and Erin was as consistent in thinking it *mostly* does. The figure also displays intriguing individual trends. Lee and Ira improved their opinions in Cycle III and Gale was the only one who seemed to think that pressure was only reduced *once in a while.* The bar graph illustrates that *once in a while* was the most popular choice in Cycle II, making up 36% of the responses, and that, after Cycle III, the *sometimes* category surged to 45%. In other words, informants generally found the Booth mode useful at reducing pressure at least in some cases, with more informants moderately relying on it in Cycle III.

3.5.2 Survey for the Excel group

In this section, we present the results from a questionnaire filled out by nine Excel formants, which reflects the informants' overall opinion, glossary tasks, and booth tasks of this study. These aspects are similar to the InterpretBank group, as a result of accumulated opinion from Cycle I to Cycle III. The questionnaire consists of several questions or statements related to each aspect, allowing the participants to express their views and experiences.

3.5.2.1 *Overall opinion* The informants' self-assessment of their interpreting performance collected notably diverse responses (**Table 18**). Similar to the survey in the InterpretBank group, we set the rating scale from 1 to 10, with 1 being the worst and 10 being the highest. The most frequent ones were 4 and 7, each with two informants. All the remaining options had one informant each. This distribution suggests some prominence for the opinions of average and excellent performance, but opinions are spread across the poles.

questions	responses
How do you feel about your interpreting performance?	2/9 responses: 7 2/9 responses: 4 each response for 1, 2, 3, 5, 6
Do you think that terminology management service will improve your term preparation efficiency?	9/9 responses for yes
Do you think that you will benefit from advanced functions of CAI tools (e.g., speech recognition, AI translation)?	8/9 responses for yes, 1 for no

Table 18. Overall opinion questions for Excel informants.

Regarding the second question of overall opinion, *Do you think that terminology management service will improve your term preparation efficiency?* The Excel informants overwhelmingly agreed that terminology management services would improve their term preparation efficiency. A substantial majority of the Excel informants (8) believed they would benefit from advanced functions of CAI tools, and only one did not trust they would bring about potential benefits (**Table 18**).

3.5.2.2 *Glossary task*

This section established four statements to investigate Excel informants' opinions on their experience with glossary tasks, categorizing responses into five options: totally disagree, disagree, not sure, agree, and totally agree. **Table 19** presents the number of informants responding to each statement. The first statement, *I prefer using applications on my phone to a PC for term retrieval,* revealed informants' preference for mobile apps over PCs for term retrieval. In this category, both *disagree* and *totally disagree* accounted for 3/9 of the responses each, indicating that a significant portion of informants were not in favor of using mobile applications for this purpose. On the other hand, *agree* and *totally agree* each represented 1/9 of the total responses, revealing a smaller but existent group who favored mobile applications. The *not sure* category included 1 informant, reflecting some indecision.

In the second statement, *while locating the term's translation, I would check its pronunciation,* most informants (6) *agreed* that they would check the pronunciation of a term while locating its translation. Another one *totally agreed* with this statement, bringing the total count of informants agreeing with the statement to seven out of nine. This indicated a prevalent practice of checking pronunciation among the Chinese respondents. Just one *disagreed* and another one was *not sure,* so a lesser group might not consider pronunciation crucial.

statements	totally disagree	disagree	not sure	agree	totally agree
I like to use applications on my phone instead of a PC for term retrieval.	3	3	1	1	1
While locating the term's translation, I would check its pronunciation.	0	1	1	6	1
I did not verify the accuracy of the translation solution given by web resources (e.g., online dictionaries, term bank)	0	5	3	1	0
I would rely on automatic term extraction from texts rather than human selection.	1	5	1	2	0

Table 19. Responses to statements related to glossary tasks.

The third statement, *I did not verify the accuracy of the translation solution given by web resources (e.g., online dictionaries, term banks)*, reflects informants' practices as in verifying translation solutions from web resources such as online dictionaries. Most of them *disagreed* that they did *not* verify their accuracy, suggesting that they indeed took steps to ensure the *correctness* of the translations. Three more were *not sure*, and only one *agreed*, so a small group might not consider verification.

In the fourth statement, *I would rely on automatic term extraction from texts rather than human selection*, the responses showed a leaning toward *disagree*, which captured more than half the responses. This suggests that the majority of informants did not prefer to rely on automatic term extraction over human selection. On the opposite end, two out of nine informants *agreed* with the statement, indicating a smaller group who favored automation. One informant each selected the *totally disagree* and *not sure* categories, adding layers of nuance to the overall perspective on this topic.

In sum, a majority seemed to prefer PCs over mobile applications for term retrieval, almost everyone considered pronunciation important when looking up translations. There was a strong inclination toward verifying the accuracy of translations from web resources, but opinions were divided on the preference for automatic term extraction over human selection.

3.5.2.3 *Booth task* This section presented four statements to examine the opinions of Excel informants regarding their experience with booth tasks (see **Table 20**). Responses were categorized into five frequency options: *never, once in a while, sometimes, mostly, always.* The first statement, *in a SI task, memorizing the term and its translation is more useful than term retrieval*, portrayed the informants' preferences regarding memorization over term retrieval in SI tasks. The data from Cycle III indicated a preference for *mostly* using memorization, capturing 5 responses out of 9 informants, while 2/9 chose *always* emphasizing a strong leaning toward memorization. Another 2 out of 9 selected *sometimes*, suggesting a more balanced approach between memorization and term retrieval. None of the informants opted for *never* or *once in a while*, indicating that memorization played a vital role in their SI tasks.

statements	never	once in a while	sometimes	mostly	always
In a SI task, memorizing the term and its translation is more useful than term retrieval.	0	0	2	5	2
I'd look for a suitable computer-assisted interpreting (CAI) tool for SI tasks	2	3	1	3	0
I do not use the existing and shared glossary instead of creating my own.	1	2	4	2	0
CAI training can influence my selection of CAI tools.	1	1	5	2	0

Table 20. Responses to statements related to booth tasks.

In the responses to the second statement, *I'd look for a suitable computer-assisted interpreting (CAI) tool for SI tasks*, the informants' inclination to search for a suitable CAI tool for SI tasks. The data showed a fairly even spread across the responses, with *mostly* and *once in a while* each capturing three informants of the total. Two informants chose *never*, and the remaining one opted for *sometimes*.

The third statement was reversed to capture whether the respondents were attentive in their answers. *I do not use the existing and shared glossary instead creating my own, which* revealed a strong leaning toward *sometimes* using existing and shared glossaries, accounting for of the responses. The *once in a while* and *mostly* options each garnered two out of nine informants, suggesting a moderate level of usage for both. Only one responded with *never*, suggesting that most informants were open to using shared resources to some extent.

The last one, *CAI training can influence my selection of CAI tools*, dealt with the impact of CAI training on the selection of CAI tools. A majority of five out of nice informants responded with *sometimes,* indicating that training had a moderate impact on their tool selection. Both *mostly* and *never* were selected by one out of nine informants in each of the responses, highlighting divergent views on the significance of training. Additionally, *once in a while* was chosen by another informant. Memorization was highly valued over term retrieval in SI tasks, while opinions on the importance of CAI tool selection and training varied among the informants from the Excel group. Most were open to using existing and shared glossaries, albeit to varying degrees.

3.6 Holistic quality assessment of informants' audios

As reported in **§ 2.5.5**, five PhD raters holistically assessed the quality of the informants' booth renderings, based on their intuitive impressions. We will now summarize the results regarding inter-rater reliability and the correlation between rating scores with indicators for fluency and accuracy. A single global assessment score is analyzed, rather than rubric-based or analytic ratings where each aspect or criterion is rated separately. Holistic rating methods can be faster but may be less reliable, due to their subjective nature (see, however, Waddington, 2001 for translation). This can ultimately result in lower inter-rater reliability. Krippendorff's alpha was calculated to be 0.016 for all five raters and 0.166 for sets of three raters, so the inter-rater reliability was really low.

3.6.1 Individual rater analysis

As a reminder, we chose 5 audio recordings each from Cycles I, II, and III, totaling 15 audios. These audios were evaluated by 5 raters for testing inter-rater reliability, resulting in each audio receiving 5 ratings (referred to *five raters group* in **Table 21**). The remaining audios were each assessed by 3 evaluators (referred to *three raters group* in **Table 21**). Further details on the selection and randomization procedures are available in **§ 2.5.5**. Also, #1 was the highest score, and #6, was the lowest one. In the analysis of the evaluation scores across two distinct groups—the *five raters group* and the *three raters group* in **Table 21**—it became evident that Luc consistently demonstrated a stringent evaluation style (see **§**

2.5.5.3). For the *five raters group*, Luc emerged as the most rigorous evaluator, recording an average score of 4.13, which leaned toward the *Poor* and *Bad* categories on the rating scale. This pattern persisted in the *three raters group*, where Luc again stood out as the strictest, posting an even higher average score of 4.23.

	Félix	**Jules**	**Luc**	**Maxime**	**Quentin**
five raters group	3.27	3.60	4.13	3.47	3.60
three raters group	2.81	3.90	4.23	3.55	3.33

Table 21. Average evaluation scores by raters and groups.

Correlation matrices of fluency and accuracy-related variables were cross-referenced with individual rater's scores per cycle (see **Table 22**). Raters were not informed of the grouping and order of speeches. Cycles and groups are not considered factors when holistically assessing recordings, but the values of fluency and accuracy-related variables depend on the source speeches in each cycle. By focusing on cycles rather than overall values, we aim to lower distortions in fluency and accuracy-related variables that influence the rating process. This strategy facilitates a deeper understanding of the underlying patterns that might be guiding the raters' evaluations, paving the way for insights grounded on individualized rater responses, examined cycle-wise. Informants' individual rating scores can be found in **Appendix J**. Since these variables are not normally distributed (see **§ 3.2** and **§ 3.3**), Kendall's Tau coefficient was used as the 'superior non-parametric measure' Mellinger & Hanson (2017, p. 191).

		Cycle	**I**		**II**		**III**	
raters		**indicators**	**Tau b**	p-value	**Tau b**	p-value	**Tau b**	p-value
Félix	**fluency**	*false start*	0.04	0.86	0.01	0.95	−0.07	0.73
		self-correction [b]	0.14	0.52	0.19	0.39	−0.25	0.23
		filler	0.68	0.00[a]	0.11	0.61	−0.54	0.01[a]
		repetition	0.05	0.81	−0.26	0.26	−0.12	0.59
		bump	−0.38	0.08	0.29	0.19	0.02	0.92
		respite	−0.06	0.77	0.10	0.65	0.04	0.85
		EVS1	−0.16	0.45	−0.15	0.49	−0.28	0.17
		EVS2	−0.24	0.27	−0.32	0.14	−0.18	0.36
	accuracy	*correct* [c]	−0.09	0.68	0.08	0.73	0.19	0.36
		adequate	0.07	0.75	−0.10	0.68	−0.03	0.91
		wrong	0.11	0.64	0.13	0.56	−0.31	0.14
		skipped	0.04	0.86	−0.14	0.53	0.01	0.96
Jules	**fluency**	*false start*	−0.26	0.24	0.26	0.25	0.03	0.91

		Cycle	I		II		III	
raters		**indicators**	**Tau b**	p-value	**Tau b**	p-value	**Tau b**	p-value
		self-correction	−0.15	0.48	0.03	0.88	0.11	0.61
		filler	0.07	0.75	−0.17	0.43	0.33	0.12
		repetition	0.16	0.50	−0.07	0.75	−0.22	0.33
		bump	−0.08	0.71	0.12	0.57	−0.11	0.61
		respite	−0.39	0.07	−0.14	0.50	−0.26	0.23
		EVS1	0.05	0.83	−0.18	0.40	−0.09	0.69
		EVS2	0.05	0.83	0.04	0.84	0.09	0.69
	accuracy	*correct*	−0.20	0.36	−0.03	0.88	−0.25	0.25
		adequate	−0.54	0.02 [a]	0.04	0.87	0.21	0.38
		wrong	0.27	0.23	0.24	0.27	0.17	0.45
		skipped	0.27	0.22	0.00	1.00	0.09	0.69
Luc	**fluency**	*false start*	0.23	0.30	−0.22	0.32	0.14	0.52
		self-correction	0.27	0.21	−0.17	0.40	−0.02	0.91
		filler	−0.08	0.71	−0.08	0.72	−0.06	0.79
		repetition	0.14	0.54	−0.22	0.31	−0.06	0.79
		bump	−0.01	0.96	−0.14	0.50	−0.25	0.24
		respite	0.08	0.71	0.08	0.72	−0.59	0.01[a]
		EVS1	0.33	0.12	−0.11	0.60	0.07	0.75
		EVS2	0.06	0.79	−0.17	0.41	−0.07	0.75
	accuracy	*correct*	0.12	0.59	0.26	0.22	−0.28	0.20
		adequate	0.12	0.61	0.18	0.42	−0.57	0.02[a]
		wrong	0.19	0.41	0.00	1.00	0.20	0.36
		skipped	−0.12	0.59	−0.29	0.17	0.29	0.19
Maxime	**fluency**	*false start*	0.20	0.36	0.12	0.60	0.27	0.22
		self-correction	0.24	0.28	0.15	0.52	0.51	0.02[a]
		filler	−0.08	0.71	−0.29	0.20	0.16	0.45
		repetition	0.10	0.65	−0.11	0.64	0.28	0.22
		bump	0.34	0.11	0.46	0.04[a]	−0.09	0.69
		respite	0.35	0.10	0.10	0.65	−0.24	0.27
		EVS1	0.03	0.87	−0.16	0.48	−0.38	0.08
		EVS2	0.19	0.36	0.01	0.95	−0.43	0.05[a]
	accuracy	*correct*	0.22	0.33	0.39	0.09	−0.34	0.13
		adequate	0.32	0.16	−0.22	0.35	−0.32	0.19
		wrong	0.11	0.62	0.16	0.51	0.05	0.81
		skipped	−0.23	0.28	−0.46	0.04[a]	0.43	0.06
Quentin	**fluency**	*false start*	0.38	0.10	−0.23	0.30	0.19	0.41

	Cycle	**I**		**II**		**III**	
raters	**indicators**	**Tau b**	p-value	**Tau b**	p-value	**Tau b**	p-value
	self-correction	0.08	0.72	−0.15	0.49	0.35	0.11
	filler	0.22	0.32	−0.28	0.19	−0.04	0.86
	repetition	0.15	0.51	−0.20	0.38	0.25	0.25
	bump	0.13	0.56	0.00	1.00	−0.21	0.33
	respite	0.27	0.22	−0.01	0.95	−0.15	0.49
	EVS1	0.07	0.77	−0.11	0.61	−0.09	0.69
	EVS2	0.35	0.11	−0.16	0.46	−0.21	0.33
accuracy	*correct*	0.34	0.14	0.37	0.09	−0.08	0.73
	adequate	−0.19	0.42	−0.07	0.77	0.00	1.00
	wrong	−0.17	0.47	0.18	0.42	−0.06	0.77
	skipped	−0.12	0.59	−0.48	0.03[a]	0.14	0.53

[a] statistically significant correlation

[b] *self-correction* is an indicator within the fluency dimension, representing the correction of the renditions.

[c] *correct* is an indicator within the accuracy dimension, indicating the correct renditions.

Table 22. Correlation between fluency and accuracy indicators among five raters.

raters	**Cycle I**	**Cycle II**	**Cycle III**
Félix	0.98	0.61	0.73
Jules	0.81	0.43	0.59
Luc	0.45	0.55	0.88
Maxime	0.58	0.92	0.94
Quentin	0.57	0.85	0.56

Table 23. Raters’ Taub variation (differences between Max and Min values).

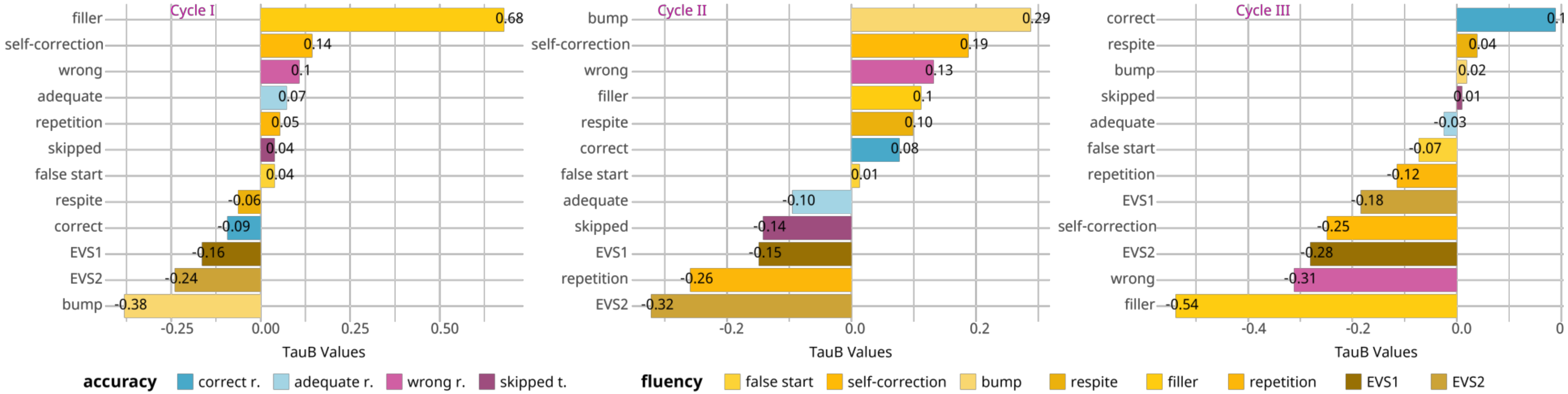


Figure 101. Tau-b values for Félix in Cycles I, II, and III.

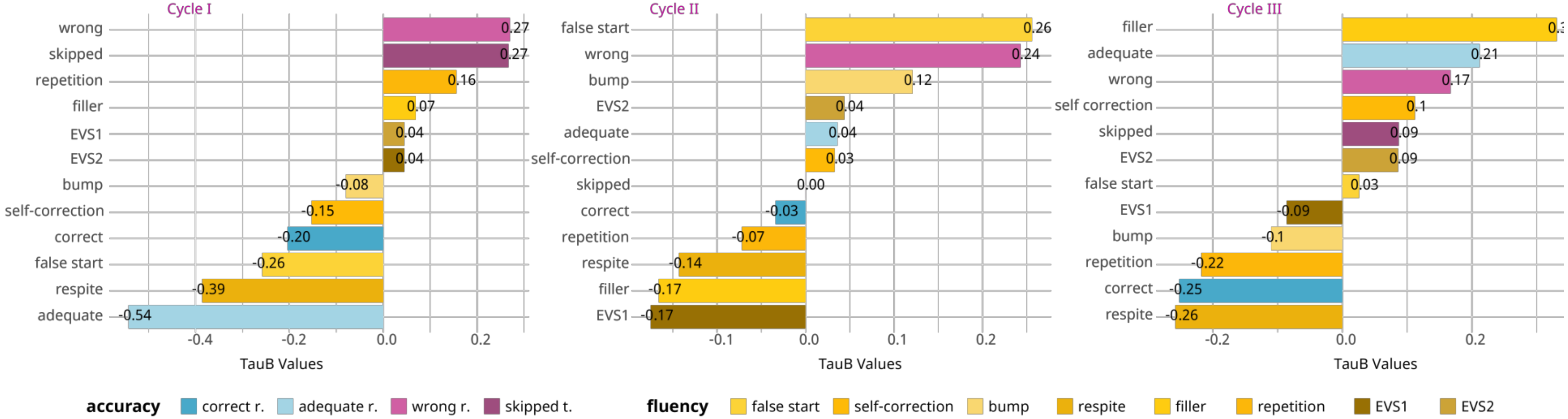


Figure 102. Tau-b values for Jules in Cycles I, II, and III.

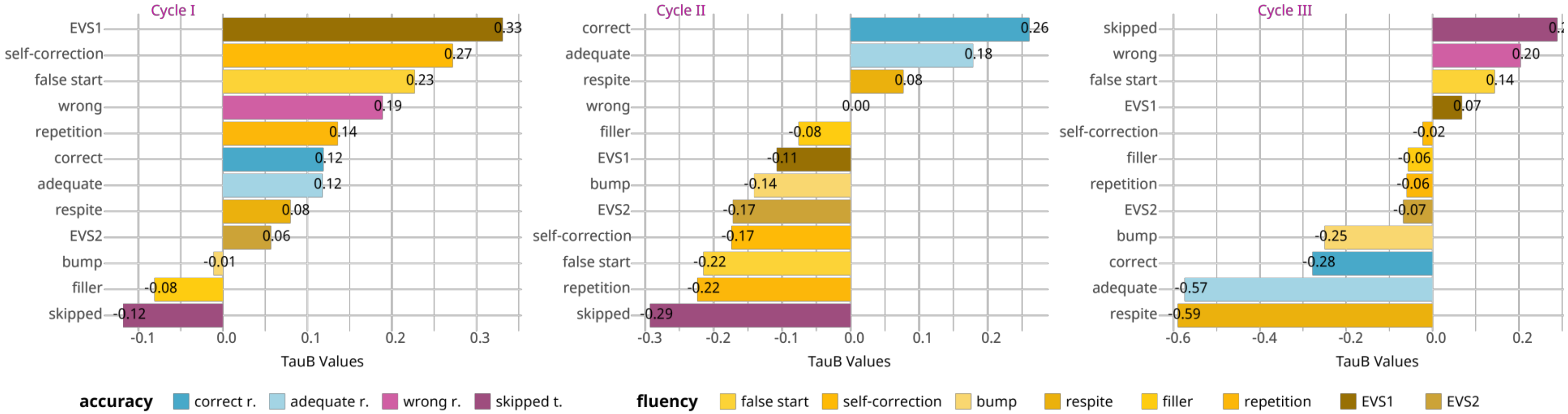


Figure 103. Tau-b values for Luc Cycles I, II, and III.

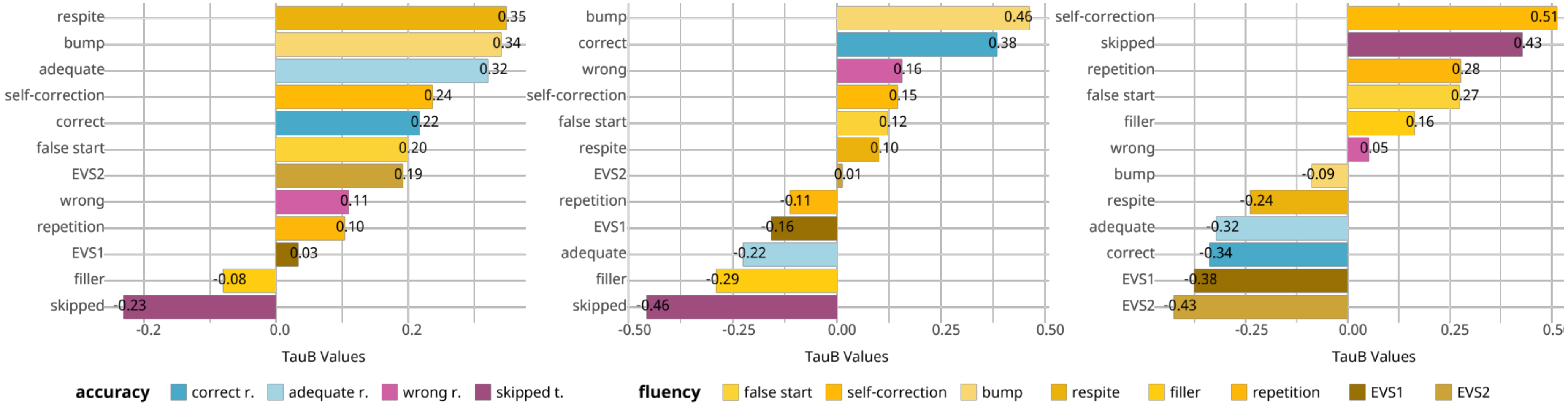


Figure 104. Tau-b values for Maxime in Cycles I, II, and III.

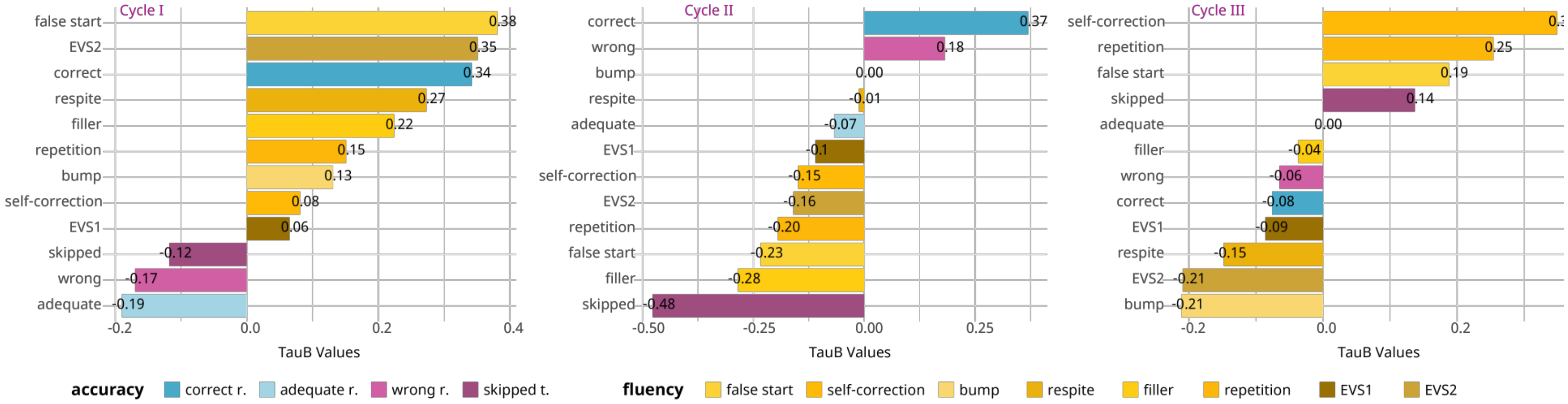


Figure 105. Tau-b values for Quentin in Cycles I, II, and III.

Data from Félix, shown in **Figure 101**, reveal within fluency the variable *filler* had a Tau-b value of 0.68 *(*p-value < 0.05) in Cycle I. This suggests a strong positive correlation between the use of *filler*s and perceived fluency in this rater. By Cycle III, however, the Tau-b for *filler*s plummeted to -0.54, (p-value < 0.01), that is, it changed to a strong negative correlation. This variation may imply that the rater continually adjusted his criteria for evaluating the role of fillers when confronted with different audios. *Accuracy* and *fluency* are displayed in the graphics but not considered so as not to mislead the reader, for raters did not necessarily consider such categories nor would they probably have the same components of the constructs that we used.

Félix's evaluations seem less decisive for accuracy. For instance, in Cycle III, *correct renditions* had a Tau-b value of 0.189 *(*p-value > 0.05, which is not statistically significant). This could suggest that Félix did not focus on terminology when considering the overall quality of the interpretation.

Jules' assessments (**Figure 102**) show less variation (from -0.54 to 0.27 in Cycle I, -0.17 to 0.26 in Cycle II, and -0.26 to 0.33 in Cycle III) in Tau-b values across cycles, and their p-value*s* often point to a lack of statistical significance. For instance, in Cycle II, the Tau-b value for *respites* is −0.14 (p-value = 0.50). While the negative Tau-b value implies an inverse correlation between *respites* and scores, the high p-value shows that it lacks statistical significance. The variable *adequate,* in Cycle I, displays a statistically significant strong inverse correlation with a Tau-b of −0.54 and a p-value of 0.02. This may reflect Jules's emphasis on adequacy as a critical factor in the overall assessment. As a reminder, the raters performed a holistic assessment where no access to the original was granted, no materials were used, and no rubrics or categories were considered. The correlation refers to a later analysis we performed on their judgments and our quantitative profiling of the renditions, but the quantities of each category are not considered, and their few instances may distort the results.

Luc's data (**Figure 103**) reveals that the variable *respites* in Cycle III had a Tau-b of −0.59 with p-value < 0.05, a strong inverse correlation that is statistically significant. So, Luc might have found longer pauses detrimental to quality, at least in the recordings for this cycle. The variable *adequate* had a Tau-b of −0.57, p-value < 0.05 in Cycle III, pointing again to a strong, inverse, statistically significant correlation.

For Maxime (**Figure 104**) the variable *bumps* in Cycle II had a Tau-b of 0.46 (p-value = 0.04), suggesting a moderate positive correlation that is statistically significant. This should mean that Maxime perceives shorter pauses as beneficial for interpreting quality, particularly in Cycle II. No accuracy variables had a p-value lower than 0.05 in any cycle, so none were statistically significant. However, the indicator *correct renditions* in Cycle II had a Tau-b of 0.39, suggesting a moderate direct correlation, even if not statistically significant. Taken together, the results for Maxime suggest that she might be more attuned to assessing in didactic environments.

Quentin in **Figure 105** had a Tau-b of 0.35 (p-value = 0.11) for the variable *self-correction* in Cycle III. This indicates a moderate direct correlation, suggesting that *self-corrections* during the speech had a positive impact on perceived quality. Nevertheless, the p-value is not under the common alpha level of 0.05. Regarding accuracy, the variable *skipped terms* in Cycle II had a Tau-b of −0.48 (p-value > 0.05), indicating a moderate to strong inverse correlation but is not statistically significant. This may suggest that Quentin regarded *skipped terms* as having a significant negative impact on the quality of the interpretation.

In conclusion, the five raters seem to have been reasonable in their assessments but each one displays unique patterns and the quantitative variables correlating with their holistic quality assessments vary to the point of not having two similar assessment profiles.

3.7 Summary

In this chapter, the study presents results from glossary and booth tasks, and analyses based on screen recordings, keylogging, and transcripts of informants' SI output. The glossary task results are organized into two main aspects: the process of compiling glossaries and the sources consulted. For glossary compilation, ***translation search*** and ***term extraction*** were identified as the most time-consuming subtasks across all cycles. Informants using Excel focused mainly on these subtasks, while those using InterpretBank engaged more diversely. InterpretBank informants also invested more time in various strategies, with search queries significantly influencing their time allocation. Regarding consultation sources, informants adopted various tools, including both local and online applications. In Cycle II, the InterpretBank group employed a more diverse range of online services compared to the Excel group. By Cycle III, both groups predominantly used search engines, with tool usage varying based on individual preferences and task specifics.

In terms of individual glossary compilation, InterpretBank informants generally compiled more terms than those in the Excel group. We will refrain from advancing conjectures about the causes here since that is one of the goals of the next chapter (but see **§ 4.1**). Additionally, a glossary review was conducted before the booth tasks, where all informants reviewed a master glossary compiled from individual ones. Particularly, in Cycle II, engagement with InterpretBank's *Memo* mode varied among informants, with some frequent users and others not using it at all. Cycle III saw shifts in usage patterns, with increased non-users and variations in engagement levels.

For booth tasks, the study presented results for fluency analysis, term accuracy analysis, search behavior for the InterpretBank group, survey analysis, and human holistic assessment. Fluency analysis encompassed various indicators like *false starts, self-corrections, fillers, repetitions, bumps, respites, EVS1,* and *EVS2.* The results showed that the Excel group had more *false starts*, *self-corrections*, *fillers*, and *respites*, than the InterpretBank group in all Cycles. except for fewer *bumps*. A modest but statistically significant negative correlation was found among the duration of source speech, *EVS1,* and *EVS2*, especially indicating that longer speech chunks led to shorter *EVS2*. However, in Cycle III, this correlation persisted only in the InterpretBank group.

Term accuracy analysis from booth tasks revealed no significant differences between the groups in *correct* renditions, *adequate* renditions, *wrong* renditions, and *skipped terms*. However, for both groups, the number of *skipped* terms exceeded the number of *correct* renditions. The InterpretBank group showed improvement from Cycle II to III in *correct* renditions and fewer *skipped* terms. Individual performances varied within each group over time.

As part of the side study, the recall effect study showed that the Excel group improved in repeated term categories *(1st2Rep, rep1, rep2)* across cycles, indicating that *1st2Rep* terms have been interpreted well in the first repetition and second repetition. The InterpretBank group experienced an increase in the correct rates of all three categories in Cycle II but a decline in Cycle III.

The search features in the InterpretBank group have been examined. Firstly, there was a decrease in the frequency of InterpretBank usage from Cycle II to Cycle III. In Cycle II, informants performed 242 searches with a 55.37% accuracy rate for correct searches with correct renditions. By Cycle III, the number of searches dropped to 176, yet the accuracy for correct searches with correct output improved to 65.34%. Secondly, the statistical analysis using Kendall's Tau-b values and p-values indicated a statistically significant but weak inverse correlation in Cycle II, which became slightly stronger (though still weak to moderate) in Cycle III. In other words, as the *ear-to-key* or *E2K* increased, the *eye-to-voice* or *I2V* tended to decrease, yet modestly. The third observation concerns the use of InterpretBank for repeated term queries, which was not particularly high in both cycles, with 5 out of 66 in Cycle II and 8 out of 66 in Cycle III.

Finally, we also examine the relationship between *search events* and *dropped chunks* among InterpretBank informants. The number of overlaps between these two events decreased in Cycle III compared to Cycle II. This may suggest a change in how informants managed *search events* and sentence delivery over time. Additionally, there was a noticeable variation in individual performance, with some informants dropping more sentences than they searched for, while others showed the opposite pattern.

In the survey, InterpretBank informants found the tool useful for timesaving and pressure reduction, with mixed opinions on its automatic term extraction feature. The Excel group believed that terminology management services would enhance their efficiency, preferring PCs over mobile apps for term retrieval and valuing memorization over term retrieval in SI tasks.

As expected, due to the low inter-rater reliability, the raters had different assessment styles: Luc was the most stringent, Jules's assessments showed less variation, and Quentin preferred self-corrections. This diversity in assessment styles reflects evolving individual characteristics. Félix's evaluations were more positive for fillers in Cycle I but more negative in Cycle III. They were less decisive for accuracy. These findings highlight the variability in rater assessments and the different quantitative variables that correlate with their holistic quality assessments.

Chapter 4

discussion

Complex results are difficult to organize. There are different alternative options, and you cannot win in the sense that no one order will be preferred by all readers. This chapter first discusses the hypotheses laid out in this project (**§ 4.1** to **§ 4.5**). Then, it steps back to take a broader view and consider the overall behavior of all participants when preparing their glossaries. Their behaviors are interpreted as indicative of likely cognitive processes involved. Section **§ 4.6** focuses on the detailed individual behaviors performance by the Excel group and InterpretBank group members when building their glossaries. Section **§ 4.7** discusses searches with InterpretBank (hence, it is restricted to the informants in the InterpretBank group) observed in Cycles II and III. Finally, the discussion shifts back to a sample level, focusing on the correlation between source speech chunks and EVS in the booth tasks in all informants (both groups). This is followed by the findings of the exploratory study on holistic assessment in the present study. Let us then get started, and do so by reminding of the hypotheses:

H1: InterpretBank improves efficiency at glossary compilation

H2: InterpretBank improves the quality of RSI rendering

H3: InterpretBank improves efficiency when producing the RSI rendering

H4: Improved documentation performance will yield better RSI rendering quality (H3)

H5: Improvements using InterpretBank but also attitudes, will lead to keeping using it

4.1 H1: InterpretBank improves efficiency at glossary compilation

Term-intensive RSI tasks are quite usual in the market and thus glossary compilation is a typical aspect of an interpreter’s workflow. Domain-specific knowledge becomes a critical component of glossary compilation, but it is the market demands, rather than the interpreters’ interests, that end up specializing professionals in different fields. Interpreting trainees tend to lack domain-specific knowledge, simply because of their age (Kurz *et al.*, 2011; M. Liu *et al.*, 2020; Chiocchetti *et al.*, 2023; Elmer & Giroud, 2023). This is probably one of the reasons why interpreting trainees often seem to find glossary compilation time-consuming and the outcome of their efforts uncertain. As a consequence, they tend to feel it is a heavy burden when not a drawback in their professional development. Hence, CAI tools are raising expectations and have in interpreting trainees more or less enthusiastic supporters who hope to benefit from their use, improving or at least alleviating term extraction.

The first hypothesis explores the impact of InterpretBank on term extraction efficiency. This study used three indicators: the time taken in glossary compilation tasks, term counts, time per term, and rate of diversity of extracted terms from individual glossaries between the two groups in Cycles II and III. The percentage of individual time spent on each subtask between groups has been shown in **Figures 33–34**, emphasizing two subtasks that consumed a significant portion of time: ***translation search*** and ***term extraction*** in Cycles II and III. In the former, a significant proportion of time was dedicated to *search queries* by both the InterpretBank and Excel informants. In the latter, *reading-first term selection* consumed a larger share of time, predominantly by Excel informants. Comparing the performance and quality variations among these resources is challenging. Furthermore, individuals may combine information from several resources.

Table 24 summarizes the main data. The InterpretBank informants tended to need relatively less time, for they used 95.72% of the time Excel informants needed in Cycle II and 90.16% in Cycle III. This is good and there are reasons to believe that with longer, more thorough training these figures would probably improve, but it is not an impressive result, definitely not one that would by itself justify investing time and money in a CAI tool. The rapid automatic term-extraction process by InterpretBank's in-built algorithm seems to have exerted some influence in glossary preparation, although the difference between the two groups was not substantial in terms of overall time taken.

Cycle	**II**		**III**	
item group	XL	IB	XL	IB
time taken	4730.1	4527.9	4554.9	4107.0
term counts	44.8	78.8	73.25	124.75
time per term	115.1	68.9	63.0	41.5
term diversity	45.30	27.66	46.43	27.76

Table 24. Glossary compilation indicators for two groups in Cycles II and III.

As **Table 24** shows, glossaries compiled with InterpretBank support had more terms, 175% in Cycle II and 170.31% in Cycle III. In principle, having more terms is not necessarily better. It very much depends on the general and domain knowledge the user has and her L1 and L2 commands. For inferential statistics about the possible impact of this data on booth task behaviors, see **§ 4.2**. Nevertheless, the difference in term counts between the Excel and the InterpretBank groups was so large that automatic extraction seems very likely to have contributed a substantial share of terms in the individual glossaries. Faced with a long list of terms, some InterpretBank informants seemed to trust the application's choices and simply accepted the term lists. We need to explore the share of terms selected by InterpretBank that have been accepted but we already see a tendency for InterpretBank's choices to increase total term counts in the individual glossaries.

Furthermore, the difference in how terms are accessed is relevant. Large Excel glossaries may lead to burdensome visual searches, with terms exceeding the screen and forcing the user to scroll up and down. In contrast, keyboard-based string searches may be less distractive at the beginning of a search, precisely because they do not display all entries on the screen. Under this scope, users with less knowledge might afford to compile larger glossaries. Even if larger glossaries are not necessarily more supportive later on in the booth, they might have a reassuring or unstressing effect. Defrancq & Fantinuoli (2021) discussed potential positive psychological effects with reference to ASR support along these lines. Here, our view is supported by survey responses from Excel informants; 5 out of 9 indicated they needed to verify the candidate renditions they found, and 6 out of 9 reported checking pronunciation after locating a term's translation. These practices in Excel glossaries—verification of translation, pronunciation checks, and switching between various information resources—contribute to additional time investment in glossary management.

The role of *reading-first term selection* when using InterpretBank should not be underestimated. Despite the automatic procedure, there is still a sizable need for human intervention to compensate for the limitations inherent to *automatic term extraction* and translation suggestions. We can assume that most of the time the InterpretBank informants devoted to glossary compilation was focused on tasks such as editing entries (e.g., modifying translations or searching for translations, and deleting irrelevant entries).

Failure to meet users' expectations may also be attributed to insufficient consideration on the part of the CAI tool developers regarding InterpretBank functionalities, specifically in relation to languages distant from the *habitual suspects* (see **§ 5.1**). For instance, InterpretBank offers translations that may be either in simplified Chinese characters or traditional Chinese characters, but restricting results to one of these options for all terms in a glossary is not possible. Hence, Chinese users often need to unify suggested translations after automatic extraction and translation.

Consequently, our informants using InterpretBank devoted more attention toward subtasks like ***translation search***, and ***entry editing***, as the glossary generated through *automatic term extraction* possibly did not meet their expectations. In any case, the combination of slightly shorter compilation times with far larger numbers of entries leads to the informant needing far less time per term—only 59.85% of the average time Excel informants needed to spot, document, and enter a term in Cycle II, and 65.91% in Cycle III (see **Table 24**).

The average *time per term* suggests that the InterpretBank informants had, on average, less duration per single term than the Excel group. *Time per term* reveals that the Excel informants, broadly speaking, devoted more time to each term, suggesting a tendency toward allocating more cognitive resources to novel information. **Table 5** shows that most InterpretBank informants used the automatic extraction feature before they even read the text. As a result, when they confronted the glossary, they were faced with a list of decontextualized terms. We can only speculate what their behavioral tendencies were, for we did not perform a detailed analysis of the dynamics of glossary review. Some informants might have devoted less time per term, for instance, because they blindly trusted InterpretBank or simply because they did not care much or did not know about the topic anyway.

However, other options are possible. In one of them, informants might devote as much time to terms as they deemed relevant but nearly none to terms, they thought superfluous. In contrast, Excel informants would be quite sure of the personal relevance of all entries in their individual glossaries. Some informants might have started devoting a comparable amount of time to terms, then realized they were running out of time and simply stopped and left the remaining term unchecked. Yet in another scenario, InterpretBank informants might have devoted increasingly shorter times to terms as they became increasingly convinced that the automatic extraction feature worked well for them. A couple of informants did indeed run the feature twice, indicating that they may have wanted to retrieve again terms they wrongfully discarded after the first round. The fact remains that, on average (we already pointed out that averages may hide the really interesting information when researching mental processes of a task that unfolds in time) InterpretBank informants devoted far shorter times to review each term than Excel informants did.

The last indicator is the diversity rate of terms in individual glossaries. That is, how much each informant differs from other informants within the same group. Individual glossaries contained duplicate terms or term variations, such as capitalized and plural forms, so the counts were performed on lemmas, or base forms of each word, disregarding any variations or affixes. A *lower* diversity rate indicates a *higher coincidence rate*. The lower the diversity rate, the more similar glossaries within a group are. The InterpretBank group has a lower diversity rate, so they coincide more with each other than Excel informants do. This was expected because informants are prompted by InterpretBank to choose from a machine-extracted, identical list of terms, some of which they might have otherwise overlooked or ignored.

The automatic procedure demands time for informants to process novel terms, reorganize them, and store them in InterpretBank. When using the application's term-extraction feature, some other tasks might even take longer, because users do not have the contextualized reading to narrow down their possible meanings. This may partially explain why InterpretBank informants still devoted considerable time to reviewing the machine-extracted terms. The informants were new users, and they began using the application right for this study and did not rely solely on it. Hence, their own prior skills in compiling glossaries played a crucial role. The only way to access the meaning of terms is by translating them outside InterpretBank. The only way to retrieve the co-text is to look at the full document (e.g., in a Word file). Given the added difficulties and considering that the original text has already been processed to extract the terms, CAI tool developers should carefully consider including a concordance feature for users to retrieve relevant source text co-texts easily.

In any case, these findings support the first hypothesis: Informants using InterpretBank to compile their glossaries for RSI displayed improved documentation behaviors, specifically in terms of term extraction efficiency *(term counts, time per term).* They spent less time on glossary compilation and included more terms in their glossaries. Future data analysis should explore how many selected terms survived from the *automatic term extraction* and human manual extraction and how many of those were actually used.

4.2 H2: InterpretBank improves the quality of RSI rendering

CAI tool use is uninteresting if it does not offer advantages or, at least, an acceptable trade-off. Either using a CAI tool improves quality or maintains it or even lowers it but offers other advantages. Discerning quality reliably is thus a core component of studies on CAI tool introduction and use. RSI quality is important for specialized audiences to receive technical information, particularly in term-intensive speeches. There are several indicators to measure RSI quality (e.g., Collados 2016; Su 2019; Chen *et al.* 2022; Han 2022a), but no standard yet. We chose not to impose a standardized rubric on the raters, who were unaware of groups and Cycles in the randomized recordings they assessed. Differences between source speeches may have led them to unconsciously group them, but they had no further clue, and they were not allowed to compare renderings. They were not handed the source speeches either, nor any materials related to them (such as scripts or transcripts). This approach did not impose any norms or third-party criteria on raters, but it also deprived them of contextual cues they might use in real-world evaluations, like knowing the conference topic beforehand. Our approach is closer than other assessment strategies to the ways unprofessional addresses tend to assess the quality of interpreters' renderings, but there is still room for improvement.

Cycle	**I**		**II**		**III**	
groups	avg.	%	avg.	%	avg.	%
Excel	4.333	27.78	3.733	37.78	3.653	39.11
InterpretBank	3.778	37.03	2.972	50.46	3.139	47.68

Table 25. Average quality ratings for the recordings.

As a reminder, most renderings were assessed by random subsets of three out of five raters to keep the volume of work manageable for these volunteers, and some recordings were assessed by all five to check inter-rater reliability. Raters categorized each RSI rendering quality according to a scale from 1 to 6, where 1 was the best and 6 was the worst category. **Table 25** displays average assessments per group and Cycle, both raw and converted into percentages, where 100 is the best and 0 is the worst, to make it more intuitive. The InterpretBank group received better evaluations (closer to 1) than the Excel group: For example, in Cycle II, the Excel group's score amounted to 37.78% and the InterpretBank, 50.46%.

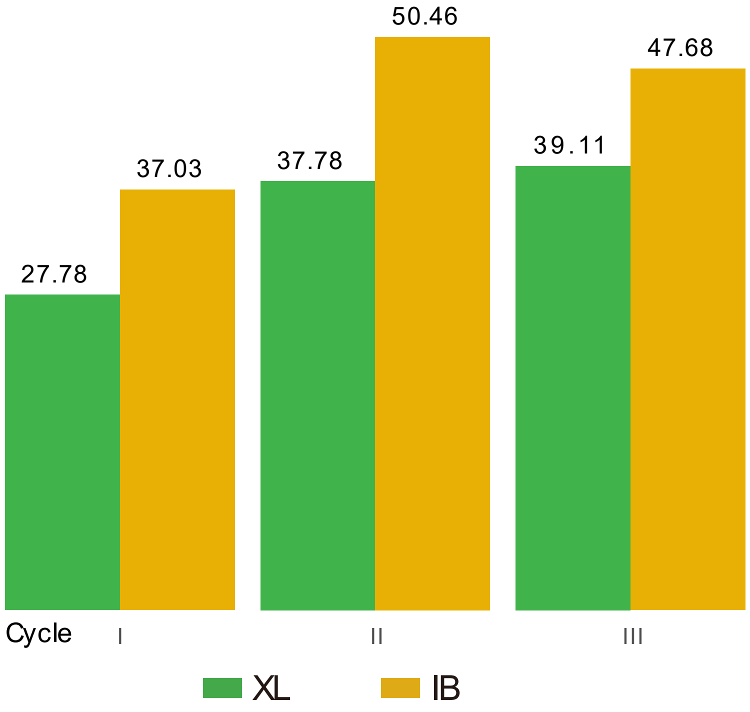


Figure 106. Average percentage of quality ratings for two groups.

Figure 106 shows a waning difference between groups of 9.25% in Cycle I, 12.68% in Cycle II, and just 8.57% in Cycle III. In view of the small sample, the expected fluctuations for uncontrolled variables—such as the day of the week when they performed the task or differences in potential cognitive demands posed by each source speech—we may consider the InterpretBank group's average score, actually more or less constant or flat. The advantages of becoming more used to using InterpretBank at the booth might have been offset by the hypothetical tendency to rely on the glossary functions and not keep the terms active in their memories. Since memory traces such as acoustic representations might be kept, ensuring that the terms are recognized later on, this might lead to mistaken recall and thus higher numbers of wrong renditions. Other indicators may also have taken their toll in the assessments, even though the inter-rater agreement was very low (see **§ 3.6**).

The results for the Excel group, however, are more difficult to explain, for both groups had to adapt to task demands. The difference between Cycles I and II is 7.19%, while the difference between the groups in Cycle II is 18.93%. The differences between Cycles II and III for the Excel group (8.40%) are about the same as the difference between the groups in Cycle III (8.79%). If this was a trend, even with a leeway of a couple of points (as we assumed when suggested that the

InterpretBank's performance was flat), then in a hypothetical Cycle IV the Excel informants might have matched the performance quality of the InterpretBank informants. This is sheer speculation, but we already had other hints to surmise that a future research project might target more than three data collection Cycles, and this insight makes a longitudinal study with more data points nearly a must. In any case, assessment data supports hypothesis 2: Using InterpretBank has a positive impact on RSI rendering quality.

4.3 H3: InterpretBank improves efficiency when producing the RSI rendering

The third hypothesis predicted that InterpretBank would contribute to enhancing the efficiency when producing RSI rendering. As a reminder, this was operationalized and measured through indicators of fluency and accuracy. For instance, the use of a placeholder like a *filler* is "motivated by constraints in cognitive processes, such as difficulty in remembering or "accessing" an appropriate lexical item" (Hayashi & Yoon, 2010, p. 42). *Bumps* (time gaps of 200−600 ms) and specially *respites* (gaps above 600 ms) might hint at the informants' approaching their cognitive limits when they allocate cognitive resources for reorganizing or formulating ad-hoc solutions, or searches in their WM (Keevallik, 2010). Facing term-intensive RSI tasks, interpreting trainees might have tried to cope with demands with strategies such as consulting their (revised) master glossaries either in MS Excel or through InterpretBank's string search feature.

4.3.1 Fluency: Content cluster

Fluency indicators were clustered into two groups for analysis: a *content* cluster, and a time cluster. The *content cluster* comprised *false starts*, *fillers*, *self-corrections, repetitions* as indicators. They are assumed to hint at cognitive efforts, although their variations may also hint at cognitive demands (e.g., Plevoets & Defrancq, 2018; Han & Yang, 2023 for fluency indicators in interpreting quality assessment). The groups exhibited differences in the average number of *false starts*, with variations observed across the Cycles. Specifically, the Excel group demonstrated fluctuations, decreasing in Cycle II, and increasing in Cycle III, in the average number of *false starts*. Furthermore, in each cycle, the Excel group consistently had a higher average of *false starts* compared to the InterpretBank group. In contrast, the InterpretBank group displayed the same pattern of changes in their average number of *false starts* across the Cycles, compared to the Excel group. Importantly, the adoption of InterpretBank did not result in a significant qualitative change in performance, as evidenced by both within-group and between-group comparisons. These observations are supported by the Friedman test for within-group analysis and the Mann-Whitney U Test for between-group analysis. Hence, the data on *false starts* does not support the hypothesis that using InterpretBank contributes to improved RSI booth performance efficiency over time.

In the results summarized in **§ 3.2.1.2**, the InterpretBank informants display a more varied pattern of *self-corrections*, compared to the Excel group. In Cycle II, the InterpretBank informants increased their aggregated *self-corrections*, which then decreased in Cycle III. One plausible explanation for this may be the learning curve associated with adapting to InterpretBank in Cycle II, so the decrease in Cycle III would hint at the group's increasingly proficient use of InterpretBank. Engaging with a specialized CAI tool like InterpretBank might introduce an additional cognitive strain on the interpreters, especially if they are engaged in the term-intense RSI tasks and in their initial phases of adapting to the tool. About half of the InterpretBank informants self-reported that they had had no exposure to InterpretBank before this research project. The other half had been only modestly exposed. The novelty might have impacted their performance, possibly leading to more *self-corrections* for some informants at SI.

However, both groups exhibit a similar pattern of change from Cycle I to Cycle II, and from Cycle II to Cycle III, with slight variations in the average counts of self-corrections. Perhaps the growing familiarity with the tool eases term retrieval while interpreting, or perhaps the added tool simply becomes less disruptive with practice. An alternative interpretation is the possible influence of individual cognitive engagement. Whichever the reason is, it leads to fewer errors and consequently fewer *self-corrections*, but it cannot be concluded that the choice of tool leads to a reduction in the frequency of self-corrections. Hence, data on *self-corrections* does not support the hypothesis that using InterpretBank may contribute to improving RSI booth performance efficiency over time. Samples 1 and 2 from our informants demonstrate that *self-corrections* are not limited solely to corrections of terms in the master glossary.

sample 1 (self-correction for verb):
2494.501 s—2496.395 s 啊，比如说我们平时*摄出*
2496.955 s—2498.519 s *摄入*的益生菌
(literal translation: em, for example, the probiotics we usually ***take out***, ***take in***.)

sample 2 (self-correction for adv):
2584.462 s—2584.827 s *应该*
2585.699 s—2587.665 s *大概*就是在中午
（literal translation: **should be**, **probably** at noon.)

As for *fillers* (see **§ 3.2.3**), both groups experienced fluctuations in their number across the Cycles. The Excel group maintained a high average number of *fillers* in Cycles I and II, which then decreased in Cycle III. This decrease could be attributed to their adaptation to the task's demands or their evolving expertise in controlling *filler* counts. In contrast, the InterpretBank group, from Cycle I onwards, consistently exhibited a lower average number of *fillers* compared to the Excel group. Although there was an overall downward trend from 12.82 *fillers* to 10.73 *fillers* from

Cycle I to III, the magnitude of this change did not seem to be substantial. The InterpretBank informants used fewer *fillers* in Cycle III, compared to Cycle II, which might suggest that they adapted to it over time, leading to a more fluent performance. Adopting InterpretBank in Cycles II and III did not result in statistically significant differences, as confirmed by the Friedman test for within-Cycle comparisons and the Mann-Whitney U Test for between-group comparisons. Differences in filler counts suggest that using InterpretBank may lead to fewer of them, but such differences were not statistically significant between groups and across Cycles. Evidence here is thus inconclusive, and the topic deserves further, more focused research. In sum, data on *fillers* neither supports nor undermines the hypothesis that using InterpretBank may contribute to improving RSI booth performance efficiency over time.

Repetitions in **§ 3.2.4** describes cases when a sequence of at least two words is repeated immediately after uttering them once (e.g., ABCDCDE...). The data show a decline in *repetitions* from Cycle I to Cycle III in the Excel group, decreasing from 4 to 2. This trend could indicate the informants' increasing familiarity with the Excel glossary and their adaptation to the demands of the interpreting booth tasks. In contrast, the InterpretBank group did not follow the same trend as the Excel group, maintaining the same median number of repetitions in Cycles I and II *(4 repetitions).* However, by Cycle III, the repetition count for the InterpretBank group also decreased to 2, possibly because of becoming more proficient with the tool and its features over time. When comparing the median *repetitions* from Cycle II to Cycle III, both groups exhibited a similar reduction. Furthermore, the Friedman test for within-group comparisons across Cycles revealed no significant differences in *repetitions* within each group. Additionally, the Mann-Whitney U test, as shown in **Figure 49**, indicated no significant inter-group differences in *repetitions*. Therefore, while *repetitions* may play a role, they do not significantly support the hypothesis that using InterpretBank contributes to the efficiency of RSI rendering at the booth.

In closing, the results of the content cluster offer mixed results. While *false starts* may suggest that InterpretBank may enhance fluency, other metrics like *repetitions*, *self-corrections,* and *fillers* indicate an increase in counts following the introduction of InterpretBank in Cycle II (post-treatment), compared to Cycle I (baseline, pre-treatment). Despite these findings, there is no conclusive evidence supporting Hypothesis 3. The good news is that the methods seem adequate to research this topic and, considered together, the data point to complex relationships between the indicators hinting at interpreters' strategic, controlled, and intuitive behaviors (whether aware or not), in accordance with the views laid out in cognitive translatological approaches.

Additionally, our findings in **§ 3.2** also reveal that there were no significant differences between the groups and within the groups for most fluency indicators across Cycles. Based on the Friedman test for within-group analysis, no significant differences were found within the groups from Cycle I to Cycle II and from Cycle II to Cycle III for all fluency indicators. According to the Mann-Whitney U Test used for inter-group analysis, there were no statistically significant differences in content cluster among fluency indicators *(false starts, self-corrections, repetitions, fillers).* Moreover, no statistically significant differences between groups were found in Cycle I and Cycle III. In sum, the fluency indicators do not support the hypothesis that the use of InterpretBank contributes to improved efficiency at RSI rendering.

4.3.2 Fluency: Time cluster

B*umps* (**§ 3.2.4**) and *respites* (**§ 3.2.6**) in this study codify minor and major gaps in the spoken language flow, which in turn are taken as proxies for possibly unnoticed vs noticeable disfluencies. The data from Cycle I offered a comparable baseline for the groups.

In **§ 3.2.4**, the median *bump* count for Excel informants was 68 in Cycle I and, for InterpretBank informants, 77. In Cycle II, the Excel group median *bump* counts rose to 89, and the InterpretBank group's median value of 117 was noticeably higher. Cycle III confirmed this difference, with the Excel group's median bump counts increasing to 91, whereas the InterpretBank group's median shrank a bit to 110. The emerging disparities in Cycles II and III are particularly significant, indicating that InterpretBank informants consistently experienced more *bumps* than their Excel counterparts. These disparities are substantiated by inter-group analysis using the Mann-Whitney U Test, revealing a statistically significant difference between the groups in Cycle II, but not in Cycles I and III. Perhaps more familiarity with InterpretBank is a possible reason for this. However, within-group comparisons for the InterpretBank group showed no significant differences between Cycles II and III. This suggestion should be taken cautiously anyway since a possible reason could be the influence of outlier performance. Again, a reminder that (in this case, group) averages may not be a reliable indicator for extended tasks.

The reduction in *bumps* for the InterpretBank group in Cycle III suggests various potential reasons, indicating that using InterpretBank may not be a decisive factor in performance concerning *bumps*. Although *bumps* codify less noticeable interruptions, their frequency may suggest higher cognitive efforts, as argued by Shreve *et al.* (2010), Shreve *et al.* (2011), and Muñoz & Apfelthaler (2022) for written translation and sight translation. The tools employed could influence the flow of speech delivery. InterpretBank might foster minor distractions or using Excel may lead to fewer hiccups in the speech flow, at least when the renditions is in Chinese. Finally, source speech in Cycle III may have been slightly easier to interpret than that of Cycle II (see **Table 1**). Be it as it may, data on *bumps* does not support the hypothesis that using InterpretBank may contribute to improving RSI booth performance efficiency over time.

Spontaneous speech often contains *respites,* and they were also consistently observed in all informants' recordings. This consistency makes it a challenge to interpret them. The groups exhibited similar counts of *respites*, with an increase in Cycle II and a decrease in Cycle III. The only plausible explanation beyond uncontrolled confounders and unknown circumstances is a higher difficulty in the task. In any case, under the light of *bumps* rising in both the second and the third

Cycles, it is the lack of a comparable rise in *respites* that is remarkable. This probably hints at the psychological reality of the intent of informants to avoid long(er) gaps in the flow. That is, additional demands or higher efforts threatening interpreting breakdowns would lead to informants strategically adapting their behavior, either with further *bumps*, or more *dropped chunks*, *repetitions*, and *fillers* to disguise time breaks devoted to facing the demands or to spread the effort along the flow.

Section **§ 3.2.6** presented data showing an increase in the median number of *respites* for both groups (Cycle I: 158, Cycle II: 178.5, Cycle III: 164.5), indicating that both the Excel and InterpretBank groups experienced an uptick in *respites* in Cycle II. The Excel group recorded more *respites* in each Cycle than the InterpretBank group. Although a downward trend was observed from Cycle II to Cycle III, no statistically significant differences were found in within-group or between-group comparisons. Therefore, while data on *respites* does not support the hypothesis that using InterpretBank may contribute to improving performance efficiency over time, it suggests that the notion of trade-off behaviors pointing to strategical steering of one's resources and skills is worthy of further study, and in agreement with the views laid out in cognitive translatology (Muñoz, 2023).

The above outcome suggested that while InterpretBank use may influence interpreters' performance, this influence was subject to individual variation within the group for booth tasks. Regarding *bumps*, and *respites*, both groups exhibited similar patterns in their frequency. Therefore, it cannot be conclusively stated that InterpretBank directly impacted these fluency factors among InterpretBank informants. Additionally, the high counts in both indicators for both groups could suggest cognitive difficulties experienced by informants when handling novel terms in speeches rich in technical terms.

Interpreters are expected to start uttering as soon as possible after the source speech but heavy cognitive demands lead interpreters to lengthen their EVS (Lee, 2004). This was expected to be particularly so in interpreting trainees dealing with term-intensive speeches. The content cluster is assumed to flag some of the interpreters' cognitive efforts in processing certain items while speaking (and listening) due to conflicting or competing demands. Another way of seeing the differences between the two clusters is that the time cluster hints at proactive mental processing and the content cluster, at reactive mental processing.

In the Excel group, a downward trend in the median duration of *EVS1* was observed from Cycle I to Cycle II, decreasing from 5.20 to 4.35 s, followed by a slight increase to 4.49 s in Cycle III. Conversely, the InterpretBank group showed an increase in *EVS1* in Cycle II (compared to the Cycle I baseline), reaching 5.56 s, which is higher than the Excel group for the same Cycle and also higher than the InterpretBank group's own baseline in Cycle I. One possible explanation for the observed variations could be the influence of introducing InterpretBank in Cycle II on the InterpretBank informants' processing source speech chunks. In Cycle III, the InterpretBank group's *EVS1* decreased to 4.24 s, marginally lower than the Excel group's 4.49 s. This suggests that informants in both groups may have been adjusting their interpreting strategies to optimize *EVS1* in Cycle III. In contrast, InterpretBank informants experienced an increase in Cycle II and a decrease in Cycle III, a pattern similar to those of other indicators in the content cluster. Integrating InterpretBank terms may have led to longer *EVS1*s, with informants perhaps taking more time between recognizing inputs and producing their renditions because they had to integrate InterpretBank use in their routines. Whether explained as related to limited working memory, additional multitasking demands, higher levels of stress, or a combination thereof, increasing EVS1s might be at least partially explained because of introducing InterpretBank.

Regarding within-group comparisons, the Excel group showed no statistical differences in *EVS1* between Cycles I-II, Cycles II-III, and Cycles I-III. In contrast, the InterpretBank group exhibited statistical differences between Cycles I-II, and Cycles II-III, but not between Cycles I-III. Although not formally tested, it is unlikely that all or most InterpretBank informants would have lesser WM than those in the Excel group. After all, *EVS1* values decreased in Cycle III in InterpretBank informants, so this makes it unlikely that the differences are due to problems "with the machine" (with brain/mind stable features or properties).

These differences might rather be attributed to longer *EVS1* durations in Cycle II for two informants (Blake and Harley), suggesting individual variability rather than a consistent trend attributable to the use of InterpretBank. Furthermore, the results hint at task performance dynamics and conditions of task execution, namely that the initial setback in *EVS1* in InterpretBank informants was due to introducing the tool and the recovery came as they became increasingly adapted to (the stressful situation and) the very use of the application. A potential confounding factor could be the informants' prior familiarity with the domain knowledge of the speeches, but it is unlikely that that would happen along the divide between the two groups.

In terms of median duration of *EVS2*, the Excel group exhibited a pattern of change similar to that observed in *EVS1*. There was a decrease to 3.63 s in Cycle II from the baseline of 4.32 s, followed by a slight increase to 3.73 s. In contrast, the InterpretBank group experienced an increase in *EVS2* during Cycle II, reaching 4.48 s, which is higher than both the Excel group in Cycle II and the InterpretBank group's baseline in Cycle I. In Cycle III, the InterpretBank group's *EVS2* decreased to 3.38 s, slightly lower than the Excel group. As in the case of *EVS1*, the rise in *EVS2* in Cycle II might be due to introducing InterpretBank, which possibly affected informants' information processing strategies.

Within-group comparisons for the Excel group did not reveal any statistically significant differences. However, in the InterpretBank group, two informants (Blake and Harley) were identified as outliers again, having longer *EVS2* durations than other informants in Cycle II. This resulted in statistical differences between Cycle I and Cycle II, as well as between Cycle II and Cycle III, indicating individual variability. In this context, the influence of expertise and InterpretBank usage appears to play a significant role in processing source speech chunks. This outcome may suggest that individual differences contribute more significantly to the overall results, highlighting the need for more in-depth intra-subject analysis to explore the underlying reasons for these variations.

In summary, the Excel group presented a pattern of adjusting their interpreting strategies for *EVS1* and *EVS2*, showing a decrease in Cycle II and a slight increase in Cycle III. In contrast, the InterpretBank group demonstrated longer median times for *EVS1* and *EVS2* in Cycle II, compared to Cycle I. This variation mainly results from two outliers. Possibly due to these two informants, significant differences of within-group differences were observed between Cycles I and II, and II and III. Nonetheless, most between-group comparisons within each Cycle and inter-group comparisons across Cycles did not yield statistically significant differences. Consequently, the analysis of time cluster does not support Hypothesis 3.

4.3.3 Term accuracy: potential problem triggers

RSI accuracy was studied by analyzing performance on rendering 39 problem triggers (**§ 2.2.2**)—33 *first-time* terms and 6 *repeated* terms—in each speech. The informants from both groups increased their accuracy through the Cycles in that they had increasingly larger counts of *correct renditions*, i.e., they used more target expressions matching those found in their revised master glossaries. In the median number of correct renditions, both groups showed an increase from Cycle I to Cycle III. The advantage of InterpretBank informants is more evident. In Cycle I, the difference in median numbers was small (7.5 correct renditions for the Excel group, 7 for InterpretBank). By Cycle II, InterpretBank surpassed the Excel group (12 for Excel, 13 for InterpretBank). In Cycle III, the gap widened (16.5 for Excel, 19 for InterpretBank). The results indicated that the InterpretBank group outperformed the Excel group in average correct terms in Cycles II and III. This growth was supported by significant differences in the Friedman test, suggesting that introducing InterpretBank potentially aided informants in delivering more correct terms. These findings align with Defrancq & Fantinuoli (2021), Prandi (2023), and Tammasrisawat & Rangponsumrit (2023). However, the Mann-Whitney test, used to compare group performance in each Cycle, detected no significant differences. This suggests that the increase in correct renditions with InterpretBank intervention might be due to more complex reasons, possibly related to individual variations in interpreting strategies adapted to the tasks across Cycles. Hence, the number of correct renditions did not support that InterpretBank contributes to a higher efficiency in RSI rendering.

The Excel informants typically rendered 1–3 potential problem triggers with adequate renditions in all Cycles, with a marginal increase in Cycle II but a decrease in Cycle III, down to the level of Cycle I (see results in **§ 3.3.2**). The InterpretBank informants exhibited greater consistency in frequency, albeit with important variations in individual performance. From Cycle I to Cycle III, they displayed fewer adequate renditions, compared to the Excel group. The frequency remained unchanged between Cycles I and II (at 1 instance) but dropped to 0 in Cycle III. Most InterpretBank informants offered 1–2 adequate renditions, with one exceptional case (Blake: 7 adequate terms in Cycles I and II, reduced to 1 in

Cycle III). On the one hand, informants might not necessarily know alternative renditions for the specialized terms, and adequate renditions might demand more effort than the correct ones. On the other hand, they might often consider the the trade-off cognitively effortful.

The informants might be consciously controlling their renditions to avoid ambiguous expressions to uphold the quality of RSI production, often resorting to the translation in the master glossary. Comparing the changes in adequate renditions with correct renditions across cycles, based on the average term counts shown in **Table 11**, we observed that while correct renditions from the Excel group consistently increased, adequate renditions peaked in Cycle II and then decreased in Cycle III. This trend suggests that Excel informants were adjusting their interpreting strategies, intentionally reducing ambiguous translations and increasing the count of correct terms. In contrast, with the InterpretBank group, despite an increase in correct term counts, adequate renditions decreased across cycles. This also reflects a conscious effort to reduce ambiguous translations and improve correct term counts. However, whether this decrease is directly related to the use of InterpretBank warrants further investigation.

Regarding wrong renditions in **§ 3.3.3**, the Excel group's median counts varied from 3 in Cycle I, to 2.5 in Cycle II, and then increased slightly to 3.5 in Cycle III. In contrast, the InterpretBank group showed an upward trend from 2 wrong renditions in Cycle I (pre-treatment) to 4 in Cycle II, before decreasing back to 2 in Cycle III. Despite these fluctuations, no statistical differences were found within each group across the cycles, nor when the groups were compared with each other. The pattern change in the median counts of wrong renditions for the InterpretBank group across Cycles was opposite to that of the Excel group. This may suggest that introducing InterpretBank in Cycle II might have increased wrong renditions, but by Cycle III, informants might have adapted to working with InterpretBank, possibly offsetting any negative impact. Individual differences across Cycles also played a role. For instance, Noel in the Excel group showed 3 wrong renditions in Cycle I, none in Cycle II, but increased to 11 in Cycle III (the highest in Cycle III). Harley from the InterpretBank group had 10 wrong renditions in Cycle I (the highest) but reduced to 4 in Cycle II and 2 in Cycle III. Kelly in the InterpretBank had zero in Cycle I, 6 in Cycle II, and 8 in Cycle III (the highest score). These individual differences were not limited to these informants and could affect both within-group and between-group analyses.

Skipped terms (see **§ 3.3.4**) both the Excel and InterpretBank groups exhibited a downward trend. The median count of skipped terms in the Excel group decreased from 27.5 in Cycle I to 17.5 in Cycle III. Similarly, the InterpretBank group showed a reduction from 30 in Cycle I to 17 in Cycle III. Since both groups followed a similar trend and the median counts of skipped terms were close (with no statistically significant differences found by the Mann-Whitney test in each Cycle for group comparisons), it is difficult to ascertain the impact of using either Excel or InterpretBank on the counts of skipped terms. However, a within-group comparison across Cycles revealed statistically significant differences for both groups. This trend suggests that the decrease in skipped terms might be attributed to increased familiarity with the task setting and demands, among other potential reasons.

Additionally, in each Cycle, the counts of skipped terms significantly exceeded the counts of the other three indicators, significantly exceed contradicts followed closely by the counts of correct renditions. This may indicate that when informants faced term-dense chunks with problematic triggers like technical terms, dropping terms was a common strategy. However, informants were not merely passive during this circumstance; this is evident when examining both wrong renditions and skipped terms, which are forms of inaccurate renditions (**Table 11**). For the Excel group, the average counts of skipped terms and wrong terms showed opposite trends (Cycle I, 27.2:2.5; Cycle II, 22.5:2.8; and Cycle III, 19.1:3.9), with skipped terms decreasing and wrong terms increasing. These opposite trends were not mirrored in the InterpretBank group (Cycle I, 26.6:2.7; Cycle II, 19.6:3.6; and Cycle III, 17:3), indicating that even when facing demanding term-dense speeches, the informants' cognitive efforts were actively and dynamically changing and continuously adjusting their interpreting strategies for the SI rendering. This adjustment might also be influenced by digital support (i.e., InterpretBank), suggesting an ongoing adaptation process. To further examine the performance related to the 39 potential problem triggers, we can analyze the data in *accurate renditions* (correct + adequate renditions) and *inaccurate renditions* (wrong renditions + skipped terms) (**Table 26**, next page).

For the Excel group's comparison of accurate and inaccurate renditions, we observed the following data for Cycles I, II, and III respectively: 9.3:29.7, 13.7:25.3, 16:23. Accurate renditions consistently increased from 9 in Cycle I to 16 in Cycle III, indicating that Excel informants improved in producing accurate renditions over time, thus raising their average counts of accurate renditions. Conversely, inaccurate renditions steadily declined from 29.7 to 23, possibly suggesting that their rendering strategy adapted to the tasks over time, helping them to handle demanding input and highlighting the active cognitive processes behind their interpreting activities. Similarly, in the InterpretBank group, there was an increase in average accurate renditions (from 9.6 in Cycle I to 19 in Cycle III) and a decrease in average inaccurate renditions (from 29.4 in Cycle I to 20 in Cycle III) across the Cycles (Cycle I, 9.6:29.4; Cycle II, 15.7:23.3; and Cycle III, 19:20). This also indicates that informants were improving their accurate renditions while reducing inaccurate renditions. As both groups exhibited similar trends, it is difficult to determine the specific impact of InterpretBank on the accurate and inaccurate renditions for the InterpretBank group. In both groups, the average counts of inaccurate renditions were higher than those of accurate renditions, possibly suggesting that informants exerted high cognitive efforts to produce the target renderings, especially when dealing with cognitively demanding term-dense speeches. As a reminder, for the purpose of further analysis and comparison between the InterpretBank and Excel groups, we merged accurate renditions (correct + adequate renditions) and inaccurate renditions (wrong + dropped renditions).

rendering	Cycles	groups	indicators	counts	total
accurate	I	**XL**	*correct*	8.3	**9.3**
			adequate	1	
		IB	*correct*	7.7	**9.6**
			adequate	1.9	
	II	**XL**	*correct*	11.6	**13.7**
			adequate	2.1	
		IB	*correct*	14.1	**15.7**
			adequate	1.6	
	III	**XL**	*correct*	15.3	**16.0**
			adequate	0.7	
		IB	*correct*	18.6	**19.0**
			adequate	0.5	
inaccurate	I	**XL**	*wrong*	2.5	**29.7**
			skipped	27.2	
		IB	*wrong*	2.7	**29.4**
			skipped	26.6	
	II	**XL**	*wrong*	2.8	**25.3**
			skipped	22.5	
		IB	*wrong*	3.6	**23.3**
			skipped	19.6	
	III	**XL**	*wrong*	3.9	**23.0**
			skipped	19.1	
		IB	*wrong*	3	**20.0**
			skipped	17	

Table 26. Average numbers for accurate and inaccurate renditions.

In summary, for the 39 potential problem triggers, the InterpretBank group demonstrated relatively better performance than the Excel group in terms of the median of correct renditions in Cycles II and III, following the employment of InterpretBank in Cycle II. This improvement is supported by significant differences in the Friedman test.

When comparing group performance in each cycle, no significant differences were detected. In adequate renditions, no significant differences were observed between-group comparisons across cycles. However, statically significant differences were observed within group comparisons across cycles. This variation within the groups across cycles could be largely influenced by one informant, Blake, in the InterpretBank group, who contributed a notably higher number of adequate renditions in Cycles II and III. For wrong renditions, there were no significant differences in either within-group or between-group comparisons. In terms of skipped terms, although both the InterpretBank and Excel groups showed significant differences within group comparisons, the median values of *skipped* terms changed in a similar pattern, decreasing from Cycle I to Cycle III. Moreover, group comparisons did not yield significant differences, indicating that the impact of InterpretBank on skipped terms is minimal and deserves further study. These outcomes align with findings from the analysis of average counts of accurate and inaccurate renditions, where both the InterpretBank and Excel groups followed a similar trend, which does not conclusively determine the impact of InterpretBank on accuracy: accurate renditions increased from Cycle I to Cycle III, while inaccurate renditions decreased over the same period.

In conclusion, the impact of InterpretBank on the accuracy of 39 potential problem triggers requires further investigation. The accuracy indicators (i.e., correct, adequate, wrong renditions, and skipped terms) do not support Hypothesis 3. Individual variations among informants were high for accuracy in potential problem triggers, underscoring the limitations of a one-size-fits-all approach to RSI performance analysis. Term accuracy may be influenced also by topic familiarity, and other uncontrolled variables. While InterpretBank did not lead to advantages in accuracy, some informants may benefit from incorporating InterpretBank into their RSI workflow, potentially improving their performance. Under this light, CAI tool developers may want to prioritize user customization. Offering more modular and more customizable features, informants would be able to adapt their tools to meet their particular style or preferences, thereby enhancing overall rendition accuracy, if only thanks to enhanced user satisfaction.

4.3.4 Term accuracy: repeated terms

Three terms—a unigram, a bigram, and a trigram—out of the 33 *first-time* terms were repeated twice in each source speech (**§ 2.2.2**). The analysis remains within the Cycle, rather than across cycles, because the correct rate may vary due to many factors. The repeated terms are categorized into first repetitions *(rep1)* and second repetitions *(rep2)*. These categories can provide a benchmark of InterpretBank's impact on performance when terms are not totally new but possibly active in working memory.

The results in **§ 3.3.5** show consistent improvements in the Excel group from Cycle I to Cycle III for repeated terms. In contrast, the InterpretBank group displayed quite an improvement in the rate of correct renditions of repeated terms in Cycle II, followed by a downturn in Cycle III in both *rep1* and *rep2*. The improvement in Cycle II may be related to InterpretBank's support. In the Excel group, the correct rate for *rep1* was slightly lower than for *1st2Rep* (first mention) and *rep2* in each Cycle. The lower correct rate for *rep1* could be due to informants not paying sufficient attention during the first repetition. However, encountering the term again in *rep2* likely led them to devote more attention and retrieve the translation from memory. The increasing rate of accuracy in *rep2* also suggests that most translations of source terms were stored in the informants' memory and required an auditory stimulus for activation.

Comparatively, the InterpretBank group showed a similar pattern for repeated terms: lower accuracy rates for *rep1* and higher for *rep2* in each cycle, suggesting insufficient attention during the first repetition. Besides, the results in **§ 3.4.3** examined InterpretBank searches and show that informants scarcely resorted to searching for repeated terms: 6 term searches out of 66 repeated terms in Cycle II (3 for *rep1*, 3 for *rep2*) and 10 term searches out of 66 repeated terms in Cycle III (4 for *rep1*, 6 for *rep2*). This suggests a lower dependency on glossary support for repeated term interpretation.

After using InterpretBank, the accuracy rate for repeated terms was relatively high (5 out of 6 term searches leading to correct renditions in Cycle II, 8 out of 10 in Cycle III). However, the act of searching itself indicates that informants believed the source term would be in InterpretBank and that it would yield good results. This suggests that, in most cases, source terms (but not necessarily their meaning) were already active in the informants' memory. It is primarily the attention mechanism that plays a role: the *1st2Rep* acts as a stimulus, *rep1* refocuses attention, thereby facilitating the retrieval of translation or source terms from memory. In sum, the use of InterpretBank did not significantly contribute to the interpretation of repeated terms, so that the performance of repeated terms cannot support Hypothesis 3.

From Cycle II to Cycle III, the contribution of InterpretBank to correct renditions of repeated terms is limited. Informants in the two groups appear to rely on their memory resources for recall. However, some informants (Gale, Kelly and Jordan) kept using InterpretBank's glossary to look for them. Biagini (2015) & Prandi (2015) used the number of queries as an indicator of ease of interaction between interpreters and CAI tools. However, performing queries while simultaneous interpreting is one of few instances where a person will simultaneously read, write, listen, and talk, making it an extreme case of language use within an extreme case of language use (cf. Obler, 2012). Under this light, we assume that the informants would try to avoid such efforts whenever reasonably possible (see term search in sample 3 for rendition checking), as it may be a case when they identified that the terms were repeated.

sample 3
[source speech] ...*The idea is that there's a* ***dorsal vagus*** *which kind of runs down the back of the spinal cord, which is involved in alertness and activation and fight or flight type stuff* ...
33:53.571—33:54.634: dorsal vagus
[target speech]
33:54.063: the informant uttered the condition of dorsal vagus（迷走神经背侧）
34:01.558—34:01.841: the informant typed character strings (i.e., *dor*).
34:01.558: suggested translation of *dorsal vagus* was shown on the screen.

Their (over-?) reliance on InterpretBank may have led them not to keep such terms active in their WMs. That is, the informants' certainty of finding a satisfactory rendition in the revised master glossary after having retrieved it at first time appearance might have led them to reduce the memory traces to just recognizing it as seen/used. Informants retain lexical information that they assume related to the source speech, when "information is kept, removed, or distorted when it is transmitted from one person to the next" (Y. Li *et al.*, 2024). Yet this reliance might not be merely about retaining lexical knowledge, but also about strategic decision-making regarding CAI-tool use.

When presented with an auditory signal from the source speech as cued-recall, it may stimulate the retrieved information for *rep1* and *rep2*, whether it be the correct translation or phonological information of the specific source term still present in the WM. Cued-recall in sequences (*1st2Rep, rep1, rep2*) should be expected to result in an improvement in the number of correct and adequate renditions in the informants. Informants might choose (even unconsciously) to always retrieve terms from the CAI tool, thereby freeing cognitive resources from rehearsing and updating WM, thus becoming a feature of strategic behavior eased by InterpretBank. In this sense, their behavior can be deemed more efficient.

4.3.5 Summary

Across cycles, the intra-subject analysis of the content cluster shows that InterpretBank informants reduced *false starts*, *fillers,* and *self-corrections,* although with important individual variations. InterpretBank informants show individual fluctuations in the performance of the content cluster, perhaps related to their varying capacities to adjust to both the tool and the task at hand—which would explain why such fluctuations are observed also at the group level from Cycle I to Cycle III. As a reminder, the variations that these data reflect happen in cognitive and self-regulatory processes that are not an explicit part of formal training. The limited evidence of reduced *repetitions* across cycles in the InterpretBank group is not so strong as to convincingly suggest that InterpretBank may lead to fewer *repetitions*.

Based on within-group testing (**Table 27**), there was no statistically significant difference in the counts of *false starts*, *fillers,* and *self-corrections, repetitions* for the Excel group and the InterpretBank group. However, the InterpretBank group had a lower statistically significant difference in *repetitions*. Furthermore, there was no statistically significant difference between the InterpretBank and Excel groups in the content cluster across cycles.

As in the case of the content cluster, in the time cluster the Excel group did not show significant differences in *bumps, respites, EVS1 and EVS2* (see **Table 27**)*.* However, the InterpretBank group, except for *bumps*, showed significant differences in *respites, EVS1* and *EVS2*. Such significant differences cannot be solely due to introducing InterpretBank in the InterpretBank group, because the *respites, EVS1* and *EVS2* showed the same changing patterns of frequencies in both the Excel and InterpretBank groups across cycles. Additionally, the high occurrence of *bumps* and *respites* in both groups suggests that informants experience cognitive challenges in processing term-dense speeches regardless of the CAI tool they use in the booth tasks.

		within-group (Friedman test)		**between-group (Mann-Whitney U test)**		
		Excel	**InterpretBank**	**Cycle I**	**Cycle II**	**Cycle III**
fluency	*false start*	0.80	0.45	0.969	0.757	0.415
	self-corrections	0.69	0.59	0.09	0.54	0.21
	fillers	0.62	0.64	0.209	0.11	0.094
	repetitions	0.31	0.01*	0.35	0.26	0.85
	bumps	0.10	0.36	0.24	0.02*	0.25
	respites	0.41	0.03*	0.88	0.76	0.27
	EVS1	0.64	0.01*	0.20	0.11	0.70
	EVS2	0.64	0.03*	0.22	0.08	0.65
accuracy	correct r.	0.01*	0.01*	0.67	0.18	0.18
	adequate r.	0.01*	0.02*	0.45	0.12	0.30
	wrong r.	0.47	0.51	0.85	0.44	0.44
	skipped t.	0.01*	0.01*	0.91	0.11	0.14

* statistically significant correlation

Table 27. P-values in statistical tests for fluency and accuracy indicators

For the between-group testing, time clusters showed no significant differences in *respites*, *EVS1* and *EVS2* across cycles. An exception was observed with the bumps in Cycle II, where the median number of bumps was higher for the InterpretBank group compared to the Excel group. Such variations were largely due to individual performances, so that the analysis of the time cluster does not support Hypothesis 2]34. The InterpretBank group showed longer median time gaps for *EVS1* and *EVS2* in Cycle II than in Cycle I, largely due to two informants (outliers). In contrast, the Excel group exhibited a change in interpreting strategies for *EVS1* and *EVS2*, with a decreasing median second in Cycle II followed by a minor increase in Cycle III.

In terms of overall accuracy *(i.e., correct, adequate, wrong renditions, and skipped terms),* there were notable individual variations among the informants. This underscores the limitations of universal approaches to RSI quality, suggesting that term accuracy may be influenced by personal factors such as prior knowledge and even attitudes toward digital technologies. Both aspects may have an impact on the individual reliance on the CAI tool. Furthermore, term accuracy might fluctuate as tasks unfold, conceivably linked to mental fatigue or the individuals' interaction with other factors, such as consoles, other apps, their smartphones, the speakers' slides, the room where they actually interpret, etc. Such individual traits, preferences, and circumstances may also be beneficial, and emphasize the potential advantages of CAI tool developers' prioritizing customization, so as to enable informants to tailor the tools to their specific needs and styles. This customization could potentially enhance both term accuracy and users' overall satisfaction.

For the 39 potential problem triggers, both groups exhibited a trend of increasing correct renditions counts across cycles. the InterpretBank group surpassed the Excel group in median correct renditions in Cycles II and III, a difference whose significance was supported by the Friedman test. Similarly, in terms of skipped terms, both groups demonstrated a trend of decreasing *skipped* terms, reflecting statistically significant within-group differences across cycles, but no significant differences were observed between groups, highlighting the need for further research to understand InterpretBank's impact. No significant differences were observed in adequate and wrong renditions either within or between groups. Both followed the same pattern: an increase in accurate renditions and a decrease in inaccurate renditions, making it unclear if InterpretBank directly influenced accuracy. Therefore, accuracy indicators cannot support this hypothesis.

On the other hand, some InterpretBank informants relied on the tool to different degrees in Cycles II and III. In terms of recall of correct terms, the contribution from InterpretBank was limited. Overall, the informants did not heavily rely on InterpretBank to render repetitions, so its impact on the accuracy of repeated terms was minimal in Cycle II and Cycle III. Instead, informants seem to draw on their memories. Still, some informants used InterpretBank, suggesting a possible reliance on the tool that would free cognitive resources. This reliance may stem from the confidence in the glossary, which would move them to drop their cognitive efforts to maintain memory traces beyond mere term recognition. InterpretBank and similar CAI tools thus have dual roles in that they not only reduce the need for memory retention but also impact decision-making and planning, that is, they are not only relevant from the perspective of the informants' capacities but also from that of their strategic behavior. As a result, some informants may focus more on other aspects, such as understanding the source speech, controlling their intonation, or keeping a moderate EVS. However, this may come at the cost of a reduced engagement with memory use, turning RSI into a more local affair.

Overall, the fluency indicators *(false starts*, *corrections, fillers, repetitions, bumps*, and *respites*, along with *EVS1* and *EVS2)* and accuracy indicators *(correct, adequate, wrong renditions,* and *skipped terms)* for the 39 potential problem triggers, as well as the side study on repeated terms, do not support Hypothesis 3.

4.4 H4: Improved documentation performance will yield better RSI rendering quality(H3)

The fourth hypothesis conjectured that RSI rendering quality might be linked to glossary compilation due to the memory traces built when compiling terms. **Table 28** summarizes average values of documentation behavior for glossary compilation and quality scores, data per cycle, and group. The InterpretBank group had a lower diversity rate (Cycle I, 29.0%, Cycle II, 27.7%, and Cycle III, 27.8%) than the Excel group, probably an effect of the automatic extraction feature of the CAI tool. The lower diversity and the larger number of entries point to a *one-size-fits-all* algorithm. Thus, while InterpretBank offers convenience and time-saving benefits, it may also inadvertently homogenize individual glossary terms. There is nothing wrong with this approach, in principle, but terms may become distractors, especially when users activate the term recognition feature, and when presented with more options to choose from when typing letter strings. Tool developers might want to consider some personalization leading to reduced noise. An expert interpreter in some specialized domain probably does not need the same glossary our informants compiled.

indicators	**group**	**Excel**			**InterpretBank**		
	Cycle	**I**	**II**	**III**	**I**	**II**	**III**
total time (s)		5749.5	4730.1	4527.9	4949.8	4554.9	4107.0
term (#)		59.1	78.8	44.8	91.3	124.8	73.3
time per term (s)		97.3	68.9	115.1	54.2	41.5	63.0
diversity rate (%)		42.8	45.3	46.4	29.0	27.7	27.8
rating score		4.3	3.7	3.7	3.8	3.0	3.1
rating (%)		33.3	40.5	48.9	44.4	59.5	57.7

Table 28. Glossary task indicators and interpreting quality for two groups.

group	name	total time (s)			term (#)			time per term (s)			diversity rate (%)			rating score			rating (%)		
		I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
XL	Morgan	9689.1	4682.8	4405.5	94	84	47	103.1	55.7	93.7	56.1	39.1	31.1	4.2	3.0	3.3	36.0	60.0	53.3
XL	Noel	1511.9	1511.9	1183.3	40	84	40	37.8	18.0	29.6	24.4	32.9	16.7	5.7	4.8	4.3	6.7	24.0	33.3
XL	Oakley	7795.1	8135.3	8281.0	79	70	59	98.7	116.2	140.4	35.1	26.4	37.7	4.3	4.3	4.7	33.3	33.3	26.7
XL	Peyton	4659.7	4468.4	2540.6	60	70	24	77.7	63.8	105.9	31.7	28.0	15.8	4.0	3.0	2.3	40.0	60.0	73.3
XL	Quinn	3791.8	5604.9	5823.9	64	94	59	59.2	59.6	98.7	43.9	37.7	40.4	3.0	3.3	3.2	60.0	53.3	56.0
XL	Riley	8405.3	7989.2	5390.8	80	90	39	105.1	88.8	138.2	44.7	37.2	31.1	4.0	2.3	2.7	40.0	73.3	66.7
XL	Sidney	5056.2	4282.4	5372.1	43	42	37	117.6	102.0	145.2	18.3	13.2	21.1	4.3	3.4	3.7	33.3	52.0	46.7
XL	Taylor	2732.1	2511.6	2292.8	46	31	11	59.4	81.0	208.4	27.5	11.6	6.6	3.8	5.3	3.3	44.0	13.3	53.3
XL	Uli	6926.0	3022.0	2642.3	54	163	86	128.3	18.5	30.7	26.7	52.0	53.9	5.0	4.0	4.0	20.0	40.0	40.0
XL	Val	6927.7	5092.4	7347.2	31	60	46	223.5	84.9	159.7	15.6	21.8	24.6	5.0	3.8	5.0	20.0	44.0	20.0
IB	Alex	2511.6	3781.2	4706.3	56	42	39	44.8	90.0	120.7	27.8	17.8	21.6	4.6	2.3	2.3	28.0	73.3	73.3
IB	Blake	4337.7	2671.8	2034.2	76	121	99	57.1	22.1	20.5	35.8	34.2	42.1	3.3	2.3	3.6	53.3	73.3	48.0
IB	Casey	2234.2	3897.8	2402.1	84	89	31	26.6	43.8	77.5	36.4	31.3	19.7	4.0	3.7	4.3	40.0	46.7	33.3
IB	Dana	6007.8	2494.4	2705.5	106	87	30	56.7	28.7	90.2	49.3	28.6	15.4	3.3	2.3	1.8	53.3	73.3	84.0
IB	Erin	7522.3	5310.7	5155.4	117	337	130	64.3	15.8	39.7	50.3	81.4	52.9	3.0	3.0	2.7	60.0	60.0	66.7
IB	Frankie	6369.5	6272.6	5287.0	72	146	55	88.5	43.0	96.1	34.4	45.6	27.0	3.7	2.0	2.7	46.7	80.0	66.7
IB	Gale	8350.1	4950.8	6367.9	91	138	82	91.8	35.9	77.7	39.7	42.2	34.4	5.0	3.0	3.3	20.0	60.0	53.3
IB	Harley	6340.5	6266.7	6226.5	82	146	118	77.3	42.9	52.8	37.4	41.9	52.1	2.7	2.3	3.0	66.7	73.3	60.0
IB	Ira	1085.2	3402.0	3256.1	93	113	93	11.7	30.1	35.0	43.0	30.5	36.7	3.7	3.3	2.8	46.7	53.3	64.0
IB	Jordan	2273.6	6243.1	4496.3	109	96	82	20.9	65.0	54.8	45.7	30.0	45.6	4.0	4.0	3.8	40.0	40.0	44.0
IB	Kelly	5478.1	845.9	716.1	107	47	38	51.2	18.0	18.8	45.4	13.5	15.1	4.4	3.7	4.0	32.0	46.7	40.0
IB	Lee	6887.3	8522.1	5930.8	103	135	82	66.9	63.1	72.3	44.7	41.1	37.8	3.7	3.7	3.3	46.7	46.7	53.3

Table 29. Documentation behavior indicators and interpreting quality.

Other than the diversity rate, attention should be paid to two primary indicators of documentation behavior: *time per term*, and *term counts* in personal glossaries. Whereas the diversity rate fosters group analysis, time per term focuses more on individual informants. Given its exploratory nature, this study adopted mainly a group-level analysis, but some individual data are intriguing (see individual average values of glossary compilation behaviors across Cycles in **Table 29**). For instance, Val from the Excel group demonstrates fluctuating *times per term* across Cycles: 223.5 s in Cycle I, 84.9 s in Cycle II, and 159.7 s in Cycle III. In Cycle I, despite having the highest time per term (223.4 s), Val's rating score percentage was only 20%, suggesting a low RSI quality compared to other Excel informants. However, in Cycle II, the informant's time per term decreased to 84.9, coinciding with an improvement in quality up by 44%.

However, this upward trend in score percentage did not continue into Cycle III, falling back to 20%, even as the time per term increased again to 159.7 s. This variation suggests that the correlation between time per term and interpreting quality can vary significantly among individuals and even in different Cycles, underscoring the importance of considering individual differences when assessing the impact of documentation behavior on SI quality. Contrastingly, Alex from the InterpretBank group, despite a consistent increase in time per term (Cycle I, 44.8 s; Cycle II, 90.0 s; and Cycle III, 120.7 s), did not show a corresponding increase in rating percentage (28.00%, 73.3%, 73.3%). This suggests a maintained high level of RSI rendering from Cycle II to Cycle III, a trend that is compared to other InterpretBank informants. However, Casey, showing a similar increasing trend in time per term (Cycle I, 26.6 s; Cycle II, 43.8 s; and Cycle III, 77.5 s), experienced a decline in rating scores after Cycle II (40.00%, 46.7%, 33.3%). This divergence exemplifies the complex linking informants' documentation behavior indicators to their rating score percentages. The variations in documentation behavior among different Cycles further emphasize that there is no straightforward correlation between these behavior indicators and the rating scores. Of course, factors such as language proficiency plus familiarity with the topic may explain the individual variations in glossary tasks. Current evidence does not strongly support a definitive association between documentation behavior indicators and RSI quality. There are, however, enough partial and indirect hints to keep working in this direction.

This initial exploration suggests that investigating individual differences might be difficult, for a Shapiro-Wilk normality test points to substantial individual variation in performance. Such variation may result from individual differences in cognitive processing, such as processing novel or repeated terms, allocating attention resources, and multitasking between comprehending source speech and retrieving terms with and within InterpretBank. Individual behavior may perhaps more clearly evolve within the task. Informants may show fatigue as they approach the end of the tasks, particularly in relation to speech chunks as measured with *EVS1* and *EVS2*.

4.5 H5: Improvements using InterpretBank but also attitudes, will lead to keep using it

Like H4, H5 is restricted to InterpretBank informants. This final hypothesis combines and derives from the previous ones, suggesting that, to continue using InterpretBank, informants need not only to have experienced improvements in their performance (even if not totally aware of them), both at glossary compilation and glossary use at the booth, but they also need to have liked it. We have discussed the findings from H1 and H3 in **§ 4.1** and **§ 4.2**, so here we address how the attitudes of the InterpretBank informants toward the tool and its use have evolved, as apparent in the survey results. Data consists of replies to three surveys. Two of them, conducted immediately after Cycles II and III, investigated the informants' opinions toward InterpretBank at glossary preparation and term retrieval at RSI tasks, and their intent to continue using InterpretBank for these tasks. As a reminder, in Cycle III all informants were allowed to choose the tools they preferred. The third survey was a follow-up, one year after the study, for the InterpretBank group.

Most informants expressed sustained satisfaction with InterpretBank for glossary preparation, but their awareness of being part of a study may have led them to conform to the predicted outcomes. This awareness could potentially result in response bias, as hinted by the divergent opinions in detailed queries about specific functionalities in both the glossary and booth tasks. The informants acknowledged time-savings and convenient aspects of InterpretBank for glossary tasks. While most of them deemed automatic extraction beneficial, a minority disagreed and concurred that maintaining a balanced approach between automated and manual extraction methods is wise(r).

There was a persistent interest among informants in using InterpretBank for booth tasks. This interest appears to rest upon the documented improvements in the correct renditions of terms. One informant, Blake, out of 11 informants was an exception to this trend: she shifted from a positive to a negative stance on using the tool (**Figure 94**), even though Blake's count of correct renditions consistently increased from Cycle I to Cycle II, and then to Cycle III (10:11:15). Blake's engagement with InterpretBank was minimal (see **Table 12** in **§ 3.4.1**), contributing to only 4 of the 242 searches in Cycle II and 4 of the 176 searches in Cycle III. Among the 11 InterpretBank informants, Blake's search frequency was the lowest. This exception suggests there may be other factors at play, which remain unidentified, potentially but not limited to individual preference or cognitive aspects related to a trade-off between demanding input and efforts for multitasking. Usage of InterpretBank was observed to decrease from Cycle II to Cycle III, with searches dropping from 242 to 176. This decline was specifically in the frequency of using InterpretBank's Booth mode, utilized for retrieving glossary entries at the booth to manage technical terms. The InterpretBank informant's training and exposure to the tool was quite limited, so the findings suggest that adopting it in the booth tasks is largely contingent on individual decision-making processes and task familiarity.

Moreover, disparate levels of confidence in the term retrieval support of InterpretBank's Booth mode offer further insights into the diverging viewpoints among informants. While technology might aid some students in correctly delivering terms, the mental toll of such computer-assisted support could not be directly observed or measured in this study. The survey reveals that InterpretBank informants' disparate levels of confidence in InterpretBank's term retrieval feature do not necessarily align with their performance improvements. Introspections may not represent users' actual experiences accurately, especially when informants are aware that they were observed and suspect the expected outcomes.

In Cycle III, one Excel informant (out of 10) switched to InterpretBank, while the rest continued using the tools they had been assigned for Cycle II. This was unexpected and informative by itself. That is, one subject was interested enough to switch to the new tool, even though she would have limited access to it and had not attended the InterpretBank workshop (the treatment)—as a reminder, recordings of both training sessions were nevertheless made freely available to all informants after Cycle III. Symmetrically, one InterpretBank informant (out of 12) switched to Excel for booth tasks. This was less surprising because it expressed an opinion on a new tool she had tried. In fact, in view of the short training, little practice, and stressful circumstances (RSI and research testing), acceptance of InterpretBank may be considered very high.

InterpretBank informants benefited from using the tool for glossary compilation, when they took less time in the glossary tasks and had more terms in the individual glossaries, compared to the Excel group (see **Table 28**). However, the diversity rate of the InterpretBank group was lower compared to the Excel group, as shown by the figures (Cycle I, 28.99: 42.80; Cycle II, 27.66: 45.30; and Cycle III, 27.76: 46.43). In other words, within the InterpretBank group, a greater number of identical terms were retained in individual glossaries. A contributing factor may be InterpretBank's *automatic term extraction* feature. The reduced time per term suggests less cognitive effort invested in each term. Contrary to the popular assumption in some CTIS research projects, here time is not naively assumed to be linearly related to cognitive effort but, in view of the relatively homogeneous backgrounds and training of the informants, the standardized test conditions and stimuli, and the overwhelming trends along group lines, it is safe to assume that here longer times hint at more actions that, in turn, needed cognitive support and hence, more cognitive effort—whether due to heightened control, problem-solving, new task integration and task coordination, or a combination of these and other factors.

Despite individual variations in documentation behavior, the InterpretBank group showed superior RSI rendering quality compared to the Excel group. Mixed attitudes were observed toward InterpretBank's support in the RSI task. In assessing the benefits of automated term extraction from InterpretBank, 7 out of 11 informants in Cycle II agreed or totally agreed on its time-saving features, while 2 were unsure and 2 disagreed. In Cycle III, the approval increased to 9 informants with *totally agree* and only 2 with *disagree*. Regarding the *reading-first term selection* feature in InterpretBank, 9 informants agreed or strongly agreed, one was unsure, and one disagreed both in Cycles II and III. However, actual usage data presents a contrasting picture. In Cycle II, only 5 informants mainly used InterpretBank for *reading-first term selection*, with time allocations varying significantly (e.g., Casey: 32.61%, Erin, Harley, Lee, Alex: all <5%). In Cycle III, the same five informants (Dana: 47.67%, Harley: 10.05%, Lee: 6.01%, Gale and Erin: both <1%) resorted to the same strategy for glossary compilation, which did not take full advantage of InterpretBank capabilities, indicating a discrepancy between survey responses and actual usage, with individual differences in time allocation for this task. Regarding the convenience of InterpretBank in compiling glossaries, all informants in Cycle II agreed and strongly agreed. By Cycle III, 10 informants remained in agreement, with one unsure. These responses may suggest that the informants claimed InterpretBank is beneficial for glossary compilation, which may be corroborated by data from **Table 6** in **§ 3.1.1.2**. The total time spent using InterpretBank by the group increased from 18,106.9 s (33.13% of total glossary time for the whole InterpretBank group) in Cycle II to 21,303.4 s (43.23%) in Cycle III, indicating a greater engagement with the tool over time.

Discussions on InterpretBank's booth tasks revealed varied opinions on the need for technical term support: in the Cycle II survey, six out of 11 informants indicated *sometimes*, with 2 for *mostly,* 2 for *always*, 1 for *never*), which evolved to 4 for *mostly*, 4 for *always*, 2 for *sometimes*, and 1 for *never* in Cycle III. Over half of the informants expressed confidence in InterpretBank for term retrieval. Inquiries about InterpretBank's ability to help locate target terms correctly showed that 6/11 for *mostly*, 3 for *sometimes*, 1 for *always*, and 1 for *never* in Cycle II; 4 for *mostly*, 4 for *sometimes*, 2 for *always*, 1 for *never* in Cycle III. Over half of the informants declared term retrieval helped them find the translation. However, the actual usage frequency of InterpretBank declined from 242 searches in Cycle II to 176 in Cycle III, indicating a discrepancy between survey responses and practical usage, possibly due to informants aiming to meet analysts' expectations. Of course, many other factors instead of merely perception of usefulness, are likely to be at play, including the cognitive efforts and resources required to use it and the prediction of issues that might appear along with the search, which could deter users, especially when they experience high demands. It could be a tendency to avert risks that leads to fewer searches, the topic, their familiarity with and confidence in RSI, etc.

When discussing the term retrieval feature's role in reducing pressure, opinions were even more divided (4 for *once in a while*, 3 for *sometimes*, 3 for *mostly*, 1 for *never* in Cycle II; 5 for *sometimes*, 2 for *mostly*, 2 for *always*, 1 for *once in a while*, one for *never*). In sum, while informants generally reported InterpretBank's benefits in glossary tasks, there was a discrepancy between declared support in survey responses and actual usage, particularly in the booth tasks. The survey also highlighted significant divergences in opinions regarding pressure reduction when working with InterpretBank. In other words, using InterpretBank was felt to be useful, yet informants were aware that using it is not necessarily easy or risk-free.

Ten out of 11 informants in the surveys even reported their intention to continue using InterpretBank for future booth tasks. However, a follow-up survey asked InterpretBank informants whether they did in fact continue using it in real-world situations one year after the third data collection Cycle was finished. Responses were received from 11 out of the original 12 InterpretBank informants. For comparative purposes, we included the informant who had not chosen InterpretBank in Cycle III, thus bringing the total back to 12 for a comprehensive overview. Among the 11 responses received, seven informants reported not having used InterpretBank after the study. Among these seven informants, four attributed this to the cost of licenses, two mentioned that they do not need it for daily practice and that the tool demands additional time and effort, and one cited that her current job did not involve interpreting. Among the remaining four informants out of eleven responses, two reported using InterpretBank very rarely (1–2 times a year), and only for *automatic term extraction*. One of those informants used it occasionally (3–7 times) for term extraction and flashcards, and the other one used it moderately (8–15 times) for the same purposes.

Overall, seven out of eleven InterpretBank informants had not used it in the year after they were introduced to it and tested about their use. Only four used it for *automatic term extraction* and, less frequently, the flashcard feature. Notably, none employed InterpretBank for booth tasks. These unexpected results contradict predictions regarding the software's high efficiency in enhancing the RSI quality and glossary documentation of RSI tasks for trainees. In spite of improved efficiency at documentation (which they may have experienced), and better assessments of rendering quality (which they were not informed about), perceived usefulness and ease of use may have contributed to these attitudes. Other factors might be the general attitude toward digital technologies, peer influence, economic burdens on students, and the need for training in professional settings (especially in RSI). More immediate reasons may be trainers' introduction and guidance of CAI tools and the duration of specific tool training. Of course, the results might be other with different samples.

We are now finished reviewing the hypotheses formulated as part of a mock quantitative study within this exploratory methodological project. We do have further results to comment upon: glossary compilation behaviors in two groups, discussion on search behavior features specific for the InterpretBank group, the relationship between duration of source speech chunks and EVS, and raters' holistic assessments.

4.6 Compilation of glossaries

As part of the analysis, informants, subtasks, and activities within them were rearranged, grouped, and color-coded in descending order of frequency or share of the aggregated task behavior (**§ 2.7.2.1**). Over time, varying degrees of reliance on tools, tool features, and services were observed, which suggested adaptation to task demands through different strategies. Our primary focus was on intra-subject analysis, then also group performance. To better understand the differences between Cycles for each participant, we adopted a percentage-based comparison. Quick reminder, we first calculated the cumulative duration of each strategy for each informant in each Cycle. Then, we calculated the percentage of these cumulative durations as a percentage of the total time for each participant. Since strategies belong to different subtasks, the percentage of a subtask is the sum of the times of its respective strategies. The results are shown below in **Tables 30–32** and we also visualized the tables into graphs in **§ 3.1.1.1**.

These figures display significant individual variations in the informants' allocation of time to different strategies and subtasks within glossary tasks. Not every strategy was involved by each informant, indicating distinct personal approaches to handling glossary tasks. While a comprehensive analysis of all evolving strategies for each informant is not feasible, we can focus on specific, essential cases to illustrate these individual differences.

For example, Noel from the Excel group demonstrated a unique approach in Cycle I, devoting 24.52% of their time to ***glossary review*** (which includes *glossary export* and *glossary formatting*)—the highest percentage for ***glossary review*** in their group during Cycle I. Additionally, Noel allocated 49.56% of their time to ***translation search***. In Cycle II, Noel continued to invest the most time within their group in ***glossary review*** (19.23%), with a decrease in ***translation search*** to 41.86%, and an increase in ***term extraction*** from 12.09% in Cycle I to 30.23%. This change became more pronounced in Cycle III, where *term extraction* rose dramatically to 84.83%, making Noel the informant who dedicated the highest percentage of time in the Excel group to this subtask. This progression reflects a strategic change in Noel's glossary preparation, initially focusing on ***translation search*** and ***glossary review***, and gradually pivoting toward ***term extraction***, indicating a continuous adjustment in his strategies in the glossary tasks.

In contrast, Gale from the InterpretBank group showed a different pattern across the three Cycles. In Cycle I, Gale devoted 18.36% of total time to term extraction and 79.43% to ***translation search*** (which includes *search result review*, *search queries*, and *translation input*, broken down as 2.25%, 66.14%, and 11.04% respectively). This distribution suggests a reliance on search queries during Cycle I. In Cycle II, Gale's focus shifted, with ***translation search*** comprising 78.97% of their time (*automatic translation* 71.25% and *search queries* 7.72%). This change indicates a strategic move toward increased reliance on *automatic translation*, a feature of InterpretBank, and a reduction in independent *search queries*. This pattern continued into Cycle III, with ***translation search*** still at 78.97% (*automatic translation* 71.25% and *search queries* 7.72%), exemplifying Gale's strategy evolution from *search queries* in Cycle I to increased use of CAI support. These cases highlight the idiosyncratic approach and strategy adaptation among informants when managing glossary tasks, underlining the diverse ways in which individuals interact with and utilize CAI tools.

subtasks and strategies	**Alex**	**Blake**	**Casey**	**Dana**	**Erin**	**Frankie**	**Gale**	**Harley**	**Ira**	**Jordan**	**Kelly**	**Lee**	**Morgan**	**Noel**	**Oakley**	**Peyton**	**Quinn**	**Riley**	**Sidney**	**Taylor**	**Uli**	**Val**
entry editing	**2.65**	**3.27**	**15.88**	**2.61**	**4.65**	**0.69**		**0.07**	**1.97**	**0.51**	**0.67**	**0.26**	**0.95**	**12.23**	**0.83**	**0.18**	**0.64**	**0.45**	**0.36**		**0.88**	
delete entries			12.96	1.53					1.57		0.37	0.08		12.23	0.02				0.36			
modify ST term	2.65			1.08	0.91	0.62			0.17	0.51		0.13	0.24								0.88	
modify translation		3.27	2.92		3.74	0.07		0.07	0.23		0.30	0.05	0.71		0.81	0.18	0.64	0.45				
glossary review	**0.98**	**52.36**	**20.13**	**1.89**	**2.43**		**1.86**	**2.93**	**0.00**		**1.29**		**1.47**	**24.52**	**0.31**	**1.40**	**2.50**	**0.06**	**0.33**			**0.32**
checking entries		52.36	18.50	0.88	2.10		0.06		0.00		0.39		1.47		0.31	1.23	1.71		0.33			0.32
glossary export			1.63	1.01							0.90			7.04								
glossary formatting	0.98				0.33		1.80	2.93						17.49		0.17	0.79	0.06				
st pre-processing	**0.94**		**15.03**	**1.14**	**6.86**	**7.29**	**0.32**	**5.08**	**0.04**		**0.70**		**19.15**	**1.61**	**0.61**		**24.74**	**3.69**	**2.39**	**3.81**	**3.32**	**14.36**
st alignment			4.04																			
st annotation			4.03					1.70					5.41				22.64	3.69				6.67
st chuncking			4.62																			
st formatting					0.72																	
text importation			2.34	0.72			0.18	0.58			0.70			1.61								
text reading	0.94			0.42	6.14	7.29	0.14	2.80	0.04				13.74		0.61		2.10		2.39	3.81	3.32	7.68
term extraction	**16.34**	**16.84**	**10.99**	**16.14**	**21.55**	**6.07**	**18.36**	**5.51**	**35.85**	**33.64**	**2.24**	**23.15**	**31.92**	**12.09**	**1.14**	**7.32**	**1.01**	**17.34**	**33.54**	**10.46**	**14.79**	**23.65**
automatic term extraction			10.99	1.38	0.77			0.64						10.93								
disfunction				2.20																		
reading-first term selection	16.34	15.62			12.06	1.00	15.78	1.92	34.37	33.64		23.15	30.28		0.34	6.00	1.01	15.83	31.91	10.46	12.38	23.65
source-text term retrieval		0.26			1.92		1.66															
task transition		0.96		12.56	6.80	5.07	0.92	2.46	1.48		2.24		1.64		0.80	1.32	0.00	1.48	1.18		2.41	
tool initialization								0.49						1.16				0.03	0.45			
translation search	**79.11**	**27.50**	**37.97**	**78.14**	**64.44**	**85.95**	**79.43**	**86.36**	**61.60**	**65.88**	**95.10**	**76.62**	**39.91**	**49.56**	**97.21**	**91.16**	**71.63**	**78.43**	**56.29**	**85.73**	**81.02**	**61.68**
search result review		0.75		0.70	0.21	0.96	2.25	1.81	22.45	1.37	0.48	2.26	5.06		3.68	1.94	21.27	5.79	1.42		3.93	
search queries	73.59	20.93	35.19	77.44	53.53	65.97	66.14	36.93	26.94	43.79	94.62	62.02	24.73	49.56	79.70	69.00	41.23	42.79	39.44	73.72	61.88	61.68
translation input	5.51	5.82	2.78		10.70	19.02	11.04	47.62	12.21	20.72		12.34	10.12		13.83	20.22	9.12	29.85	15.43	12.00	15.21	
total percentage	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Table 30. Individual subtasks and activities percentage in Cycle I.

subtasks and strategies	Alex	Blake	Casey	Dana	Erin	Frankie	Gale	Harley	Ira	Jordan	Kelly	Lee	Morgan	Noel	Oakley	Peyton	Quinn	Riley	Sidney	Taylor	Uli	Val
entry editing	**5.90**	**15.98**	**7.06**	**63.37**	**13.45**	**2.96**	**12.32**	**16.31**	**12.75**	**12.78**	**2.17**	**5.25**	**5.59**	**6.75**		**0.06**	**0.69**	**0.42**		**0.70**	**20.43**	
delete entries	4.95	5.28	4.88	13.25	0.47	1.11	10.63	2.41	0.81	1.08	2.17	0.54		6.75							20.43	
modify ST term	0.95	3.39		46.35				3.39	6.82	4.72		0.48					0.35	0.27		0.70		
modify translation		7.31	2.18	3.77	12.98	1.85		10.51	5.12	6.98		4.23	5.59			0.06	0.34	0.15				
remove punctuation marks							1.69															
glossary review	**1.34**	**2.84**	**4.57**		**2.41**	**0.58**	**7.99**	**1.84**	**3.23**	**1.62**	**13.23**	**0.07**	**1.36**	**19.23**		**0.96**	**0.08**	**0.67**		**0.67**	**3.27**	
checking entries	0.75	1.99	3.50		1.60	0.58	7.99	1.84	3.23	1.62	11.70	0.07		1.71							3.27	
glossary export	0.59	0.63	1.07		0.81						1.54			11.29								
glossary formatting		0.22											1.36	6.23		0.96	0.08	0.67		0.67		
st pre-processing	**0.79**	**0.48**	**5.51**	**5.69**	**0.80**	**17.65**	**0.43**	**6.54**	**0.24**	**2.63**	**5.32**	**1.61**		**1.92**			**1.58**	**14.31**		**0.69**	**0.13**	**15.05**
st annotation																	0.31	14.31				
text importation	0.27	0.48	1.71	2.58	0.09	0.30	0.15	0.25	0.24	1.91	5.32	1.61		1.92								
text reading	0.52		3.80	3.11	0.71	17.35	0.28	6.29		0.72							1.27			0.69	0.13	15.05
term extraction	**11.29**	**51.10**	**57.16**	**1.52**	**4.01**	**1.84**	**0.30**	**14.41**		**17.00**	**66.77**	**8.55**	**12.60**	**30.23**	**44.29**	**18.95**	**45.41**	**23.57**	**71.10**	**14.82**	**62.84**	**23.23**
automatic term extraction	0.29			1.37		1.84	0.30				60.64			2.51								
disfunction	0.25		0.02								6.13	0.01										
reading-first term selection	0.38		32.61		4.01			1.57				1.41	12.60		43.64	18.95	36.35	23.57	71.10	14.82	6.59	12.13
source-text term retrieval	7.59	46.87						12.84		8.87												
task transition	2.44	4.23	24.48	0.15						8.13		6.34		26.86	0.65		9.06				56.25	11.10
tool initialization	0.34		0.05									0.79		0.86								
translation search	**81.11**	**29.89**	**25.68**	**29.43**	**79.33**	**76.95**	**78.97**	**60.94**	**83.78**	**68.57**	**12.51**	**84.49**	**80.46**	**41.86**	**55.69**	**80.03**	**52.23**	**60.97**	**28.91**	**83.09**	**13.32**	**61.73**
automatic translation	33.16	1.85	2.25	14.94	6.87	1.65	71.25	3.16		0.25	9.25	0.91										
search result review	3.75	1.28	1.39					1.15	0.25	1.15		0.61			2.66			0.10			13.04	1.35
search queries	31.94	26.39	20.05	14.49	72.46	75.30	7.72	56.63	83.41	67.17		72.32	70.01	41.86	32.99	67.08	39.54	46.91	14.74	72.20	0.28	51.66
translation input	12.26	0.37	1.99						0.12		3.26	10.65	10.45		20.04	12.95	12.69	13.96	14.17	10.89		8.72
total percentage	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Table 31. Individual subtasks and activities percentage in Cycle II.

subtasks and strategies	**Alex**	**Blake**	**Casey**	**Dana**	**Erin**	**Frankie**	**Gale**	**Harley**	**Ira**	**Jordan**	**Kelly**	**Lee**	**Morgan**	**Noel**	**Oakley**	**Peyton**	**Quinn**	**Riley**	**Sidney**	**Taylor**	**Uli**	**Val**
entry editing	**0.69**	**1.19**	**10.14**	**6.32**	**11.26**	**20.64**	**0.73**	**21.32**	**24.05**	**12.19**	**0.53**	**3.06**	**2.21**	**2.49**			**2.24**	**0.42**	**0.19**		**0.72**	
delete entries		0.25	8.34	4.66	0.64	19.56	0.73	8.10	5.81	3.27	0.53	3.06		2.49								
modify ST term					3.17			1.82	10.06	0.14			0.25						0.19			
modify translation	0.69	0.94	1.80	1.66	7.45	1.08		11.40	8.18	8.78			1.96				2.24	0.42			0.72	
glossary review		**3.71**	**34.15**	**2.14**	**0.27**		**0.21**			**0.63**	**15.21**	**0.11**		**3.91**			**0.12**	**1.41**				**0.66**
checking entries		2.98	25.97	2.14							11.17											0.66
glossary export		0.73	8.18		0.27		0.21			0.63	4.04	0.11		3.91								
glossary formatting		0.00															0.12	1.41				
st pre-processing	**64.45**	**1.53**	**5.59**	**3.63**	**8.38**	**6.20**	**8.33**	**0.29**	**1.11**	**1.30**	**5.13**	**2.89**	**1.21**	**6.99**		**1.57**	**2.30**		**4.04**	**0.03**		**9.09**
st alignment			1.73																			
st annotation			0.81														2.30					0.17
st chuncking			2.11																			
text importation	2.85	1.53	0.94	3.63	8.38	1.03	1.07	0.01	1.11	0.91	5.13	0.68	0.55	6.99								
text reading	61.60			0.00		5.17	7.26	0.28		0.39		2.21	0.66			1.57			4.04	0.03		8.92
term extraction	**3.61**	**60.18**	**12.48**	**53.02**	**1.08**		**83.23**	**16.37**	**2.72**	**61.65**	**76.10**	**12.23**	**44.02**	**84.83**	**53.89**	**13.87**	**48.13**	**75.59**	**66.74**	**6.44**	**39.51**	**13.88**
automatic term extraction			5.48						2.43		52.09	0.37		4.80								
disfunction	1.83				0.38		79.85				24.01	1.38		39.31								
reading-first term selection				47.67	0.51		1.09	10.05				6.01	15.52	20.04	53.45	13.31	48.13	1.91	66.74	6.44	39.51	13.88
source-text term retrieval		57.38		1.90				6.32		4.90			28.03									
task transition	0.91	2.80	7.00				1.30		0.29	56.14		4.17		20.68		0.56		73.68				
tool initialization	0.87			3.45	0.19		0.99			0.61		0.30	0.47		0.44							
translation search	**31.23**	**33.37**	**37.63**	**34.89**	**79.02**	**73.16**	**7.52**	**62.06**	**72.17**	**24.63**	**3.03**	**81.74**	**52.55**	**1.78**	**46.13**	**84.49**	**47.20**	**22.55**	**29.01**	**93.54**	**59.81**	**76.30**
automatic translation	6.10							1.61		0.21	2.03	10.52	30.68	1.78								
Chatgpt prompt modification																					24.95	
search result review						0.44	0.20		0.42	0.84			3.64		2.72			0.42				
search queries	17.59	33.37	35.25	34.89	78.00	72.72	7.32	60.45	70.39	23.29		60.89	14.74		22.00	74.58	47.20	11.49	9.72	88.13	30.79	71.61
translation input	7.54		2.38		1.02				1.36	0.29	1.00	10.33	3.49		21.41	9.91		10.64	19.29	5.41	4.07	4.69
total percentage	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Table 32. Individual subtasks and activities percentage in Cycle III.

Data from **Tables 31** to **32** reveals that both groups devoted more time to ***translation search*** and ***term extraction*** subtasks, compared to other subtasks ***(i.e., st pre-processing, glossary review, entry editing).*** For instance, in Cycle II, InterpretBank was instrumental in many subtasks (e.g., ST pre-processing, glossary review, and entry editing). The seemingly simple process of automatic time extraction may appear to save a sizable amount of time. However, processing isolated terms selected by InterpretBank requires additional time and cognitive resources after *automatic term extraction*, in subtasks other than term extraction. One possible outcome is that *automatic term extraction* does not truly save time and effort but rather spreads it through the later steps, resulting, for instance, in increased multitasking or more frequent subtask switching. In contrast, the Excel group relied on *reading-first term selection* from source texts, their time and energy focused on term spotting, extraction, searching, and translation. *Reading-first term selection* consistently demanded substantial time, more than any other subtasks. Prolonged reading and translation search may lead to mental fatigue, increased stress levels, or induce a sense of aversion toward the task.

As **Figure 32** and **Figure 34** show, ***translation search*** was consistently the most time-consuming subtask for both groups. This includes *automatic translation, search queries*, *translation input*, *search result review*, *translation input,* and *ChatGPT prompt modification*. *Search queries* were in turn the most time-consuming activity across Cycles. A detailed analysis of informants' information seeking behavior for *search queries* is beyond the scope of this research project, but a few trends in tool use are obvious and addressed in **§ 3.1.1.2**. In Cycle I and Cycle III, *translation input* followed *search queries*. *Translation input* refers specifically to informants typing translations into a glossary entry, which constitutes a significant portion of the time spent and may suggest an attempt by informants to reinforce their memory of the translations through this activity. Otherwise, they might as well simply copy and paste it.

In Cycle II, the second position for the InterpretBank group was for *automatic translation.* This suggests that InterpretBank informants might have decided that *automatic translation* did not merit extensive time or effort and, consequently, they spared time there and redistributed it to other subtasks. Besides, *ChatGPT prompt modification* was used by only a small number of informants, its impact is minimal and not indicative of broader trends.

4.6.1 Sources

Cycle I data was the baseline benchmark of the glossary task. All informants preferred popular tools, including Google Translate, the Youdao multilingual dictionary (both app and web versions), the Oulu App (also a multilingual dictionary), and Baidu. Among them, the Youdao Dictionary and Oulu App are very popular among Chinese users. Terms are automatically translated in both directions of the pair. They also furnish bilingual examples sourced from the internet and provide English audio pronunciation. In most cases, the informants relied on translations or bilingual examples.

The choice of sources and their use remained consistent in Cycles II and III, so they were likely not contingent options but rather the informants' usual preferences. They very likely were accustomed to using them, so no additional learning or training efforts were required. In our view, these tools provide convenience and speed, but they may have their drawbacks too. Informants' searches are not always deep enough and, in some of them, translations are not guaranteed to come from professional or accurate sources.

The informants allocated a substantial amount of time to using search engines and bilingual dictionaries (**§ 3.1.1.2**). Often, individuals successively enter different keywords in search engines like Google and Baidu to refine their search results. Furthermore, most selected renderings came from bilingual texts on web pages found through search engines. Informants appeared to trust translations obtained from genuine bilingual contexts more than those from other sources (e.g., straightforward machine translation of isolated terms or with no co-text). As for bilingual dictionaries, aside from appearing to be frequent resources, the ones the informants used also provide links to bilingual texts, easing comparisons of translation options.

Another notable feature in glossary compilation observed from Cycle II to Cycle III was an increased reliance on InterpretBank in the experimental group. They progressively used it both when selecting terms and when looking for renditions, suggesting a growing dependency. However, the informants' reliance or dependency was not often blind: most InterpretBank informants diversified their approach, that is, they paired InterpretBank use with that of other resources to verify renditions or test hypothesized equivalences.

Some Excel informants used web-based terminology management systems (or *CAI tools,* depending on your definition), like LingoSail, TermBox, and Interpreter's Help.[18] TermBox is designed for Chinese users and offers automated term extraction and machine translation. Interpreter's Help is free and it is "designed not only to manage multilingual glossaries but also to manage job assignments and clients" (Costa et al., 2017, p. 67). One possible explanation for enlarging the pool of resources with these sources and tools (see **Table 6**) is that some informants may have been trained or at least introduced to them in their training programs, as supported by their replies in the sociodemographic profiling questionnaire (see in **Appendix A**) prior to the study.

In sum, the reasons to prefer some sources of information and to stick to them likely include the informants' prior experience with them and their exposure to them in training. Collectively, these factors contribute to a diverse array of tool choices across Cycles, reflecting a varied and dynamic approach to the glossary task. There is no room to further articulate all potential influences in the informants' decision-making process of term retrieval with different tools.

[18] Respectively, http://termbox.lingosail.com/ and https://interpretershelp.com/

4.6.2 Contents

As we saw (**Figure 38**), the InterpretBank informants generally compiled more terms than the Excel informants did, probably due to using the *automatic term extraction* feature from InterpretBank (see **§ 3.1.1.1**). This tool generates a larger pool of terms for selection before importing them into glossary entries, potentially resulting in a more time-consuming entry validation process (see **Figures 33 and 34**). Validating possibly unknown, isolated terms is particularly difficult, and informants had to resort to the source script now and then. Term counts fluctuated notably in informants such as Erin, Uli, and Ira, Noel across Cycles (**§ 3.1.1.3**). This hints at potential inconsistencies in their approaches to glossary compilation but might also reflect individual differences in topic familiarity and source text comprehension.

4.6.3 Flashcard use (InterpretBank group)

As a reminder, InterpretBank has a flashcard mode, and using flashcards is argued to be a useful method for term preparation for SI assignments (Goldsmith, 2023; Prandi, 2023; see also De Groot, 2000, pp. 54–60). To the best of our knowledge, however, there is no specific study on flashcard use in interpreter training. During the treatment—the InterpretBank training session—informants were introduced to this feature. The percentage of informants using the flashcard feature decreased from Cycle II to Cycle III (**§ 3.1.2**). Hence, the informants did not seem to convinced that it was worth using (see **§ 3.1.2**).

InterpretBank flashcards differ from traditional, user-created ones. For instance, users may choose the direction in the language pair. This may contribute to prioritizing memorizing source terms, which, in turn, may influence how well terms are remembered during RSI (Xu, 2015). One key distinction is that InterpretBank flashcards are automatically created (here, out of the master glossary, as explained in **§ 2.5.1**). Pan *et al.* (2023) argue that user-made flashcards are effective in enhancing memory retention and comprehension of textual materials. As mentioned, our findings did not really support this notion for automatically-generated flashcards— their use to prepare for the booth tasks was scarce—perhaps due to the relatively short exposure of the informants to the InterpretBank potential or else because of the different memory traces or their lack made using flashcards cognitively more demanding. This would explain why flashcard use declined between Cycles II and III.

4.7 Findings on search behavior from the InterpretBank group

Human-computer interaction prompts both novel approaches and debates such as the role of CAI systems in interpreter training (R. O. Tarasenko *et al.*, 2021). This section presents the unique features of search behavior by the InterpretBank group. It mainly focuses on the individual variation in workflow with InterpretBank, response times (i.e., ear-key span and eye-voice span) term accuracy with

InterpretBank search, and the correlation between *search duration* and *dropped chunks*.

4.7.1 Searching workflow

To explore the informants' interaction with InterpretBank, it is worthwhile to closely reflect on the underlying cognitive processes behind each search event in the workflow of RSI tasks with InterpretBank. First, we provide evidence of high variation in individual performance with InterpretBank during booth tasks when term retrieval is needed. Second, InterpretBank contributes to a high rate of correct term retrievals. Prior research on InterpretBank use did not perform any training or did not report how informants were trained, except for Prandi (2017). Some studies only provide limited exposure to InterpretBank without explicit training (e.g., Pisani & Fantinuoli, 2021). Alternatively, some studies have chosen to let interpreters use InterpretBank's advanced functions directly (e.g., Defrancq & Fantinuoli, 2021). In any case, there is still limited discussion on how to successfully integrate informants' training with InterpretBank to conduct empirical research.

In the present study, InterpretBank informants were trained to use this CAI tool for term retrieval in RSI tasks, but the isochronic training session was a one-off, short experience. The recording of the session was made available to all informants, and they had a week to train on their own, as suggested in the face-to-face training session, but all this may not be adequate or enough to ensure adoption and good use. This was intentional, in that it actually resembles more the preparation that a novel user will have outside an interpreter training program. Hence, informants were exposed to using it in a realistic task before they (and this researcher) could know whether the provided training was sufficient. Familiarity with and acceptance of InterpretBank varied, possibly leading to individual differences in allocating cognitive resources in the RSI tasks involving InterpretBank. As a reminder, informants' actions on their computers were screen-recorded, and no alternative strategies to learn or memorize terms were spotted. They were also requested to turn in any papers and additional materials they had used, but none of them did.

Using a machine-extracted glossary, rather than a personal, read-first glossary, leads informants to focus on more novel terms, whose understanding also varies. In any case, the behavior of InterpretBank informants differs from that of Excel informants, who mostly compiled glossaries through read-first approaches. Elgort *et al.* (2023) show that readers with prior exposure to definitions of new words allocated less attention to them when reading. The informants preparing a glossary are actually performing a task with written language, and Shreve *et al.* (1993) and Jakobsen & Jensen (2008) show that reading for translating is a cognitively deeper (perhaps more effortful?) task, because it focuses on meaning in ways deeper than just perhaps interlinguistic and metalinguistic. All the more so, we would contend, if we are cherry-picking terms that may later become potential problem triggers when hearing and uttering them.

Interpreting between distant languages adds another turn of the screw. Trainees whose L1 is distant from English may not necessarily guess how the original term will sound. This will often add a layer of cognitive complexity because that lack of familiarity with a given term—with its phonological representation—may make word recognition more difficult later on at the booth. Milligan & Schotter (2024) found that the phonological preview facilitates early word recognition during reading and we may safely assume that the same effects may be at work when hearing the word.

Furthermore, in logographic languages like Chinese and Japanese, written characters are not phonetic but rather represent meaning-based word constituents or full words, like 们 '[plural marker]', 你 'you', 脑 'brain', 人 'person', 木 'wood' and 爱 'love'.[19] These are very frequent characters that may be safely assumed to be generally known but, if you did not look at the footnote and you do not speak Chinese, you are probably clueless about how to pronounce them. Specialized terms in logographic languages are mainly coined by combining existing characters or through phonetic imitation of the term in another language, but chances are that reassurance or confirmation on how to pronounce it is often necessary.

Pronunciation checking was observed as a common action in the informants' translation search process from their screen recordings, with 7 out of 10 informants in the Excel group also stating in the survey that they need to do it during glossary preparation (see **§ 3.5.2.2**). Commonly observed in the screen recordings was the use of local dictionary applications or online dictionaries for hearing audio pronunciations. However, InterpretBank did not offer audio pronunciation for terms in glossary entries; version 8 notably lacked this feature.

Another frequent kind of retrieval activity is searching for *non-target terms*. Although these terms did not appear for the first time, informants looked for them again. Despite their recurrence, informants continued to search for them, whether because they were still unfamiliar with them, under stress, overwhelmed, seeking reassurance, or simply had consciously or not chosen not to keep the term active in their memories, once they had realized the glossary would readily offer a rendition. Of course, it is not clear how much one can willfully leave out of memory, and some traces need to remain to identify the term as located in the glossary. The point is that accessing it again in the glossary may become a strategy to free cognitive resources or to lower cognitive efforts.

Mistyped sequences (strings with misspellings or typos) often led to dropping terms in the RSI renditions when using InterpretBank. In Cycle II, there were 24 mistyped sequences, and in Cycle III, 15 (see **Figure 84** and **Figure 85**). In some cases, informants must have been quite certain that the terms they were searching for were in the master glossary. Mistyping may be due to poor typing skills or mechanical accidents—like placing the hand off for one key when back to the home row position—but it may also be the consequence of high cognitive efforts and stress, potentially impacting the rendering. Clearing the InterpretBank window and retyping the sequence might ultimately overburden the informants' WM while facing a potential problem trigger. Additionally, some character strings may match certain terms in the glossary that are not the correct ones (and there will be more

[19] Respectively pronounced [mén], [nǐ], [nǎo], [rén], [mù] and [ài].

chances, as the number of entries grows). Consequently, some informants may verbalize the incorrect term's renditions displayed on the screen, due to various reasons. For instance, the pronunciation and spelling of two terms were similar, so they retrieved a term similar to the one displayed. They might have also looked for a term in InterpretBank but had no clue as to whether the retrieved term was correct. In another scenario, they were uncertain of the term but, as they were pressed for a response, they decided to trust the suggested translation from InterpretBank. Finally, even when they know it is incorrect, they might feel pressed to keep the flow and use the wrong terms as a filler, that is, just to keep the audience's attention, in the hope they will ignore the wrong rendition and replace it contextually with the right one.

Hence, *mistyped sequences* may lead to dropping terms (and sentences, to compensate for the delay) or to uttering incorrect renditions, as a consequence of InterpretBank's inability to accommodate misspellings and some phonetic variations, in contrast to what other applications do, such as search engines. Implementing fuzzy search algorithms to also cater for the mis and alternative spellings might help alleviate the stringent requirements for precise typing. This is particularly important, again, for lower levels of command of the input language and for language pairs whose singletons are quite separate from each other. This is also particularly important for Chinese interpreters, who need to juggle several graphic and phonological representations when facing the keyboard and the microphone at once.

Another finding from the study is the rate of *correct* renditions counts and the efficiency of InterpretBank searches, that is, the proportion of successful term retrievals relative to the overall rendition accuracy. **§ 3.4.1** presents evidence of InterpretBank's high success ratios for correct renditions: in Cycle II, out of 153 search results from InterpretBank, 134 led to a correct rendition, yielding an effectiveness of 87.58%. In Cycle III, there were 115 correct renditions out of 127 searches (90.55%). Simply put, if (1) InterpretBank had been fed a useful glossary and (2) the appropriate sequences of characters were typed in correctly, the informants could successfully incorporate the suggested translations into their RSI renderings. Condition 1 relates to the users' glossary compilation success, while condition 2 is right at the moment of production, making the apparently separate processes a whole, situated, dynamic one.

Any novel information needs to be transferred from the external sources into the CAI system, and the user needs to select, reorganize, assess, link, and enter this information. InterpretBank informants, especially trainees, may feel tempted, as the data shows, to use their glossaries strategically to free up WM, which would only hold 3-4 information chunks for no more than 20 s (Cowan, 2001). Hence, InterpretBank informants opting for this strategy might only have or retain a superficial knowledge of the glossary entries. Such superficial knowledge might help explain some of the mistypes at term retrieval.

In this light, introducing InterpretBank in interpreter training programs might well be more beneficial if interpreting trainees do not overly rely on

InterpretBank as an external memory repository, but rather exercise and enhance their own memories to keep terms active in their minds, very much like regular interpreters do when unaided. The automatic extraction feature might be a double-edged sword and would need to accommodate concordance features to provide more context information for informants before they make decisions on term selection at glossary compilation. Further improvements might include references to standard (and dialectal!) pronunciation of both original terms and their renditions and some leeway for the sequences typed in to link to the right term even if mistyped. InterpretBank should also ease wiping out the search window in the Booth mode so that users can retype with no further ado.

4.7.2 Ear-key span and eye-voice span

Previous studies (e.g., Christoffels & De Groot, 2004; Timarová *et al.*, 2011; Chmiel & Lijewska, 2022; Guo & Han, 2024) used time lag as a single indicator of cognitive load in SI. To do so, they studied the interaction between the acoustic signals of the source speech audio with the corresponding utterances in the output. Here, we added keystroke events on the keyboard, and changes on the screen. The combined interactions made it possible to use ear-key spans (E2Ks) and eye-voice spans (I2Vs) to pinpoint behaviors likely to correlate with cognitive events in the interval between hearing a term in English, retrieving the entry in InterpretBank, and uttering the rendition in Chinese.

The E2K in Cycle II averaged 1.925 s and in Cycle III, 1.639 s (**§ 3.4.2**). The I2V averaged 2.309 s in Cycle II and 1.503 s in Cycle III. These results suggest rapid, moment-to-moment reactions. Confronted with auditory or visual stimuli, the informants exhibit reaction times with minimal average variation within a short timescale. The response durations, both the E2Ks and I2Vs, appear to result from the continuous processing of ongoing behavior. The scatter plots (**Figure 86**) show wide ranges for both E2Ks and I2Vs across Cycles, corroborated by expansive 95% confidence intervals, which suggest a high degree of variation. This variation may be a distortion due to the reduced sample size or reflect individual behavioral trends. Variation may also occur over time within tasks. Differences can arise from task difficulty, informants' reliance on tools, or even their prior knowledge. This may be traced in E2K and I2V values, suggesting that they are apt constructs to capture RSI dynamics.

Most E2Ks and I2Vs in this study yield positive numbers, indicating that, in most cases, actions follow their ideal sequential order, where one behavior leads to another and finishes before the next one starts, maintaining temporal consistency. In the E2K, keystrokes are prompted by the end of an acoustic signal of unfamiliar terms. For instance, in the I2V, utterances are prompted by the entry displayed on the screen. The temporal consistency demonstrates that the informants' responses to stimuli generally maintain the characteristics of a time series. Yet our study revealed negative values in the I2V (but not in E2K). Theoretically, E2K values may be negative, but only when both ASR and user-initiated search coincide.

A negative E2K indicates that an informant initiated a keypress for a search before the end of the sound wave for the term in the source speech soundtrack. This suggests that the informant had already identified the term in her mind before the term had been fully presented. This behavior is essentially a form of prediction, and it could be driven by an interest in saving time in term renditions or a need for verification of a known solution. Similarly, a negative I2V suggests that informants begin uttering the rendition even before they finish typing. This behavior may be due to typing beyond their needs, as attention is reallocated from typing to speaking, showcasing multitasking. Taken together, E2K and I2V portray the search process as a multifaceted human-computer interaction involving prediction, automatic processing, and multitasking, and they contribute to our understanding of the multimodal nature of live interpreting tasks.

Contrary to scatter plot results, Kendall's Tau-b tests reveal a statistically significant inverse correlation between E2Ks and I2Vs. As one variable increases, the other one tends to decrease, suggesting that longer typing times lead informants to try to compress the time available for rendering delivery. This in turn suggests that informants engage in self-monitoring as they are looking for information, adjusting not just their pace when delivering their renditions but also the speed of term retrieval. The search process with InterpretBank entails thus simultaneous controlled (self-monitoring) and automatic processes (typing beyond needs when already rendering the term). This reflects the dynamic complexity of interpreters working with CAI tools. Unlike passive information processors, interpreters actively adjust their strategies in response to the support and constraints provided by CAI tools, indicating an interactive and situated effort working with the technology.

E2Ks and I2Vs do not correspond directly. The presence of E2Ks does not necessarily imply I2Vs. In other words, the term retrieval event may stop after typing, possibly due to changes in strategy. For instance, to catch up with production time lags, such as in extended (sentence-level) *EVS1*, informants may pace up their production in source speech gaps but also forgo rendering terms even when already displayed on the screen. This may also result from mistyped sequences that lead nowhere in the glossary or lead to wrong entries identified as such. The outcomes vary among informants, and they also depend on the specific circumstances. In any case, the absence of I2Vs often flags an incomplete term retrieval, suggesting that the informant faced difficulties before or after the search event. The relationship between *search events* and *dropped chunks* will be explored later on (this point is planned but not reported here). To mitigate the adverse effects of the frequently mistyped sequences, CAI tool developers might want to resort to apply fuzzy recognition into the search bar. This would allow users to type character strings closer to the correct sequence of the target terms or accommodate cases where users may have an incorrect sequence in their memory.

In brief, E2Ks and I2Vs typically yield positive values, but there are instances of negative values (see **Figure 86**) that indicate an altered sequence of events. This raises questions about the interrelationship between these spans. For example, does a negative E2K influence the absolute value of the I2V? Can a potential spillover effect be indirectly measured? Due to the limited time frame and sample size of our study, these aspects remain unexplored, presenting potential areas for future research.

4.7.3 Potential problem triggers with InterpretBank search

This project aimed to be as naturalistic as possible. Collecting subject profiling information might have caused a white coat effect, that we wanted to avoid. Thus, the indirect confirmation of language command and a relatively developed simultaneous interpreting skill were assumed in view of their condition as at least second-year MA trainees in competitive training programs. Still, in view of the results, further detail on subject profiles seems convenient. As a reminder, we employed a correlative numerical labeling system for problem triggers that linked repetitions with their first appearances and also clarified their position in the sequence by adding the letters A *(first-time terms)*, B *(rep1)*, and C *(rep2)*.

InterpretBank's use by the group is inconsistent across terms. This is exactly what we expected. Language command and interpreting performance is after all different from person to person, the same way that no two interpretations of Debussy's *Claire de Lune* are ever exactly the same. In our case, it could be attributed to many combined factors, such as different text and term complexities, differences in informants' task preparation, such as the flashcard mode of InterpretBank, or the very use of ad-hoc glossaries in Excel or InterpretBank. The opposite, absolute homogeneity (or diversity), is what really needs explaining. For instance, the near-unanimous accuracy for term 01, which is the topic term, suggests that informants quickly and correctly grasped the central theme of the speech and possibly that they were fresh and ready for action.

Interestingly, some of the terms where informants performed exceptionally well were not supported by any glossary consultation with InterpretBank. This raises questions about the claims on CAI tools' contributions based on overall and imprecise counts of correct renditions, in view that often in the RSI renditions excelling in accuracy comes without the informants using them.

Furthermore, the renditions of repeated terms (those whose labels end with B and C, in **Table 33**) do not consistently rely on the usage of InterpretBank. Based on the classifications identified in **§ 3.4.3***,* the category *inaccurate renditions but IB not used* contained most *rep1* and *rep2*. This suggests a trend where certain terms, despite being repeated, were infrequently searched for with InterpretBank, and often inaccurately rendered. On the other hand, the category *accurate renditions but IB not used* included only a few terms. In Cycle II, this was true for T21_10B and T39_10C, and in Cycle III, for T39_32C. This indicates that while these terms were not often referred to when they were searched, the accuracy of their usage was high.

This may be influenced by learning effects, adaptation to InterpretBank, and increasing familiarity with problem triggers. Sankey diagrams (**Figures 89 and 90**) plot how each InterpretBank informant fared with term repetitions in Cycles II and III. The counts were obtained from screen recording files and cross-verified with keylogging data. Informants scarcely used InterpretBank for term repetitions, but they rendered them correctly in most cases when using InterpretBank (10 correct renditions of 12 term searches and 2 wrong renditions in Cycle II, 13 correct renditions of 18 term searches, 4 wrong renditions, 1 skipped term in Cycle III). However, the overall frequency of InterpretBank usage for term repetitions was low (12 term searches out of 66 repeated terms, and 18 term searches out of 66 in Cycle III). This high rate of success suggests that InterpretBank is effective when used for handling repetition terms. The striking insight is that the tool is not commonly used for repetitions, as informants seemingly do not need to search for the term, probably active in their WMs.

Cycle	term coded	term type	name
II	14A	*1st2Rep*	Gale
	26_14B	*rep1*	Kelly
	10A	*1st2Rep*	Jordan
	21_10B	*rep1*	Gale
	39_10C	*rep2*	Gale
	39_10C	*rep2*	Jordan
	18A	*1st2Rep*	Gale
	18A	*1st2Rep*	Jordan
	18A	*1st2Rep*	Kelly
	36_18C	*rep2*	Gale
III	20_07B	*rep1*	Jordan
	35_07C	*rep2*	Gale
	32A	*1st2Rep*	Jordan
	32A	*1st2Rep*	Erin
	32A	*1st2Rep*	Frankie
	32A	*1st2Rep*	Gale
	32A	*1st2Rep*	Ira
	36_32B	*rep1*	Jordan
	36_32B	*rep1*	Lee
	39_32C	*rep2*	Lee
	39_32C	*rep2*	Gale
	39_32C	*rep2*	Ira
	39_32C	*rep2*	Dana

Table 33. Correct renditions of *repeated* terms and *1st2Rep* in Cycles II and III

In **Table 33**, the distribution of correct renditions from 10 successful term searches in Cycle II and 13 term searches in Cycle III among informants is shown. In Cycle II, compared to five instances of correct renditions for *1st2Rep* search, repeated terms only received two out of 12 correct renditions for rep1, and three out of 12 for rep2. This suggested that a few informants successfully rendered the repeated terms using the interpreting bank. The informants showed a low degree of reliance on the interpreting bank for renditions of repeated terms. This was supported by the fact that only 3 out of 11 informants (Kelly, Gale, and Jordan) provided correct renditions with InterpretBank used. However, a notable case is Gale, who engaged with all three categories of searches both *repeated* terms and *1st2Reps*: one search for 14A, two searches for 21_10B and 39_10C (same term), and one search for 18A and 36_18C (same term again). This may indicate Gale's reliance on InterpretBank for repeated terms or challenges with personal recall.

In Cycle III, 13 correct renditions with successful searches covered 7 repeated terms (3 for *rep1*, 4 for *rep2*), and 5 for *1st2Rep*. Although the proportion of repeated terms of the term searches with correct renditions (7/13) increased, the differences between individual categories *(rep1, rep2, and 1st2Rep)* were not substantial (3:4:5). Additionally, the number of correct renditions with searches for *rep1* and *rep2* was lower than that of *1st2Rep*. This suggested that the InterpretBank did not noticeably facilitate correct renditions of repeated terms. Regarding the number of informants involved, five out of 11 informants used the InterpretBank for repeated terms with correct renditions, also including Kelly, Gale, and Jordan, which further demonstrated that individual reliance variation on the InterpretBank for assisting with repeated term renditions.

The role of a hypothesized regulation of cognitive efforts cannot be dismissed. Testing repeated terms seems to be able to shed light on memory use, and perhaps on the interpreters' personal cognitive styles, so it is a line that holds promise and should be further pursued. We turn now to explore this path and its impact on performance, particularly in relation to *dropped chunks* and search behaviors.

4.7.4 Search duration and dropped chunks

The frequency of searches often equals or surpasses that of dropped chunks. Their overlaps suggest an intricate balance between cognitive demands and multitasking challenges. Informants like Alex, Blake, Dana, Erin, Frankie, and Gale displayed distinct search behaviors (see **Figure 91**). Variation in the use of InterpretBank suggests that other cognitive factors may have been at work. For instance, Blake (B)'s search behavior in Cycle II involved 4 searches but 8 dropped chunks in Cycle II. Jordan, on the other hand, conducted the highest number of searches in Cycle II (47 searches) but also had 8 dropped chunks. This comparison suggests that Jordan may have been more adept at integrating InterpretBank searches into his renditions, as evidenced by a higher search count relative to the number of dropped chunks, compared to Blake. In Cycle III, Jordan exhibited the highest number of searches, 27, alongside 7 dropped chunks. This pattern of integrating InterpretBank searches into the rendition was similarly observed in Frankie's performance. In Cycle II, Frankie conducted 37 searches with 7 dropped chunks, and in Cycle III, he executed 26 searches, just behind Jordan (27 searches), with only 4 dropped chunks. This may hint at Frankie's comparable proficiency in utilizing

InterpretBank, managing a substantial number of searches while maintaining a relatively low number of dropped chunks.

The distribution of overlaps between searches and dropped chunks is not even through the tasks in Cycles II and III (see **Figure 91**). This supports our interest in studying behavior over time rather than performing a simple product analysis. All InterpretBank informants' RSI renditions were aligned into a universal timeline for Cycles II and III. Each universal timeline was divided into three segments: *beginning* (0–200 s), *middle* (200–600 s), and *toward the end* (600–800-s) (see **§ 3.4.4**). In Cycle II, the session began with 13 overlaps in the first 200 s, then witnessed 12 overlaps in the middle 400 s, with 13 overlaps in the final 200 s. Contrastingly, Cycle III started with only one overlap, with an increase to 7 overlaps during the *middle*, ending with 6 overlaps in the final segment. The Cycle comparison may suggest that the demand for searches is not consistent across different phases of the task. In Cycle II, overlaps were more frequent at the beginning and end, while in Cycle III, the beginning had nearly no overlaps. The impact of these overlaps on the quality of SI rendering in each phase warrants further investigation. It might be a matter of informants getting used to using InterpretBank (supported by the mostly even counts of overlaps), in Cycle II, and facing a somewhat easier source speech with a bit more experience in Cycle III (where searches would thus grow after a few minutes into the talk).

The overlaps of search actions and dropped chunks events highlight the challenges of multitasking for informants, especially when working with InterpretBank. Engagement with InterpretBank seems to increase task demands, requiring informants to exert high effort in processing multimodal data (auditory from source speeches, text from the monitor, and rendering delivery). To reduce effort, informants might opt to simply drop sentences, pointing to the complex dynamics of managing cognitive effort. The findings cannot be generalized but they contribute to our understanding of the complex interaction of cognitive abilities and tool use under pressure and time constraints in term-dense speeches.

4.8 Duration of source speech chunks and EVS

Longer source speech chunks may contain more syllables, and probably carry more information. Informants may be influenced both by sessions with a large amount of information and the time lag between the chunks. Results in this study highlight the relationship between chunk-initial ear-voice span (*EVS1*) and chunk-final ear-voice span (*EVS2*). The results (see **§ 3.2.9**) revealed a barely statistically significant negative correlation between the duration of the source speech and the *EVS1* and *EVS2* for both groups in the first two Cycles. However, in the third Cycle, the Excel group exhibited almost no correlation whereas the InterpretBank group maintained this inverse relationship. This lack of correlation in the Excel group during Cycle III raises questions about potential factors influencing their performance. It is possible that by this Cycle, informants in the Excel group had adapted to the information processing demands of term-dense speeches and did not have to cope with additional multitasking demands to handle InterpretBank.

The research line started by Fantinuoli & Montecchio (2023) determine maximum latency for ASR features was here extended to graphic information in view of the multimodal human communication. We took the notions of maximum delay of the machine and outlined human-computer interaction at both of beginning and at the end of each chunk, also drawing from Zhou *et al.* (2021) In our view, the frequent overlaps of searches and dropped chunks proved the approach useful. The observed inverse correlation between *EVS2* and the duration of the source speech in the first two Cycles for both groups and in Cycle III for the InterpretBank group may reveal that longer source speech durations led to shorter *EVS2*, perhaps because informants struggled to process previous information quickly enough to handle new input. This does not necessarily imply a loss of information or a halt in the middle of a sentence; rather, the shorter *EVS2* could be an active strategy employed by interpreters to manage their cognitive efforts. Such a strategy might involve selectively shortening *EVS2* to prioritize certain elements of the source speech. Even if informants dropped information, it might be because they suspected that the research was at least partially focused on term accuracy rather than overall quality, leading them to concentrate on term renditions.

4.9 Holistic assessment by raters

We opted for blind holistic assessments by third parties (see **§ 2.5.5**) and were not surprised to find very low inter-rater reliability, so we studied the potential relationship between customary quantitative indicators of quality and raters' judgments. To our surprise, no quantitative indicators correlated with the assessments of any rater, casting doubts on quantitative, narrow approaches to rendering quality when using CAI tools. Holistic assessments seem closer to those performed by the audience whereas the quantitative rubrics seem more useful in training environments. We hope to contribute to the ongoing discussion of what the best approach may be to assess quality in research scenarios. Here, however, we typical and new quantitative indicators complement our views on the impact of InterpretBank on the informants' behavior.

Hence, this study directly investigated raters' individual differences in their assessments of speeches and changes in main indicators. First, large variations emerged across raters. Luc, for instance, *consistently* adopted a stringent approach, skewing scores toward the lower end of the rating scale. Divergent emphases among raters suggest the need for further investigation, even when some variables did not reach statistical significance but showed moderate correlations.

Furthermore, the relationship between their assessments and quantitative indicators was also inconsistent. For instance, the indicator *filler* for Félix exhibited a strong positive correlation in the first cycle but reversed to a strong inverse correlation in the third Cycle. This fluctuation prompts questions about the evolving role of such quality indicators over time.

4.10 Summary

In this chapter, we discussed the findings related to the five hypotheses proposed in the introduction. The first hypothesis assessed the efficiency of term extraction with InterpretBank. Findings reveal that the InterpretBank group generally required less time for glossary compilation, produced more terms per individual glossary, and exhibited a lower time-per-term ratio compared to the Excel group. These findings confirm the first hypothesis. InterpretBank seems to reduce total glossary compilation time while it enhances efficiency. However, the term diversity in the InterpretBank group was lower, indicating a narrower range of unique terms in the individual glossary, which might suggest a reliance on automatically extracted terms from InterpretBank and less direct usefulness of the glossary, in that it does not respond to the particular needs of any individual.

The second hypothesis examines the impact of InterpretBank on fluency and accuracy in SI. The within-group analysis, using the Friedman test, showed no significant differences in fluency indicators (e.g., *false starts*, *self-corrections*, *fillers*, *repetitions, respites*) across cycles. The only exception was *bumps* in Cycle II for the InterpretBank group where informants produced more *bumps* than the Excel group. In terms of *EVS1* and *EVS2*, significant within-group differences were noted between Cycles possibly due to the outliers, but between-group comparisons did not show significant differences, indicating a minimal impact of InterpretBank on fluency. Therefore, fluency indicators did not support the second hypothesis.

Regarding accuracy indicators *(correct, adequate, correct renditions, and skipped terms)* for 39 potential problem triggers, the InterpretBank group showed better performance in correct renditions in Cycles II and III. However, no significant group differences were detected per Cycle. In the case of adequate renditions, while some significant within-group variations were observed across Cycles, there were no significant between-group differences across Cycles. This inconsistency in the data suggests the influence of individual informant behaviors—particularly from one informant, Blake. Wrong renditions and skipped terms did not exhibit significant differences in either within-group or between-group analyses, suggesting the need for further investigation into InterpretBank's impact on these aspects. So far, the findings from accuracy indicators did not support the second hypothesis.

Additionally, we also examined group performance in accurate (correct + adequate) and inaccurate (wrong + skipped) renditions. We observed that the average counts of accurate renditions tended to increase, while inaccurate renditions decreased from Cycle I to Cycle III for both groups. Despite this improvement, both groups initially had higher counts of inaccurate than accurate renditions, indicating the cognitive effort required in the delivery output in term-dense speeches. Analysis of 33 potential problem triggers also showed that InterpretBank's contribution to correct renditions was limited for repeated terms, with informants often relying on memory. However, some, like Gale, Kelly, and Jordan, consistently used InterpretBank's term retrieval.

The third hypothesis explored InterpretBank's impact on SI rendering quality. In this study, quality ratings were assigned on a scale from 1 to 6, with the InterpretBank group consistently receiving better evaluations than the Excel group across different cycles, as evident in the percentages shown in **Table 25**. The diminishing difference in ratings between the groups across cycles, particularly a smaller margin in Cycle III, suggests potential adaptation to task demands and varying cognitive strategies employed by the informants. Despite the low inter-rater agreement and the potential influence of various uncontrolled variables, the results broadly indicate that using InterpretBank positively impacts RSI rendering quality, supporting the study's third hypothesis.

The fourth hypothesis investigated whether improved documentation performance correlates with better RSI rendering quality. This hypothesis of the study speculated a connection between glossary compilation and RSI rendering quality. Analysis of documentation behavior indicators like *time per term* and *term diversity* in the Excel and InterpretBank groups revealed that, while InterpretBank generally led to lower *time per term* and higher *term counts*, it also resulted in lower *term diversity,* possibly due to its *automatic extraction* feature in InterpretBank. This finding indicates that InterpretBank supports efficiency, but that it might also lead to a certain homogenization of glossary terms. Additionally, individual variations in documentation behavior, such as fluctuating *times per term* for some informants, suggest that the correlation between these behaviors and interpreting quality is complex and not straightforward, highlighting the need for further investigation into individual differences. The master glossaries—including all terms from InterpretBank automatic term extraction plus entries added by at least two Excel informants, plus potential problem triggers—were reviewed and sometimes modified by the informants. Therefore, the only remaining differences were in their ways of compiling their glossaries and their possibly related memory traces. In other words, we tried to isolate the impact of glossary *compilation* on glossary *use*. However, the findings do not support the fourth hypothesis.

The fifth hypothesis concerned the continued use of InterpretBank based on improvements and attitudes. While InterpretBank was efficient in glossary compilation, survey results indicated varied opinions on its necessity and effectiveness for term retrieval. A follow-up survey a year later revealed that only a minority of informants (four informants out of eleven responses) continued using InterpretBank, primarily for glossary preparation with a few times, indicating limited employment in interpreting tasks. Therefore, our findings support the first and third hypotheses. However, the second, fourth, and fifth hypotheses are not supported by our findings.

In terms of individual performance, we observed variations in glossary compilation strategies and engagement with subtasks across Cycles. The primary focus was on intra-subject analysis and group performance, with a percentage-based comparison used to better understand the differences between Cycles for each participant. The results showed significant individual variations in the informants' allocation of time to different strategies and subtasks within glossary tasks, highlighting the diverse ways in which individuals interact with and utilize CAI tools. In Cycle I of the study, informants set a baseline for their glossary task, heavily relying on popular tools such as Google Translate, Youdao multilingual dictionary, Oulu App, and Baidu. These tools, especially Youdao and Oulu, which are preferred by Chinese users, offered features like automatic translations, bilingual examples, and English audio pronunciations. A significant amount of time was devoted to using search engines and bilingual dictionaries, with a marked preference for translations sourced from authentic bilingual contexts.

The study also incorporated the analysis of keystroke events and screen changes, employing ear-key spans (E2Ks) and eye-voice spans (I2Vs) to explore behaviors linked to cognitive events in the process of term retrieval and interpretation. The results indicated rapid response times and minimal variation, implying a continuous and seamless processing of behavior. Nonetheless, the observed variations in E2Ks and I2Vs, including the occurrence of negative values, suggest individual differences in behavioral patterns and the influence of task-specific factors. These findings underscore the prediction, multitasking, and dynamic interaction with CAI tools involved in live interpreting tasks, highlighting the complexity and multi-faceted nature of this field. The study found a statistically significant negative correlation between the duration of source speech chunks and ear-voice spans (*EVS1* and *EVS2*) in the first two Cycles for both groups, suggesting longer source speeches lead to shorter *EVS2* as informants may struggle to process information quickly. However, in the third cycle, this correlation diminished for the Excel group but persisted for the InterpretBank group, indicating potential adaptation to information processing demands and strategic management of cognitive load by selectively shortening *EVS2*.

This study evaluated the use of InterpretBank in rendering 33 terms and 6 repetitions across three cycles. While InterpretBank's usage varied across terms, a pattern emerged: certain repeated terms were *inaccurate renditions, but InterpretBank not used*, whereas others were *accurate renditions but InterpretBank not used.* This suggests individual differences in recall ability and a possible adaptation to InterpretBank over time. Despite the low overall frequency of InterpretBank usage for term repetitions, when used, it showed a high rate of success, indicating its effectiveness in handling repeated terms, yet informants did not commonly rely on it, probably due to active retention in memory resources.

The study revealed that the frequency of searches often matches or exceeds dropped chunks, suggesting a complex interplay of cognitive demands from source speeches and multitasking challenges working with InterpretBank. Informants like Alex, Blake, Dana, Erin, Frankie, and Gale showed varied search behaviors with InterpretBank, indicating individual differences and strategic adaptations to the tool's use, as evidenced by varying frequencies of searches and dropped chunks across different task phases and cycles.

In view of the importance of quality assessment in the present study, we took additional steps to support scientific rigor. The results revealed significant individual differences among raters in assessing speeches, with low inter-rater reliability indicating varied rating approaches, such as Luc's consistently stringent scoring. Additionally, the relationship between raters' assessments and quantitative indicators was inconsistent, exemplified by the fluctuating correlations for indicators like *filler* across different cycles, highlighting the need for further investigation into the evolving role of these quality indicators over time. We prudently think that the framework, methods, and constructs we used were, as a whole, adequate, and led to an informative result. In this sense, the methodology goals of this research project have modestly been reached.

Chapter 5

conclusion, limitations, and further research

The aim of this exploratory research project was to develop, adapt, and test research methods drawing from cognitive translatology (a situated cognition approach) to study interpreting. The project adopted a perspective of human-computer interaction to focus on aspects of remote simultaneous interpreting, and used InterpretBank as a paramount example of 3rd generation CAI tools (Prandi, 2023). Of particular interest was whether purported benefits and drawbacks remain the same when the users are trainees and when they have an L1 (Chinese) distant from the source language (English). The star features of InterpretBank, automatic retrieval of terms through voice recognition, were intentionally left out, to avoid too large a number of variables.

In order to have a full test bed that allows for comparisons with prior research projects on the use of CAI tools, a mixed methods study was designed and carried out that examined the behaviors of 22 Chinese interpreting trainees at glossary compilation and also performing remote simultaneous interpreting, both supported with InterpretBank features. Three cycles of data collection were separated one week from each other, and each one entailed compiling a glossary and performing RSI on a popular science speech ca 13 min each. After Cycle I, the sample was split into a control group, using Excel, and an experimental group, using InterpretBank. Data sources were keylogging, screen recording, SI output, and survey questionnaires: a sociodemographic survey to profile the informants, two questionnaires on tool use and self-assessment after data collection rounds, and a follow-up questionnaire to check on InterpretBank use one year later.

Even though it was an exploratory project, the study adopted a pre-test post-test design. Different from a typical confirmatory research project was that the treatment consisted of a training workshop whose contents differed: multimodal searching for the Excel group and an introduction to the InterpretBank features to be used for the experimental group. Recordings of both workshops were made available to all informants, after data collection was complete. Also different from typical pre-test post-test designs, there were two data collection rounds after the treatment, which differed in that the use of either Excel or InterpretBank was compulsory in Cycle II but free for informants to choose from in Cycle III. Data from Cycle I became the baseline benchmark for cycles II and III.

Keylogging data, screen recordings, and audio recordings of source speeches and SI outputs remotely collected from the informants' PCs and Apple computers were aligned and synchronized to the millisecond onto a universal time scale. This enabled comparative analyses of the informants' actions, based on their timings, durations, and placement. A battery of indicators for interpreting fluency and accuracy was also used to study the impact of InterpretBank and was cross-referenced with the holistic quality assessments of 5 or 3 raters. Our findings hold potential implications for various stakeholders, including users (professional interpreters, non-European language speakers, interpreting trainees), CAI developers, trainers, and researchers.

Glossary tasks were set up to observe how informants (in two groups) compiled individual glossaries. Key indicators for data analysis such as total time spent on glossary compilation, number of terms, time per term, and diversity rate were analyzed across Cycles I, II, and III for both Excel and InterpretBank groups. The InterpretBank group showed a lower diversity rate and a larger number of glossary entries compared to the Excel group, suggesting that the automatic extraction feature of InterpretBank might save time but lead to a more homogenized, larger set of glossary terms. Individual performance varied significantly, as seen in the case studies of Val (Excel group) and Alex (InterpretBank group), where the relationship between time spent per term and interpreting quality was inconsistent across cycles. As a reminder, we expected *read-first* glossaries (where glossary compilers read the original in various ways and mark, copy or type the entries as they spot them) leave stronger memory traces in the informants, and we used *time per term* as a proxy of focused attention fostering these traces, such that shorter times are assumed to hint at the weaker presence of terms in memory.

For the booth tasks based on the pre-treatment test (Cycle I) and post-treatment tests (Cycle II and Cycle III) and for most fluency indicators *(false starts, self-corrections, fillers, repetitions),* there were no significant differences both within and between the Excel and InterpretBank groups across the three cycles of the study. This conclusion is based on the Friedman test for within-group analysis and the Mann-Whitney U Test for inter-group analysis. However, in Cycles I and III, no significant differences were observed between the groups for this indicator. This suggests that, except for the noted instance in Cycle II, using either Excel or InterpretBank did not significantly compromise mental processing for the informants. Regarding the time cluster *(bumps, respites, EVS1 and EVS2),* both groups exhibited no significant differences in between-group comparisons, except for *bumps* in Cycle II where the InterpretBank group exhibited a higher median number of bumps compared to the Excel group. They also displayed similar patterns in the frequency of *bumps* and *respites.* The Excel group adjusted their strategies for *EVS1* and *EVS2* across the cycles, while the InterpretBank group showed longer durations of EVS1 and EVS2 than the Excel group in Cycle II due to two outliers. These variations did not show significant differences in most between-group comparisons. So InterpretBank's impact on fluency indicators, whether positive or negative, is limited.

In terms of accuracy indicators *(correct, adequate, wrong, and skipped),* both groups showed growth in the median number of correct renditions from Cycle I to Cycle III, with InterpretBank informants showing a more pronounced improvement. In Cycle I, the difference between the groups (all using Excel) was small, but by Cycle II, InterpretBank was introduced for the experimental group and surpassed Excel, a trend that widened in Cycle III. InterpretBank informants showed fewer adequate renditions compared to the Excel group across cycles. Regarding wrong renditions, both groups showed fluctuations across the cycles without significant statistical differences. For skipped terms, both groups exhibited a downward trend, with no significant differences between the groups.

The study also compared accurate and inaccurate renditions for 33 problem triggers. Both groups showed an increase in accurate renditions *(correct + adequate)* and a decrease in inaccurate renditions *(wrong + skipped)* over time. The results show that the average counts of inaccurate renditions were higher than those of accurate renditions. Since the number and the distribution of problem triggers were similar across cycles, the differences in accurate rand inaccurate renditions might be explained by informants actively trying to minimize inaccurate renditions while maximizing accurate renditions as they interpreted term-dense speeches. SI rendering quality was holistically assessed by five PhD raters. The InterpretBank group consistently received better quality ratings than the Excel group across cycles based on the average rating scores, hinting that using InterpretBank positively impacts RSI rendering quality.

Surveys conducted after Cycles II and III and a follow-up survey one year later assessed informants' opinions on InterpretBank regarding glossary preparation and term retrieval at RSI tasks. While initial surveys indicated satisfaction with InterpretBank's time-saving features for glossary tasks and term retrieval for booth tasks, actual usage data suggested discrepancies between positive survey responses and low frequency of practical tool usage in booth tasks. InterpretBank informants generally reported benefits in glossary compilation and expressed intent to continue using the tool for future booth tasks. However, the follow-up survey after one year revealed that seven out of eleven informants did not use InterpretBank after the study, citing reasons such as cost, lack of need in daily practice, and job requirements not involving interpreting.

5.1 General conclusions and implications

For interpreters, this study underscores that the successful use of CAI tools entails a significant learning curve and a training period of (probably) at least weeks. An expanding and diversifying market of remote simultaneous interpreting underscores the need for extensive information literacy skills (Drechsel, 2019) which minimally involve managing information from a diverse range of sources, efficiently and effectively retrieving information, critically and competently assessing it, and accurately and inventively using it to address specific issues or problems. Compiling glossaries before SI assignments can enhance the interpreters' ability to process information at the booth and use electronic resources (Jiang, 2013). Incorporating CAI tools into SI routines involves adapting to new workflows, learning sophisticated functions, and meeting larger multitasking demands. We have suggested some strategies to improve InterpretBank's effectiveness in version 8. On the other hand, using InterpretBank may also have an impact on cognitive styles of interpreters that might lead to higher technological dependence or reliance.

For interpreting trainees, technology may act as cognitive support and some applications may even become genuine "technologies of the extended mind" (Reiner & Nagel, 2017; O'Brien, 2023). Evidence shows that using InterpretBank leads to improvements in fluency, but intra-subject analysis reveals considerable individual variation. Several informants seemed to systematically search for repetitions as much as they did search for first-time terms. The early introduction of InterpretBank in training programs might implicitly encourage trainees to rely less on their WMs, fostering a dependency on CAI tools. That is, getting trainees not to use their WMs might induce a switch in their cognitive styles and even end with the often presumed and still debated enlarged WM capacity (Mellinger & Hanson, 2019; Ghiselli, 2022). Using InterpretBank, on the other hand, did not lead to an overall improvement in all indicators. For instance, there was a higher rate of correct renditions of potential problem triggers, but also more sentences were dropped whose source version contained those terms.

InterpretBank seems to have been primarily developed for European language users, and Chinese interpreters should approach translation suggestions with caution. For instance, traditional and simplified Chinese characters are not separate in InterpretBank. We may make an analogy of the difference as to that between old Blackletter or Gothic script (littera textualis or textus fractus) vs modern Roman or Latin script (this text). In the 1950s and 1960s, China introduced simplified characters, and education and contemporary written materials primarily employ simplified characters. In contrast, Taiwan and Hong Kong have consistently adhered to the use of traditional characters, also in education. That is, both are currently used. This is a common challenge. Search engines may not always distinguish effectively between Chinese and Japanese content, but a sophisticated professional tool for interpreters should.

In many cases, traditional and their corresponding simplified characters are pronounced the same. Crucially, beyond frequent characters, even very educated speakers may sometimes be unable to tell the nature of isolated characters, let alone know how to pronounce them. Recognizing a specialized term heard in the source language and pronouncing its rendition in the target language is a challenge. In the pair Chinese-English, it is unclear whether glossary compilation with the automatic term extraction feature actually saves time or instead simply dumps it onto other subtasks, such as searching for renditions and pronunciations. The meaning, relevance, and pronunciation of renditions need to be backed somehow.

Insights gained from logged behavior should help CAI tool developers create or improve applications that align with users' "actual behavior rather than their presumed behavior" (Dumais *et al.*, 2014). For instance, the workflows in **§ 4.7.1** demonstrated in CAI tools influence users' behaviors. Understanding user habits is crucial. For example, some Chinese users habitually do not switch to English character input before booth tasks.

More to the point, InterpretBank informants had larger and more homogeneous glossaries, because informants generally accepted most results from automatic term extraction. This is not necessarily good. A one-size-fits-all approach to the InterpretBank informants' needs obviates the user's language command and their domain knowledge. More entries might become distractors or noise—perhaps especially when the automatic recognition and retrieval function is at work, which was not the case. However, the larger the number of entries, the more chances there are of finding competing renditions to choose from. Fine-tuning the algorithm through machine learning is a must for specialized interpreters with accumulated experience.

As part of the data management of this study, we have also faced issues when importing files into InterpretBank from system directories with non-Roman characters, such as paths containing Chinese characters. This often resulted in errors and failed file imports due to clashes. A practical solution to this issue is to save the files in directories with paths exclusively in English, thereby avoiding the complications associated with non-Roman characters. But no information is to be found from InterpretBank sources, perhaps does not exist.

Many researchers have investigated didactic aspects of CAI tools (Amelina & Tarasenko, 2020; Prandi, 2020; Mellinger, 2023). Some research projects relied on quantitative rubrics and indicators and probably underestimated holistic assessment in evaluating CAI tool users' performance (Wang & Wang, 2019; Pisani & Fantinuoli, 2021; J. Zhang, 2021). Of course, holistic assessment may also be subject to the instructions and consensus scoring but we aimed at naturalistic assessments where such strategies that improve inter-rater reliability may be considered distorting. As a precaution, we chose raters to be interpreters who are PhD interpreting researchers (see **§ 2.5.5.1**). After all, holistic, intuitive assessment are exactly how users of interpreting services are going to assess their quality. Here the holistic assessments performed by relative specialists show a large variation, and precisely because of that they have significant implications for our understanding of human factors in assessing performance, when studying the impact of CAI tools.

There were no clear linear trends in all informants across cycles, raising questions about the nature of CAI-supported rendering assessment. Longitudinal studies are sorely needed, focused on CAI orientation and training, as well as integrated syllabi, before the real value of CAI tools can be settled. This study suggests that InterpretBank is indeed effective in improving term renditions accuracy, but unquantified variables such as prior knowledge, motivation, and attitude clearly very likely play a role as well. Most InterpretBank informants declared their intention of using InterpretBank in the future, but one year later only two really had used it, and not too often, considering that they are interpreting trainees. There may be many reasons for this, but these two only used InterpretBank to compile glossaries.

Finally, this study tested ways to explore human-computer interaction based on data from a cognitively-situated perspective. It is the reader—in particular, the viva committee—who will judge whether we succeeded in our goal of contributing to developing such perspectives to study interpreting. We modestly think that this project managed to suggest novel ways of using known tools, such as keylogging, to study interpreting behavior. Two constructs, *ear-key span,* and *eye-voice span*, proved useful to trace the interpreting trainees' information-seeking behaviors. Aligning different sources catered for a multimodal grid that portrayed the interaction of the informants with the tool. We further think that this approach to study some aspects of remote simultaneous interpreting might be extended to other aspects and tasks.

This methodological study adopted some ways of confirmatory designs, both to organize the complex quantitative threads of this multimethod project and to enhance rigor. Contradictory results are not discouraging; the goal was to formulate ways to apprehend the dynamics within a task, the interaction of the informants with their tools, and the environmental impacts on a task. Determining whether InterpretBank is useful in supporting the process and results of Chinese interpreting trainees was the testbed, rather than a primary goal by itself. All in all, our situated approach has proved to be able to yield a richer picture of the strategical behavior of the informants in their interaction with their environment to face some demands while pursuing their goals.

5.2 Limitations

The sample size in this study was small from the perspective of the numbers typical in other areas of Cognitive Science, but in CTIS there is no population with homogeneous characteristics to generalize any result. Rather, in CTIS we build knowledge by replicating tests and obtaining similar results in different populations, which is usually taken as a hint that such results *may be* a general case. For instance, in earlier CAI tools study, samples of informants were 22 MA participants in Xu (2018), six MA participants in Defrancq & Fantinuoli (2021), three professional interpreters in Fantinuoli *et al.* (2022), nine advanced interpreting students in Prandi (2023), and ten conference interpreters in Frittella (2023).

This study included informants from three Chinese MA programs where competition for registration is intense due to their reputed research, and part of the application process includes proving they have excellent L2 (English) command. In combining informants from several programs, I aimed to minimize differences stemming from the training in their home institutions, or a poor educational background. Informants from more average backgrounds might exhibit significantly different behaviors.

We focused on determining the potential advantages of using InterpretBank under naturalistic conditions (close to ideal, real conditions) by reasonably skilled informants, and the results cannot be generalized to *all* Chinese interpreting trainees. Here the ideal conditions are realistic and naturalistic, but not completely real and natural, from the moment the task was not real (it lacked an audience), and the informants were conscious of the experimental scenario, so the behavior observed in these interpreting trainees might not be exactly typical. For instance, in a real situation, the informants might have striven to render at least some of the sentences they dropped, perhaps enlarging the number of incorrect renditions.

Under these circumstances, the conclusions derived from the analysis apply to this testing environment, whose distortions did coincide with those usual in interpreter training programs. Working from home may be a professional feature, but doing it with no boothmate, on negotiated dates, and knowing you will be evaluated are all typical didactic conditions. There were no specific restrictions on the testing environment, as we did not want to intimidate participants. We used unobtrusive methods to allow participants to behave naturally for the sake of ecological validity. Potential biases that could influence this outcome include informants being aware of the purpose of the research, which could unconsciously move them to adjust their performance, especially as they became used to the task setting. We cannot exclude confounders like the effect of new tasks or individual prior knowledge. Therefore, while the data from our study hints at the potential influence of using InterpretBank on RSI task performance from English to Chinese, drawing very concrete conclusions remains a challenge.

The study employed various indicators and conducted observational log analysis based on basic statistics in order to comprehend the overall information-search and -management behavior of the informants over time, but such measures proved insufficient. We still lack precise hints at the underlying factors influencing each choice and strategy. The reasons behind each informant's emotions, interests, and attention shifts remain undisclosed. For instance, one informant accessed the glossary in Excel but did not perform any actions for five minutes; no mouse movement or keystrokes were observed. This could be due to the informant being uncertain about the next steps or using her mobile phone to look up words without being observed. We even did not know whether the informant was confused in the study. So, values alone cannot speak out the whole fact. Consequently, this cognitive analysis is in part an artifact and cannot avoid some bias, as we cannot turn down alternative reasons behind some observations.

To prompt informants' term retrieval needs and investigate how informants respond to problem triggers, we manipulated variables, such as the number of problem triggers, the speed of delivery, discourse features, and the length of the source speech. Although the text has been revised by an L1 English speaker who is both a conference interpreter and an interpreting trainer, ensuring its relevance for interpreting training, this design raises issues about the comparability of *artificial* texts applied in experimental settings. Yet many texts to be publicly delivered are *artificial* in that many a hand participates in their final form, they are often carefully worded, way above regular drafting, and increasingly computers help as well, from grammar and spell checking to actually write or revise excerpts.

Moreover, we adopted an intra-subject approach and group approaches, but there were more confounders in the environment, and we could not even know them all, let alone control them all, especially when we opted for remote collection and several cycles spanning one month. The complex setup forced us to keep source speeches in the same order, so there may have been order effects we could not mitigate, hence the third data collection cycle. In spite of the efforts to control them, there may be factors at work that require further research, such as the informants' familiarity with the topics, their comfort with their respective tools, and even the order and difficulty level of speeches in each cycle. Hence, the number of *fillers* observed in both groups across cycles could be attributed to several factors.

The difference in results between InterpretBank and Excel may result from other factors such as the participants' familiarity with the software or the nature and complexity of the speeches in Cycle II. Additionally, the participants' inherent cognitive abilities and prior training experience with RSI may also influence the frequency of respites. Furthermore, adapting to new tools takes time, and proficient use of a tool necessitates systematic, longer training. Consequently, the informant surveys only provide a partial view, reflecting immediate opinions (right after booth tasks in my case) rather than the complete range of perspectives. Additionally, InterpretBank's booth mode was generally perceived as a useful feature for mitigating stress among interpreting trainees. However, the extent to which it was relied upon varied individually over time, warranting further exploration into the psychological implications of technology use in this context.

5.3 Future work

This study was not seeking to confirm any truths but rather trying to determine if there are truths to be confirmed, whether they can be approached the ways we did, and how these ways could be refined in future approaches to consolidate a grasp on interpreting trainees' information seeking behavior when using CAI tools. However, we formally tested whether InterpretBank was good for Chinese interpreting trainees and found some food for thought as well.

An analysis of silent intervals in the speech flow can be useful to explore the cognitive processes behind *bumps* and *respites*—unintentional and unnoticed short gaps vs unintentional but potentially noticeable, longer gaps. InterpretBank informants tended to exhibit more bumps and respites than Excel informants, which may be attributed to the tool they used. Further investigation is needed to understand the cognitive circumstances leading to these bumps and respites. For instance, do InterpretBank informants display increased controlled attention to respites in the task or is the control of bumps and respites a function of regular, intuitive self-monitoring? Can measuring bumps and respites over time serve as a quantitative means to indirectly evaluate self-monitoring efficiency? Answering these questions will open new areas of discussion and chart pauses for interpreting strategies related to cognitive functions, particularly when trainees deal with novel terms.

Pronunciation-checking behaviors and their aspects require further investigation. Phonological knowledge is crucial in SI preparation, particularly in a technological work environment. Interpreters need to retain pronunciations in memory, for they play an important role in production. For instance, incorrect pronunciation of a name may result in the loss of relevant information and the offer of information different from that originally provided (Scaglioni, 2013). One of the typical features, supported by Chung's study (2023), is that experienced interpreters demonstrate better mastery than non-interpreters in processing phonological information when embedded in the text. Future research should examine the correlation between the accessibility of audio pronunciation in CAI tools and interpreting trainees' memory capacity for technical terms, assessing their correlation with term accuracy in interpreting tasks.

Attention to keyword-based searching (Dennis *et al.*, 2002)—actually, of typed-in-string searches—might reveal an uncharted area for discussing in detail how informants use keywords to navigate their online information sources. Different keyword combinations to explore the renditions of potential problem triggers may possibly reflect users' search skills or uncertainty, but it may also be a problem associated with the online resources a language has. As we walk down the peak of a handful of world languages, digital documentation loses part of its glitter.

Future research may alleviate the increasing difficulties for users by anticipating the relevance of selected keywords for an effective search (Azzopardi *et al.*, 2017). The situated observation of online information-seeking behavior may also reflect the users' strategies for keyword query formulation, including steps of identifying keywords and extracting relevant information from text to achieve search results that match the expected rendition of terms. Future work may lead to improved efficiency and accuracy in text searching while reducing cognitive efforts.

This research project also examined the screen recording and keylogged behavior log data across cycles, as categorized for this project. The disparities between pre- and post-treatment for individual participants were important. However, due to the limitations in time and scope, the trainee informants were not compared to professional interpreters as to their use of CAI tools. For instance, the study did not investigate how domain experts employ glossaries, resources, and strategies, compared to inexperienced trainees who lack knowledge in a specific domain. Interpreting directions within a language pair might also yield interesting differences and searching different but possibly systematic patterns of behavior in both directions may yield precious insight for CAI tool developers.

Moreover, the informants' approaches, sources, and tools for glossary compilation were quite varied even within the InterpretBank group. Isolating the impact of each one on the number of terms in individual glossaries is a real challenge, impossible with such a small sample of informants and recordings, and way beyond the scope of this project. The question remains whether specific tools other than InterpretBank, such as TermBank, also foster larger glossaries or if other factors also play a role, such as investigating individual search patterns. Future research should aim to explore these aspects, potentially uncovering insights into how tool choice impacts the quality and comprehensiveness of glossaries for RSI tasks.

This study combined multimodal data collection methods involving keylogging, screen recording, and SI renditions recording, with a special focus on keylogging as an unobtrusive way to register informants' observational behavior during tasks. To our knowledge, at the time of writing this dissertation, this is the first empirical study to incorporate it into interpreting research. Keylogging behavior is unobtrusive. In this case, it was solely based on keydown events. This was a technical requirement of the equipment but, introducing keyup events and hence telling apart the duration of keypresses from those of the time gaps between keypresses (i.e., inter keystroke intervals) in future studies, might let us tap from knowledge accumulated by translation researchers regarding revisions, planning, certainty, and decision making.

The number of terms and repetitions in this mini-test was limited, the observed recall effects cannot be generalized but are certainly food for thought. Automatic term recognition and retrieval may be increasingly felt as a way to put words in the mouth of an interpreter and be felt increasingly annoying with repetitions. This is definitely worth further study. In general, some CAI tool developers might benefit from studying what interpreters do and then coming up with ways to support interpreters in what they do, rather than coming up with something computers can do and then seeing if interpreters benefit from that.

The combined results of the Excel and InterpretBank groups concerning both *first-time* and *repeated* terms (*rep1* and *rep2*) indicate that the choice of tool impacts rendering quality. This is likely because informants must adapt to the specific workflow of InterpretBank, thereby altering their cognitive strategies and focused attention. The decline in the InterpretBank group during Cycle III suggests that there is room for improvement in the tool. Future research should explore the learning curve of CAI tools and user adaptations because this study found hints that it may affect the trainees' cognitive style to approach the task when it is under development. Future studies might seek to examine whether long-term use and familiarity with CAI tools mitigate these effects.

references

Aldridge, M., & Fontaine, L. (2022). Keystroke logging data: What can it tell us about mode and written language production? In E. Asp & M. Aldridge (Eds.), *Empirical Evidences and Theoretical Assumptions in Functional Linguistics* (1st ed., pp. 37–59). Routledge. https://doi.org/10.4324/9780429031427

Altarriba, J., & Basnight-Brown, D. M. (2007). Methodological considerations in performing semantic- and translation-priming experiments across languages. *Behavior Research Methods*, *39*(1), 1–18. https://doi.org/10.3758/BF03192839

Amelina, S., & Tarasenko, R. (2020). *Using Modern Simultaneous Interpretation Tools in the Training of Interpreters at Universities* (A. Bollin, H. C. Mayr, A. Spivakovsky, M. Tkachuk, V. Yakovyna, A. Yerokhin, & G. Zholtkevych, Eds.; Vol. 1, pp. 188–201). CEUR-WS.org. http://ceur-ws.org/Vol-2740/20200188.pdf

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Angelone, E. (2021). Broadening the scope of error categories in translation assessment through screen recording. *Across Languages and Cultures*, *22*(2), 143–157. https://doi.org/10.1556/084.2021.00021

Annalisa, S. (2015). Becoming an interpreter: The role of computer technology. *MonTI. Monografías de Traducción e Interpretación*, 111–138. https://doi.org/10.6035/MonTI.2015.ne2.4

Anthony, L. (2022). *AntConc* (https://www.laurenceanthony.net/software; 4.1.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Atabekova, A. A., Gorbatenko, R. G., Shoustikova, T. V., & Valero-Garcés, C. (2018). Cross-cultural mediation with refugees in emergency settings: ICT use by language service providers. *Journal of Social Studies Education Research*, *9*(3), Article 3. https://jsser.org/index.php/jsser/article/view/274

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (Vol. 2, pp. 89–195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Azzopardi, J., Benedetti, F., Guerra, F., & Lupu, M. (2017). Back to the Sketch-Board: Integrating Keyword Search, Semantics, and Information Retrieval. In A. Calì, D. Gorgan, & M. Ugarte (Eds.), *Semantic Keyword-Based Search on Structured Data Sources* (pp. 49–61). Springer International Publishing. https://doi.org/10.1007/978-3-319-53640-8_5

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Ballier, N., Pacquetet, E., & Arnold, T. (2019). Investigating Keylogs as Time-Stamped Graphemics. *Proceedings of Graphemics in the 21st Century, Brest 2018*, 353–365. https://doi.org/10.36824/2018-graf-ball

Bao, L., Qian, Z., & Zhang, Q. (2023). The multiple phonological activation in Chinese spoken word production: An ERP study supporting cascaded model. *Behavioural Brain Research*, *451*, 114523. https://doi.org/10.1016/j.bbr.2023.114523

Barik, H. C. (1973). Simultaneous Interpretation: Temporal and Quantitative Data. *Language and Speech*, *16*(3), 237–270. https://doi.org/10.1177/002383097301600307

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 1313–1316. http://www.lrec-conf.org/proceedings/lrec2004/pdf/509.pdf

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2008). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Biagini, G. (2015). *Glossario cartaceo e glossario elettronico durante l'interpretazione simultanea: Uno studio comparativo [Printed glossary and electronic glossary in simultaneous interpretation: A comparative study]* [Università degli studi di Trieste.]. https://www.academia.edu/23759751/Glossario_cartaceo_e_glossario_elettronico_durante_linterpretazione_simultanea_uno_studio_comparativo

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (6.4) [Computer software]. http://www.praat.org/

Bowen, N. E. J. A., & Thomas, N. (2020). Manipulating texture and cohesion in academic writing: A keystroke logging study. *Journal of Second Language Writing*, *50*, 100773. https://doi.org/10.1016/j.jslw.2020.100773

Bower, K. (2015). Stress and Burnout in Video Relay Service (VRS) Interpreting. *Journal of Interpretation*, *24*(1), 18.

Brice, A. (2021). Interactive Language Development. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science* (pp. 4178–4181). Springer International Publishing. https://doi.org/10.1007/978-3-319-19650-3_1351

Cabral, G. G., & Minku, L. L. (2023). Towards Reliable Online Just-in-Time Software Defect Prediction. *IEEE Transactions on Software Engineering*, *49*(3), 1342–1358. https://doi.org/10.1109/TSE.2022.3175789

Cacciamani, L., Mojica, A. J., Sanguinetti, J. L., & Peterson, M. A. (2012). Meaning can be Accessed for the Groundside of a Figure. *Journal of Vision*, *12*(9), 305–305. https://doi.org/10.1167/12.9.305

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5), 489–503. https://doi.org/10.1023/A:1008084120205

Calatrava, M., De Irala, J., Osorio, A., Benítez, E., & Lopez-del Burgo, C. (2022). Matched and Fully Private? A New Self-Generated Identification Code for School-Based Cohort Studies to Increase Perceived Anonymity. *Educational and Psychological Measurement*, *82*(3), 465–481. https://doi.org/10.1177/00131644211035436

Carlson, B. W. (2024). Simpson's paradox. In *Encyclopedia Britannica*. https://www.britannica.com/topic/Simpsons-paradox

Chen, H., Wang, Y., & Brown, T. P. (2021). The effects of topic familiarity on information completeness, fluency, and target language quality of student interpreters in Chinese–English consecutive interpreting. *Across Languages and Cultures*, *22*(2), 176–191. https://doi.org/10.1556/084.2021.00013

Chen, J., Yang, H., & Han, C. (2022). Holistic versus analytic scoring of spoken-language interpreting: A multi-perspectival comparative analysis. *The Interpreter and Translator Trainer*, 1–19. https://doi.org/10.1080/1750399X.2022.2084667

Chen, S., & Kruger, J.-L. (2023). The effectiveness of computer-assisted interpreting: A preliminary study based on English-Chinese consecutive interpreting. *Translation and Interpreting Studies*, *18*(3), 399–420. https://doi.org/10.1075/tis.21036.che

Chicca, J. (2022). Screencasts as a way to enhance online learning environments in nursing. *Teaching and Learning in Nursing*, *17*(1), 130–131. https://doi.org/10.1016/j.teln.2021.07.007

Chiocchetti, E., Lušicky, V., & Wissik, T. (2023). Terminology standards and their relevance for legal interpreters and translators: Results of an exploratory study in Austria and Italy. *Digital Translation*, *10*(2), 156–179. https://doi.org/10.1075/dt.00006.chi

Chmiel, A. (2018). Meaning and words in the conference interpreter's mind: Effects of interpreter training and experience in a semantic priming study. *Translation, Cognition & Behavior*, *1*(1), 21–41. https://doi.org/10.1075/tcb.00002.chm

Chmiel, A., & Lijewska, A. (2022). Reading patterns, reformulation and eye-voice span (IVS) in sight translation. *Translation and Interpreting Studies*. https://doi.org/10.1075/tis.21021.chm

Cho, H., & Hirst, D. (2006). The contribution of silent pauses to the perception of prosodic boundaries in Korean read speech. *Proceedings of Speech Prosody 2006*, 185–188. https://doi.org/10.21437/SpeechProsody.2006-28

Christoffels, I. K., & De Groot, A. M. B. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, *7*(3), 227–240. https://doi.org/10.1017/S1366728904001609

Chung, H.-Y. (2023). Context-embedded phonological memory in interpreters. *Lebende Sprachen*, *68*(1), 75–95. https://doi.org/10.1515/les-2022-1030

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. JSTOR. http://www.jstor.org.ezproxy.unibo.it/stable/3328150

Collados Aís, Á. (2016). *Quality assessment and intonation in simultaneous interpreting: Evaluation patterns*. https://doi.org/10.6035/MonTI.2016.ne3.8

Conde Ruano, J. T. (2009). *Proceso y resultado de la evaluación de traducciones* [Http://purl.org/dc/dcmitype/Text, Universidad de Granada]. https://dialnet.unirioja.es/servlet/tesis?codigo=69150

Conde Ruano, J. T. (2012a). Quality and quantity in translation evaluation: A starting point. *Across Languages and Cultures*, *13*(1), 67–80. https://doi.org/10.1556/Acr.13.2012.1.4

Conde Ruano, J. T. (2012b). The Good Guys and the Bad Guys: The Behavior of Lenient and Demanding Translation Evaluators. *Meta*, *57*(3), 763–786. https://doi.org/10.7202/1017090ar

Corpas Pastor, G. (2016). *A Survey of Interpreters' Needs and Practices Related to Language Technology* (FFI2012-38881-MINECO/TI-DT-2016–1; p. 60). Málaga: University of Málaga.

Corpas Pastor, G. (2021). Interpreting and Technology: Is the Sky Really the Limit? *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*, 15–24. https://doi.org/10.26615/978-954-452-071-7_003

Corpas Pastor, G. (2022). Interpreting Tomorrow? How to Build a Computer-Assisted Glossary of Phraseological Units in (Almost) No Time. In G. Corpas Pastor & R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 62–77). Springer International Publishing.

Costa, H., Pastor, G. C., & Durán-Muñoz, I. (2017). Assessing Terminology Management Systems for Interpreters. In G. Corpas Pastor & I. Durán-Muñoz (Eds.), *Chapter 3: Assessing Terminology Management Systems for Interpreters* (pp. 57–84). Brill. https://doi.org/10.1163/9789004351790_005

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191. APA PsycArticles. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (1st ed., pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2000). Processing limits of selective attention and working memory: Potential implications for interpreting. *Interpreting*, *5*(2), 117–146. https://doi.org/10.1075/intp.5.2.05cow

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. Cambridge Core. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2005). *Working memory capacity*. Psychology Press.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Towse, J. N., Hamilton, Z., Saults, J. S., Elliott, E. M., Lacey, J. F., Moreno, M. V., & Hitch, G. J. (2003). Children's working-memory processes: A response-timing analysis. *Journal of Experimental Psychology: General*, *132*(1), 113–132. https://doi.org/10.1037/0096-3445.132.1.113

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, *43*(1), 19–37. https://doi.org/10.1080/03054985.2016.1232245

Cui, Y., & Zheng, B. (2022). Extralinguistic Consultation in English–Chinese Translation: A Study Drawing on Eye-Tracking and Screen-Recording Data. *Frontiers in Psychology*, *13*, 891997. https://doi.org/10.3389/fpsyg.2022.891997

Davoudi, M., & Moghadam, H. R. H. (2015). Critical Review of the Models of Reading Comprehension with a Focus on Situation Models. *International Journal of Linguistics*, *7*(5), 172–187. https://doi.org/10.5296/ijl.v7i5.8357

De Groot, A. M. B. (2000). A Complex-skill Approach to Translation and Interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting* (Vol. 37, pp. 53–68). John Benjamins Publishing Company. https://doi.org/10.1075/btl.37.06gro

de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162

Defrancq, B., & Fantinuoli, C. (2021). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target. International Journal of Translation Studies*, *33*(1), 73–102. https://doi.org/10.1075/target.19166.def

Dennis, S., Bruza, P., & McArthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, *53*(2), 120–133. https://doi.org/10.1002/asi.10015

Díaz-Galaz, S. (2011). The effect of previous preparation in simultaneous interpreting: Preliminary results. *Across Languages and Cultures*, *12*(2), 173–191. https://doi.org/10.1556/Acr.12.2011.2.3

Díaz-Galaz, S. (2020). Listening and comprehension in interpreting: Questions that remain open. *Translation and Interpreting Studies*, *15*(2), 304–323. https://doi.org/10.1075/tis.20074.dia

Díaz-Galaz, S., & Torres, A. (2019). Comprehension in interpreting and translation: Testing the phonological interference hypothesis. *Perspectives*, *27*(4), 622–638. https://doi.org/10.1080/0907676X.2019.1569699

Dijk, T. A. : van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press. http://ezproxy.unibo.it/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05251a&AN=at.UBO0421910&site=eds-live&scope=site

Direnga, J., Timmermann, D., Lund, J., & Kautz, C. (2016, September 12). *Design and Application of Self-Generated Identification Codes (SGICs) for Matching Longitudinal Data*. 44th Annual Conference of the European Society for Engineering Education - Engineering Education on Top of the World: Industry-University Cooperation, Tampere, Finland. https://www.sefi.be/proceeding-author/j-direnga/

Doherty, S. (2020). Multisensory integration in audiovisual translation. In *Multilingual Mediated Communication and Cognition* (pp. 155–170). Routledge. https://doi.org/10.4324/9780429323867-7

Dong, Y., & Li, P. (2020). Attentional control in interpreting: A model of language control and processing control. *Bilingualism: Language and Cognition*, *23*(4), 716–728. https://doi.org/10.1017/S1366728919000786

Drechsel, A. (2019). Technology literacy for the interpreter. In D. B. Sawyer, F. Austermühl, & V. Enríquez Raído (Eds.), *American Translators Association Scholarly Monograph Series: Vol. XIX* (pp. 259–268). John Benjamins Publishing Company. https://doi.org/10.1075/ata.xix.12dre

Du, Z., & Muñoz Martín, R. (in preparation). *Holistic Evaluation Methods for Simultaneous Interpreting Quality Assessment*.

Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding User Behavior Through Log Data and Analysis. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 349–372). Springer New York. https://doi.org/10.1007/978-1-4939-0378-8_14

Dyreson, C. E., & Snodgrass, R. T. (1993). Timestamp semantics and representation. *Information Systems*, *18*(3), 143–166. https://doi.org/10.1016/0306-4379(93)90034-X

Elgort, I., Wetering, R. V. D., Arrow, T., & Beyersmann, E. (2023). Previewing Novel Words Before Reading Affects Their Processing During Reading: An Eye-Movement Study With First and Second Language Readers. *Language Learning*, 1–33. https://doi.org/10.1111/lang.12579

Elmer, S., & Giroud, N. (2023). Simultaneous interpreting, brain aging, and cognition: A review and future directions. *Translation, Cognition & Behavior*, *6*(2), 118–140. https://doi.org/10.1075/tcb.00082.elm

Enríquez Raído, V. (2013). *Translation and Web Searching* (0 ed.). Routledge. https://doi.org/10.4324/9780203798034

Enríquez Raído, V., & Cai, Y. (2023). Changes in web search query behavior of English-to-Chinese translation trainees. *Ampersand*, *11*, 100137. https://doi.org/10.1016/j.amper.2023.100137

Enríquez-Raído, V. (2013). Using Screen Recording as a Diagnostic Tool in Early Process-oriented Translator Training. In S. Hansen-Schirra, D. Kiraly, & K. Maksymski (Eds.), *New prospects and perspectives for educating language mediators / Don Kiraly, Silvia Hansen-Schirra, Karin Maksymski (eds.)* (pp. 121–138). Narr. http://diglib.cib.unibo.it/orti.php?id=BID_4264910

Erazo-Toscano, R., & Osan, R. (2023). Synaptic propagation in neuronal networks with finite-support space-dependent coupling. *Physical Review E*, *107*(3), 034403. https://doi.org/10.1103/PhysRevE.107.034403

Eysenck, M. W., & Keane, M. T. (2020). *Cognitive Psychology: A Student's Handbook* (8th ed.). Psychology Press. https://doi.org/10.4324/9781351058513

Fantinuoli, C. (2017). Computer-assisted preparation in conference interpreting. *The International Journal of Translation and Interpreting Research*, *9*(2), 24–37. https://doi.org/10.12807/ti.109202.2017.a02

Fantinuoli, C. (2023). Towards AI-enhanced computer-assisted interpreting. In G. Corpas Pastor & B. Defrancq (Eds.), *IVITRA Research in Linguistics and Literature* (Vol. 37, pp. 46–71). John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.37.03fan

Fantinuoli, C., Marchesini, G., Landan, D., & Horak, L. (2022). KUDO Interpreter Assist: Automated Real-time Support for Remote Interpretation. *Proceedings of Translating and the Computer 43*, 68–77. https://doi.org/10.48550/arXiv.2201.01800

Fantinuoli, C., & Montecchio, M. (2023). Defining maximum acceptable latency of AI-enhanced CAI tools. In Ó. Ferreiro-Vázquez, A. T. Varajão Moutinho Pereira, & S. L. Gonçalves Araújo (Eds.), *Technological Innovation Put to the Service of Language Learning, Translation and Interpreting: Insights from Academic and Professional Contexts* (pp. 213–225). Peter Lang Verlag. https://doi.org/10.3726/b20168

Foster, M., & Scheinost, D. (2024). Brain states as wave-like motifs. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2024.03.004

Frittella, F. M. (2022). CAI Tool-Supported SI of Numbers: A Theoretical and Methodological Contribution. *International Journal of Interpreter Education*, *14*(1), 32–56. https://doi.org/10.34068/ijie.14.01.05

Frittella, F. M. (2023). Usability research for interpreter-centred technology. In *Language Science Press*. Language Science Press. https://doi.org/10.5281/zenodo.7376351

Frittella, F. M., & Rodríguez, S. (2022). Putting SmartTerp to Test: A tool for the challenges of remote interpreting. *INContext: Studies in Translation and Interculturalism*, *2*(2), Article 2. https://doi.org/10.54754/incontext.v2i2.21

Fulford, H. (2001). Exploring terms and their linguistic environment in text: A domain-independent approach to automated term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *7*(2), 259–279. https://doi.org/10.1075/term.7.2.08ful

Gaber, M., Corpas Pastor, G., & Omer, A. (2020). Speech-to-Text Technology as a Documentation Tool for Interpreters: A new approach to compiling an ad hoc corpus and extracting terminology from video-recorded speeches. *TRANS. Revista de Traductología*, *24*, 263–281. https://doi.org/10.24310/TRANS.2020.v0i24.7876

Ge, T. (2023). *Usability of Terminology—Assistance in Chinese to English Simultaneous Interpretation—Taking interpretBanias an Example* [Master's Thesis, Beijing Foreign Studies University]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFDTEMP&filename=1023063456.nh&v=

Gervits, F., Johanson, M., & Papafragou, A. (2023). Relevance and the Role of Labels in Categorization. *Cognitive Science*.

Ghiselli, S. (2022). Working memory tasks in interpreting studies: A meta-analysis. *Translation, Cognition & Behavior*. https://doi.org/10.1075/tcb.00063.ghi

Gieshoff, A. C., & Schuler, M. (2022). *The augmented interpreter: A pilot study on the use of augmented reality in interpreting*. 3rd HKBU International Conference on Interpreting : Interpreting and Technology, Hong-Kong (online), 7-9 December 2022. https://digitalcollection.zhaw.ch/handle/11475/26724

Gilbert, A. S., Croy, S., Hwang, K., LoGiudice, D., & Haralambous, B. (2022). Video remote interpreting for home-based cognitive assessments: Stakeholders' perspectives. *Interpreting. International Journal of Research and Practice in Interpreting*, *24*(1), 84–110. https://doi.org/10.1075/intp.00065.gil

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649–683. https://doi.org/10.1080/01690960500181332

Goldsmith, J. (2023). Tablet interpreting: A decade of research and practice. In G. Corpas Pastor & B. Defrancq (Eds.), *IVITRA Research in Linguistics and Literature* (Vol. 37, pp. 27–45). John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.37.02gol

Gough, J. (2023). Individual variations in information behaviour of professional translators: Towards a classification of translation-oriented research styles. *Translation Studies*, 1–22. https://doi.org/10.1080/14781700.2023.2231933

Grinschgl, S., & Neubauer, A. C. (2022). Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.908261

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, 20.

Groh, J. M., & Gazzaniga, M. S. (2003). How the Brain Keeps Time. *Daedalus*, *132*(2), 56–61. JSTOR. http://www.jstor.org.ezproxy.unibo.it/stable/20027840

Guo, M., & Han, L. (2024). From manual to machine: Evaluating automated ear–voice span measurement in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*. https://doi.org/10.1075/intp.00100.guo

Guo, M., Han, L., & Anacleto, M. T. (2022). Computer-Assisted Interpreting Tools: Status Quo and Future Trends. *Theory and Practice in Language Studies*, *13*(1), 89–99. https://doi.org/10.17507/tpls.1301.11

Han, C. (2022a). Assessing spoken-language interpreting: The method of comparative judgement. *Interpreting. International Journal of Research and Practice in Interpreting*, *24*(1), 59–83. https://doi.org/10.1075/intp.00068.han

Han, C. (2022b). Interpreting testing and assessment: A state-of-the-art review. *Language Testing*, *39*(1), 30–55. https://doi.org/10.1177/02655322211036100

Han, C., & Yang, L. (2023). Relating utterance fluency to perceived fluency of interpreting: A partial replication and a mini meta-analysis. *Translation and Interpreting Studies*, *18*(3), 421–447. https://doi.org/10.1075/tis.20091.han

Harley, T. A. (2014). *The psychology of language: From data to theory* (Fourth edition). Psychology Press, Taylor & Francis Group.

Hayashi, M., & Yoon, K.-E. (2010). A cross-linguistic exploration of demonstratives in interaction: With particular reference to the context of word-formulation trouble: In N. Amiridze, B. Davis, & M. Maclagan (Eds.), *Fillers, Pauses and Placeholders* (pp. 33–66). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.93.03hay

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, *14*(3), 577–598. https://doi.org/10.1007/s11097-014-9355-1

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, *3*. https://doi.org/10.3389/neuro.09.031.2009

Hinbarji, Z., Albatal, R., O'Connor, N., & Gurrin, C. (2016). LoggerMan, a Comprehensive Logging and Visualization Tool to Capture Computer Usage. In Q. Tian, N. Sebe, G.-J. Qi, B. Huet, R. Hong, & X. Liu (Eds.), *MultiMedia Modeling* (pp. 342–347). Springer International Publishing.

Ho, C.-E. (2021). What does professional experience have to offer? An eyetracking study of sight interpreting/translation behaviour. *Translation, Cognition & Behavior*, *4*(1), 47–73. https://doi.org/10.1075/tcb.00047.ho

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196. https://doi.org/10.1145/353485.353487

Holm, H., Skein, E., & Sullivan, K. P. H. (2022). Using computer keystroke logging in the second language composition classroom. In J. Qin & P. Stapleton, *Technology in Second Language Writing* (1st ed., pp. 167–181). Routledge. https://doi.org/10.4324/9781003279358-11

Huang, J., & Wang, J. (2022). Post-editing machine translated subtitles: Examining the effects of non-verbal input on student translators' effort. *Perspectives*, 1–21. https://doi.org/10.1080/0907676X.2022.2026424

Huseynov, H., Kourai, K., Saadawi, T., & Igbe, O. (2020). Virtual Machine Introspection for Anomaly-Based Keylogger Detection. *2020 IEEE 21st International Conference on High Performance Switching and Routing (HPSR)*, 1–6. https://doi.org/10.1109/HPSR48589.2020.9098980

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*(3), 265–288. https://doi.org/10.1016/0364-0213(95)90020-9

Hvelplund, K. T. (2019). Digital resources in the translation process – attention, cognitive effort and processing flow. *Perspectives*, *27*(4), 510–524. https://doi.org/10.1080/0907676X.2019.1575883

Jakobsen, A. L. (1999). Logging target text production with Translog. In G. Hansen (Ed.), *Probing the process in translation: Methods and results (Copenhagen Studies in Language 24)* (pp. 9–20). Samfundslitteratur. https://research.cbs.dk/en/publications/logging-target-text-production-with-translog

Jakobsen, A. L., & Jensen, K. T. H. (2008). Eye Movement Behaviour Across Four Different Types of Reading Task. *Copenhagen Studies in Language*, *36*, 103–124. https://research.cbs.dk/en/publications/eye-movement-behaviour-across-four-different-types-of-reading-tas

Jiang, H. (2013). The interpreter's glossary in simultaneous interpreting: A survey. *Interpreting. International Journal of Research and Practice in Interpreting*, *15*(1), 74–93. https://doi.org/10.1075/intp.15.1.04jia

Jong, N. H. D., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *The 6th Workshop on Disfluency in Spontaneous Speech (DISS)*, *17–20*.

Keevallik, L. (2010). The interactional profile of a placeholder: The Estonian demonstrative see. In N. Amiridze, B. Davis, & M. Maclagan (Eds.), *Fillers, Pauses and Placeholders* (pp. 139–172). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.93.07kee

Kintsch, W. (1998). Learning from text. In *Comprehension: A paradigm for cognition* (pp. 282–331). Cambridge University Press.

Kintsch, W. (2013). Revisiting the Construction–Integration Model of Text Comprehension and Its Implications for Instruction. In D. E. Alvermann, N. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (Sixth edition, pp. 807–839). International Reading Association.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Krickel, B. (2023). Extended Cognition and the Search for the Mark of Constitution – A Promising Strategy? In M.-O. Casper & G. F. Artese (Eds.), *Situated Cognition Research: Methodological Foundations* (pp. 129–146). Springer International Publishing. https://doi.org/10.1007/978-3-031-39744-8_8

Kuang, H., & Zheng, B. (2023). Note-taking effort in video remote interpreting: Effects of source speech difficulty and interpreter work experience. *Perspectives*, *31*(4), 724–744. https://doi.org/10.1080/0907676X.2022.2053730

Kukreja, U., Stevenson, W. E., & Ritter, F. E. (2006). RUI: Recording user input from interfaces under Windows and Mac OS X. *Behavior Research Methods*, *38*(4), 656–659. https://doi.org/10.3758/BF03193898

Kurz, I., Liu, M., Shlesinger, M., & Obler, L. K. (2011). *Interpreting work buffers against aging? Reporting on the AIIC Lifespan Study*. aiic.net

Lee, T.-H. (2004). Ear Voice Span in English into Korean Simultaneous Interpretation. *Meta*, *47*(4), 596–606. https://doi.org/10.7202/008039ar

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Levering, K. R., & Kurtz, K. J. (2019). Concepts: Structure and Acquisition. In R. J. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought: An Introduction* (pp. 55–70). Heidelberg University Publishing (heiUP). https://doi.org/10.17885/heiup.470

Lewis, J. D. (2022, July 28). *Using Techsmith Capture videos to analyze a student's information seeking behaviors*. https://rrpress.utsa.edu/handle/20.500.12588/1112

Li, S., Wang, Y., & Rasmussen, Y. Z. (2022). Studying interpreters' stress in crisis communication: Evidence from multimodal technology of eye-tracking, heart rate and galvanic skin response. *The Translator*, *28*(4), 468–488. https://doi.org/10.1080/13556509.2022.2159782

Li, Y., Breithaupt, F., Hills, T., Lin, Z., Chen, Y., Siew, C. S. W., & Hertwig, R. (2024). How cognitive selection affects language change. *Proceedings of the National Academy of Sciences*, *121*(1). https://doi.org/10.1073/pnas.2220898120

Liu, J. (2022). The Impact of Technology on Interpreting: An Interpreter and Trainer's Perspective. *International Journal of Chinese and English Translation & Interpreting*. https://doi.org/10.56395/ijceti.v1i1.14

Liu, M., Kurz, I., Moser-Mercer, B., & Shlesinger, M. (2020). The interpreter's aging: A unique story of multilingual cognitive decline? *Translation, Cognition & Behavior*, *3*(2), 287–309. https://doi.org/10.1075/tcb.00045.liu

Lovallo, W. R. (2016). Stress and Health: Biological and Psychological Interactions. In *Stress and Health: Biological and Psychological Interactions* (Third Edition, pp. 115–136). SAGE Publications, Inc. https://doi.org/10.4135/9781071801390

Lu, S., Xiangling, W., & Shuya, M. (2022). Investigating the relationship between online information seeking and translation performance among translation students: The mediating role of translation self-efficacy. *Frontiers in Psychology*, *13*, 944265. https://doi.org/10.3389/fpsyg.2022.944265

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Lung, R. (2005). On the history of interpreting in China. *Perspectives*, *13*(2), 143–150. https://doi.org/10.1080/09076760508668983

M. Shreve, G., Angelone, E., & Lacruz, I. (2014). Efficacy of screen recording in the other-revision of translations: Episodic memory and event models. *MonTI. Monografías de Traducción e Interpretación*, 225–245. https://doi.org/10.6035/MonTI.2014.ne1.7

Madore, K. P., Khazenzon, A. M., Backes, C. W., Jiang, J., Uncapher, M. R., Norcia, A. M., & Wagner, A. D. (2020). Memory failure predicted by attention lapsing and media multitasking. *Nature*, *587*(7832), 87–91. https://doi.org/10.1038/s41586-020-2870-z

Matis, N. (2010). Terminology Management during Translation Projects: Professional Testimony. *Linguaculture*, *1*(1), 107–115. https://doi.org/10.47743/lincu-2010-1-0226

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287–330. https://doi.org/10.1037/0033-295X.86.4.287

Megyesi, B., & Gustafson-Capková, S. (2002). Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish. *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2153–2156. https://doi.org/10.21437/ICSLP.2002-588

Mellinger, C. D. (2015). On the applicability of Internet-mediated research methods to investigate translators' cognitive behaviour. *Translation & Interpreting*, *7*(1), 59–71.

Mellinger, C. D. (2023). Incorporating Translation and Interpreting into the Business Language Classroom. *Global Business Languages*, *23*, 58–73. https://doi.org/10.4079/gbl.v23.5

Mellinger, C. D., & Hanson, T. A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge.

Mellinger, C. D., & Hanson, T. A. (2018). Interpreter traits and the relationship with technology and visibility. *Translation and Interpreting Studies*, *13*(3), 366–392. https://doi.org/10.1075/tis.00021.mel

Mellinger, C. D., & Hanson, T. A. (2019). Meta-analyses of simultaneous interpreting and working memory. *Interpreting*, *21*(2), 165–195. https://doi.org/10.1075/intp.00026.mel

Miller, K. S., & Sullivan, K. P. H. (2006). Keystroke Logging: An Introduction. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer Keystroke Logging and Writing: Methods and Applications: Vol. 1st ed* (pp. 1–9). Elsevier. http://ezproxy.unibo.it/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=166795&site=eds-live&scope=site

Milligan, S., & Schotter, E. R. (2024). Do readers here what they sea?: Effects of lexicality, predictability, and individual differences on the phonological preview benefit. *Journal of Memory and Language*, *135*. https://doi.org/10.1016/j.jml.2023.104480

Mizuno, A. (2005). *Process Model for Simultaneous Interpreting and Working Memory*. *50*, 739–752. https://doi.org/10.7202/011015ar

Morales, M., Patel, T., Tamm, A., Pickering, M. J., & Hoffman, P. (2022). Similar neural networks respond to coherence during comprehension and production of discourse. *Cerebral Cortex*, *32*(19), 4317–4330. https://doi.org/10.1093/cercor/bhab485

Morgan, J. H., Cheng, C.-Y., Pike, C., & Ritter, F. E. (2013). A Design, Tests and Considerations for Improving Keystroke and Mouse Loggers. *Interacting with Computers*, *25*(3), 242–258. https://doi.org/10.1093/iwc/iws014

Morrison, A. B., & Richmond, L. L. (2020). Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. *Cognitive Research: Principles and Implications*, *5*(1), 1. https://doi.org/10.1186/s41235-019-0201-4

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 555–563. https://doi.org/10.1037/0278-7393.28.3.555

Moss, J., & Schunn, C. D. (2015). Comprehension through explanation as the interaction of the brain's coherence and cognitive control networks. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00562

Muñoz Martín, R. (2010). On paradigms and cognitive translatology. In G. M. Shreve & E. Angelone (Eds.), *American Translators Association Scholarly Monograph Series: Vol. XV* (pp. 169–187). John Benjamins Publishing Company. https://doi.org/10.1075/ata.xv.10mun

Muñoz Martín, R. (2014). Situating translation expertise: A review with a sketch of a construct. In J. W. Schwieter & A. Ferreira (Eds.), *The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science* (1st ed., pp. 2–56). Cambridge Scholars Publisher.

Muñoz Martín, R. (2023). *Traductología cognitiva. Tratado general*. ULPGC Ediciones. https://spdc.ulpgc.es/libro/traductologia-cognitiva_150298/

Muñoz Martín, R., & Apfelthaler, M. (2021). Spillover Effects in Task-Segment Switching: A Study of Translation Subtasks as Behavioral Categories Within the Task Segment Framework. In R. Muñoz Martín, S. Sun, & D. Li (Eds.), *Advances in Cognitive Translation Studies* (pp. 19–45). Springer Singapore. https://doi.org/10.1007/978-981-16-2070-6_2

Muñoz Martín, R., & Apfelthaler, M. (2022). A task segment framework to study keylogged translation processes. *The International Journal of Translation and Interpreting Research*, *14*(2), 8–31. https://doi.org/10.12807/ti.114202.2022.a02

Muñoz Martín, R., & Conde Ruano, J. T. (2007). Effects of Serial Translation Evaluation. In H. E. Jungst & P. A. Schmitt (Eds.), *Translationsqualitat* (pp. 428–444). Lang.

Muñoz Martín, R., & González, C. (2021). Cognitive Translatology: A primer, revisited. *语言、翻译与认知 [Studies in Language, Communication & Cognition]*, *1*(4), 131–165. https://hdl.handle.net/11585/828295

Muñoz Martín, R., & Rojo López, A. M. (2018). Meaning. In S.-A. Harding & O. C. Cortés (Eds.), *The Routledge Handbook of Translation and Culture* (1st ed., pp. 61–78). Routledge. https://doi.org/10.4324/9781315670898-4

Muñoz Martín, R., & Tiselius, E. (in press). *Written words speak as loud. On the cognitive differences between translation and interpreting*.

Navarrete, E., & Costa, A. (2005). Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language*, *53*(3), 359–377. https://doi.org/10.1016/j.jml.2005.05.001

Norman, D. A., & Shallice, T. (1986). Attention to Action: Willed and Automatic Control of Behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation* (pp. 1–18). Springer US. https://doi.org/10.1007/978-1-4757-0629-1_1

Oakes, L. M., & Rakison, D. H. (2019). Developmental Cascades: A New Framework to Understand Change. In L. M. Oakes & D. H. Rakison, *Developmental Cascades* (pp. 100–122). Oxford University Press. https://doi.org/10.1093/oso/9780195391893.003.0005

Obler, L. K. (2012). Conference interpreting as extreme language use. *International Journal of Bilingualism*, *16*(2), 177–182. https://doi.org/10.1177/1367006911403199

O'Brien, S. (2023). Human-Centered augmented translation: Against antagonistic dualisms. *Perspectives*, 1–16. https://doi.org/10.1080/0907676X.2023.2247423

Olalla-Soler, C., Spinolo, N., & Muñoz Martín, R. (2023). Under Pressure? A Study of Heart Rate and Heart-Rate Variability Using SmarTerp. *HERMES - Journal of Language and Communication in Business*, *63*, Article 63. https://tidsskrift.dk/her/article/view/134292

Oléron, P., & Nanpon, H. (1965). Research into simultaneous translation. In F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 42–50). Routledge.

Olson, G. M., & Olson, J. S. (2003). Human-Computer Interaction: Psychological Aspects of the Human Use of Computing. *Annual Review of Psychology*, *54*(1), 491–516. https://doi.org/10.1146/annurev.psych.54.101601.145044

Onishi, N., & Yamada, M. (2020). Why translator competence in information searching matters: An empirical investigation into differences in searching behavior between professionals and novice translators. *Invitation to Interpreting and Translation Studies*, *22*, 1–23.

Palmér, M. (2023). *pynput: Monitor and control user input devices* (1.7.6) [Python; MacOS :: MacOS X, Microsoft :: Windows :: Windows NT/2000, POSIX]. https://github.com/moses-palmer/pynput

Pan, S. C., Zung, I., Imundo, M. N., Zhang, X., & Qiu, Y. (2023). User-generated digital flashcards yield better learning than premade flashcards. *Journal of Applied Research in Memory and Cognition*, *12*(4), 574–588. https://doi.org/10.1037/mac0000083

Paneth, E. (1957). *An investigation into conference interpreting: With special reference to the training of interpreters* [MA dissertation, University of London]. WorldCat.

Pérez-Pérez, P. S. (2018). The use of a corpus management tool for the preparation of interpreting assignments: A case study. *The International Journal of Translation and Interpreting Research*, *10*(1), 137–151. https://doi.org/10.12807/ti.110201.2018.a08

Peterson, M. A., & Skow-Grant, E. (2003). Memory and Learning in Figure–Ground Perception. In *Psychology of Learning and Motivation* (Vol. 42, pp. 1–35). Academic Press. https://doi.org/10.1016/S0079-7421(03)01001-6

Phung, D. V., & Michell, M. (2022). Inside Teacher Assessment Decision-Making: From Judgement Gestalts to Assessment Pathways. *Frontiers in Education*, *7*, 830311. https://doi.org/10.3389/feduc.2022.830311

Pisani, E., & Fantinuoli, C. (2021). Measuring the Impact of Automatic Speech Recognition on Number Rendition in Simultaneous Interpreting. In C. Wang & B. Zheng (Eds.), *Empirical Studies of Translation and Interpreting* (1st Edition, pp. 181–197). Routledge.

Plevoets, K., & Defrancq, B. (2018). The cognitive load of interpreters in the European Parliament: A corpus-based study of predictors for the disfluency *uh(m)*. *Interpreting. International Journal of Research and Practice in Interpreting*, *20*(1), 1–32. https://doi.org/10.1075/intp.00001.ple

Prandi, B. (2015). The Use of CAI Tools in Interpreters' Training: A Pilot Study. *Proceedings of the 37th Conference Translating and the Computer*, 48–57. https://aclanthology.org/2015.tc-1

Prandi, B. (2017). Designing a Multimethod Study on the Use of CAI Tools during Simultaneous Interpreting. *Proceedings of the 39th Conference Translating and the Computer*, 76–88. www.asling.org

Prandi, B. (2020). The use of CAI tools in interpreter training: Where are we now and where do we go from here? *inTRAlinea*, *Special Issue: Technology in Interpreter Education and Practice.* http://www.intralinea.org/specials/article/2512

Prandi, B. (2023). Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology. In *Language Science Press*. Language Science Press. https://doi.org/10.5281/zenodo.7143055.

Rai, S., Choubey, V., Suryansh, & Garg, P. (2022). A Systematic Review of Encryption and Keylogging for Computer System Security. *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 157–163. https://doi.org/10.1109/CCiCT56684.2022.00039

Reiner, P. B., & Nagel, S. K. (2017). *Technologies of the extended mind: Defining the issues* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198786832.003.0006

Robbins, P., & Aydede, M. (Eds.). (2009). *The Cambridge handbook of situated cognition*. Cambridge University Press.

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Rosman, T., Mayer, A.-K., & Krampen, G. (2016). A longitudinal study on information-seeking knowledge in psychology undergraduates: Exploring the role of information literacy instruction and working memory capacity. *Computers & Education*, *96*, 94–108. https://doi.org/10.1016/j.compedu.2016.02.011

Sadler, D. R. (2009). Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 1–19). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8905-3_4

Sagiroglu, S., & Canbek, G. (2009). Keyloggers: Increasing threats to computer security and privacy. *IEEE Technology and Society Magazine*, *28*(3), 10–17. https://doi.org/10.1109/MTS.2009.934159

Salaets, H., & Brône, G. (Eds.). (2020). *Linking up with video: Perspectives on interpreting practice and research*. John Benjamins Publishing Company.

Sales, D., Pinto, M., & Fernández-Ramos, A. (2018). Undressing information behaviour in the field of translation: A case study with Translation trainees. *Journal of Librarianship and Information Science*, *50*(2), 186–198. https://doi.org/10.1177/0961000616666131

Scaglioni, G. (2013). Simultaneous Interpreting from German into Italian: The Importance of Preparation on a Selection of Cultural Items. *The Interpreters' Newsletter*, *18*, 81–103. http://hdl.handle.net/10077/9753

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories — new models. *Interpreting. International Journal of Research and Practice in Interpreting*, *13*(2), 176–204. https://doi.org/10.1075/intp.13.2.02see

Shreve, G. M., Lacruz, I., & Angelone, E. (2010). Cognitive effort, syntactic disruption, and visual interference in a sight translation task. In G. M. Shreve & E. Angelone (Eds.), *Translation and Cognition* (pp. 63–84). John Benjamins Publishing Company. https://doi.org/10.1075/ata.xv.05shr

Shreve, G. M., Lacruz, I., & Angelone, E. (2011). Sight translation and speech disfluency: Performance analysis as a window to cognitive translation processes. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and Strategies of Process Research: Integrative approaches in Translation Studies* (pp. 93–120). John Benjamins Publishing Company. https://doi.org/10.1075/btl.94.09shr

Shreve, G. M., Schäffner, C., Danks, J. H., & Griffin, J. (1993). Is There a Special Kind of "Reading" for Translation?: An Empirical Investigation of Reading in the Translation Process. *Target. International Journal of Translation Studies*, *5*(1), 21–41. https://doi.org/10.1075/target.5.1.03shr

Smeaton, A. F., Krishnamurthy, N. G., & Suryanarayana, A. H. (2021). Keystroke Dynamics as Part of Lifelogging. In J. Lokoč, T. Skopal, K. Schoeffmann, V. Mezaris, X. Li, S. Vrochidis, & I. Patras (Eds.), *MultiMedia Modeling* (pp. 183–195). Springer International Publishing.

Sousa, D. A. (2022). How the Brain Processes Information. In D. A. Sousa, *How the Brain Learns* (Sixth Edition, pp. 37–70). Corwin. https://doi.org/10.4135/9781071855324

Spink, A., Ozmutlu, H. C., & Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology*, *53*(8), 639–652. https://doi.org/10.1002/asi.10124

Sprevak, M. (2019). Extended Cognition. In T. Crane (Ed.), *The Routledge Encyclopedia of Philosophy Online* (1st ed.). Routledge. https://doi.org/10.4324/9780415249126-V049-1

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Strömqvist, S., & Karlsson, H. (2002). *ScriptLog for Windows: User's manual* [Technical Report]. University of Lund: Department of Linguistics, and University College of Stavanger: Centre for Reading Research.

Su, W. (2019). Interpreting quality as evaluated by peer students. *The Interpreter and Translator Trainer*, *13*(2), 177–189. https://doi.org/10.1080/1750399X.2018.1564192

Su, W. (2020). *Eye-Tracking Processes and Styles in Sight Translation*. Springer Singapore. https://doi.org/10.1007/978-981-15-5675-3

Tammasrisawat, P., & Rangponsumrit, N. (2023). The Use of ASR-CAI Tools and their Impact on Interpreters' Performance during Simultaneous Interpretation. *New Voices in Translation Studies*, *28*(2), 25–51. https://doi.org/10.14456/nvts.2023.19

Tarasenko, R., & Amelina, S. (2020). A Unification of the Study of Terminological Resource Management in the Automated Translation Systems as an Innovative Element of Technological Training of Translators. In O. Sokolov, G. Zholtkevych, V. Yakovyna, Y. Tarasich, V. Kharchenko, V. Kobets, O. Burov, S. Semerikov, & H. Kravtsov (Eds.), *Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops* (Vol. 2732, pp. 1012–1027). CEUR. https://ceur-ws.org/Vol-2732/#20201012

Tarasenko, R. O., Amelina, S. M., & Semerikov, S. O. (2021). Conceptual Aspects of Interpreter Training Using Modern Simultaneous Interpretation Technologies. In E. Vadim, E. David, M. Heinrich C., N. Mykola, B. Sergiy, Z. Grygoriy, Y. Vitaliy, & S. Aleksander (Eds.), *Proceedings of the 17th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume I: Main Conference, PhD Symposium, and Posters* (p. 14).

Timarová, S., Dragsted, B., & Gorm Hansen, I. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Benjamins Translation Library* (Vol. 94, pp. 121–146). John Benjamins Publishing Company. https://doi.org/10.1075/btl.94.10tim

Waddington, C. (2001). Should translations be assessed holistically or through error analysis? *HERMES - Journal of Language and Communication in Business*, *26*, Article 26. https://doi.org/10.7146/hjlcb.v14i26.25637

Wan, H., & Yuan, X. (2022). Perceptions of Computer-assisted Interpreting Tools in Interpreter Education in Chinese Mainland: Preliminary Findings of a Survey. *International Journal of Chinese and English Translation & Interpreting*. https://doi.org/10.56395/ijceti.v1i1.8

Wang, X., & Wang, C. (2019). Can computer-assisted interpreting tools assist interpreting? *Transletters. International Journal of Translation and Interpreting*, *2*, Article 2. https://www.uco.es/ucopress/ojs/index.php/tl/article/view/11575

Wannagat, W., Steinicke, V., Tibken, C., & Nieding, G. (2022). Same topic, different genre: Elementary school children's mental representations of information embedded in narrative and expository texts. *Learning and Instruction*, *80*, 101559. https://doi.org/10.1016/j.learninstruc.2021.101559

Wehrle, S. (2023). Conversation and intonation in autism: A multi-dimensional analysis. In *Language Science Press*. Language Science Press. https://doi.org/10.5281/zenodo.10069004

Will, M. (2008). Knowledge Management for Simultaneous Interpreters in LSP Conferences. In H. Gerzymisch-Arbogast, G. Budin, G. Hofer, & European Marie Curie Conference (Eds.), *LSP translation scenarios: Selected contributions to the EU Marie Curie Conference Vienna 2007* (pp. 65–100). Books on Demand.

Will, M. (2020). Computer Aided Interpreting (CAI) for Conference Interpreters. Concepts, Content and Prospects. *ESSACHESS – Journal for Communication Studies*, *13*(1(25)), Article 1(25). http://www.essachess.com/index.php/jcs/article/view/480

Witte, T. (2018). Mouse Underlaying: Global Key and Mouse Listener Based on an Almost Invisible Window with Local Listeners and Sophisticated Focus. *ICST Transactions on Security and Safety*, *5*(15), 155740. https://doi.org/10.4108/eai.15-10-2018.155740

Woesler, M. (2021). Modern Interpreting with Digital and Technical Aids: Challenges for Interpreting in the Twenty-First Century. In R. Moratto & M. Woesler (Eds.), *Diverse Voices in Chinese Translation and Interpreting* (pp. 191–217). Springer Singapore. https://doi.org/10.1007/978-981-33-4283-5_8

Xu, R. (2015). *Terminology Preparation for Simultaneous Interpreters* [Doctoral Thesis, University of Leeds]. http://etheses.whiterose.ac.uk/10164/

Xu, R. (2018). Corpus-based terminological preparation for simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *20*(1), 33–62. https://doi.org/10.1075/intp.00002.xu

Yang, Z. (2021). Effective Computer-assisted Terminology Management Through SDL MultiTerm. *Journal of Physics: Conference Series*, *1861*(1). https://doi.org/10.1088/1742-6596/1861/1/012106

Yao, B. (2021). Mental Simulations of Phonological Representations Are Causally Linked to Silent Reading of Direct Versus Indirect Speech. *Journal of Cognition*, *4*(1), Article 1. https://doi.org/10.5334/joc.141

Yuan, L., & Wang, B. (2023). Cognitive processing of the extra visual layer of live captioning in simultaneous interpreting. Triangulation of eye-tracked process and performance data. *Ampersand*, *11*, 100131. https://doi.org/10.1016/j.amper.2023.100131

Zhang, J. (2021). *An Experiment Report on the Impact of Computer-Aided Interpreting Tools on Simultaneous Interpreting* [Master's Thesis, China Foreign Affairs University]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202201&filename=1021596437.nh&v=

Zhang, Q., Zhu, X., & Damian, M. F. (2018). Phonological activation of category coordinates in spoken word production: Evidence for cascaded processing in English but not in Mandarin. *Applied Psycholinguistics*, *39*(5), 835–860. https://doi.org/10.1017/S0142716418000024

Zhong W., & Xin Z. (2021). Application of Screen Recording in Translation Teaching—A Process-oriented Experiment in Translation Teaching and Its Implications for Teaching. *Way to Translation*, *1*(3), 54–61. https://doi.org/10.35534/wtt.0103008

Zhou, H., Weng, Y., & Zheng, B. (2021). Temporal Eye-Voice Span as a Dynamic Indicator for Cognitive Effort During Speech Processing: A Comparative Study of Reading Aloud and Sight Translation. In R. Muñoz Martín, S. Sun, & D. Li (Eds.), *Advances in Cognitive Translation Studies* (pp. 161–179). Springer Singapore. https://doi.org/10.1007/978-981-16-2070-6_8

Zhou, L. (2019). *The Impact of Computer-Aided Interpreting Tools on Simultaneous Interpreting Performance: Taking InterpretBank as an Example* [Master's Thesis, Xiamen University]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202002&filename=1019069326.nh&v=

Zhu, M., Zhang, M., & Gu, L. (2023). Insights into Editing and Revising in Writing Process Using Keystroke Logs. *Language Assessment Quarterly*, 1–24. https://doi.org/10.1080/15434303.2023.2291478

Zwaan, R. A., & Rapp, D. N. (2006). Discourse Comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics (Second Edition)* (pp. 725–764). Academic Press. https://doi.org/10.1016/B978-012369374-7/50019-5

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: An international vs. a national view. *The Interpreters' Newsletter*, *15*, 127–142. https://www.openstarts.units.it/handle/10077/4754

appendices

Appendix A Profiling survey for informants

Intent to Enroll Form

Any information you provide will be protected and not be used in any other way.

1. Please fill in your name here. (For example: CHEN LILI)

2. Your gender?
☐ Woman ☐ Man ☐ Other/ Prefer not to say

3. Your age?

4. What's your affiliation? (For example: University of Oxford)

5. What is your major?
☐ Translation ☐ Interpreting ☐ Other

6. How many semesters of your MA degree have you already COMPLETED?
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

7. What kind of computer (operating system) are you using for the study?
☐ Windows ☐ Mac ☐ Other

8. Do you own or have access to a headset and will use it for all SI tasks in the study (as required)?
☐ Yes ☐ No

9. Can you have and use a quiet place to perform all tasks without interruptions, distractions, or noise?
☐ Yes ☐ No

10. Please let me have your email address. (To be used only for matters strictly related to this study, then deleted)

11. Do you have any experience using InterpretBank?
☐ Yes ☐ No

12. Please elaborate on when and for how long you used it

13. Do you have any experience in using other computer-assisted interpreting tools, like Intragloss?

☐ Yes ☐ No

14. Could you please tell me how many hours of using computer-assisted interpreting tools you have?

15. ID code: we need this code to avoid duplicate responses and will not be used in any other way or published.

Please write:
First TWO letters of mother's Last Name?
Number of brothers or sisters (living and deceased)?
Number representing the month you were born.
Example:

- the daughter (born in April) of CHEN LI, has two sisters, and she will write CH0204.
- the son (born in June) of WANG YUE, has no brothers, and he will write WA0006.

Thanks for your interest.

I will contact you shortly to tell you whether you are shortlisted.
Please feel free to write whatever questions you may have: zhiqiang.du@studio.unibo.it.

Appendix B Informants' profiles

groups	informants	gender	age	number of completed semesters in MA	operating system	experience with InterpretBank usage	experience with other CAI tools
InterpretBank	Alex		23	2	Mac		
	Blake		25	4			
	Casey		24	2	Mac		
	Dana		22	2		Yes	Yes
	Erin		23	2			
	Frankie		23	2	Mac	Yes	Yes
	Gale		23	2			
	Harley		24	2		Yes	Yes
	Ira		23	2		Yes	Yes
	Jordan		23	2		Yes	Yes
	Kelly	male	27	4	Mac	Yes	Yes
	Lee		24	2			
EXCEL	Morgan		27	5			
	Noel		24	3			
	Oakley		31	3			
	Peyton	male	24	2	Mac		
	Quinn		27	6			
	Riley		23	2	Mac		
	Sidney		23	2			
	Taylor		23	2			
	Uli		24	2			
	Val		34	2			

note: most informants were females using windows with no experience using InterpretBank and other CAI tools.

Appendix C Potential problem triggers

term coded	term	n-gram
01	time perception	2
02	fast time-restricted feeding	3
03	gene expression	2
04	norepinephrine	1
05	entrainment	1
06	circannual rhythms	2
07A	melatonin	1
08A	light-dark cycle	2
09	circadian time cycle	3
10_07B	melatonin	1
11	clock genes	2
12_08B	light-dark cycle	2
13	ultradium rhythm	2
14	slow-wave sleep	2
15	REM sleep	2
16_07C	melatonin	1
17	acetylcholine	1
18	dopamine	1
19	basic rest-activity cycle	3
20_08C	light-dark cycle	2
21A	resting blood glucose	3
22	human nervous system	3
23_21B	resting blood glucose	3
24	neuromodulators	1
25	neural circuits	2
26	Serotonin	1
27	frame rate	2
28	cannabinoid receptor activation	3
29_21C	resting blood glucose	3
30	overclocking	1
31	hippocampus	1
32	neocortex	1
33	cognitive behavioral therapy	3
34	spontaneous eye-blink rate	3
35	metabolism	1
36	brown fat stores	3
37	mesolimbic reward pathway	3
38	nucleus accumbens	2
39	ventral tegmental area	3

Table 34. Potential problem triggers in Cycle I.

term coded	term	n-gram
01	hormones	1
02	immune system	2
03	cortisol	1
04	epinephrine	1
05	estrogen	1
06	cholesterol	1
07	dietary cholesterol	2
08	stress hormone	2
09	adrenaline	1
10A	neuroplasticity	1
11	corticotropin releasing hormone	3
12	pituitary	1
13	insomnia	1
14A	blood vessels	2
15	Arteries	1
16	stress response	2
17	net effect	2
18A	sympathetic chain ganglia	3
19	chronic cortisol elevation	3
20	non-sleep deep rest	3
21_10B	neuroplasticity	1
22	stress threshold	2
23	high-intensity interval training	3
24	abdominal fat accumulation	3
25	immune response	2
26_14B	blood vessel	2
27	neural energy	2
28	chronic stress	2
29_18B	sympathetic chain ganglia	3
30	negative feedback loop	3
31	melanocytes	1
32	sympathetic nervous system	3
33	hair stem cells	3
34	melanocyte stem cells	3
35	low-density lipoprotein cholesterol	3
36_18C	sympathetic chain ganglia	3
37	psychological stress	2
38_14C	blood vessel	2
39_10C	neuroplasticity	1

Table 35. Potential problem triggers in Cycle II.

term coded	term	n-gram
01	emotions	1
02	carbohydrates	1
03	micronutrients	1
04	vagus nerve	2
05	10th cranial nerve	3
06	neurons	1
07A	reward prediction error	3
08	heart rate	2
09	polyvagal theory	2
10	dorsal vagus	2
11	spinal cord	2
12	hypothalamus	1
13	lateral hypothalamus	2
14	locus coeruleus	2
15	amino acid	2
16	neurochemicals	1
17	intestines	1
18A	L-tyrosine	1
19	plant-based foods	2
20_07B	reward prediction error	3
21	raphae nucleus	2
22	antidepressants	1
23_18B	L-tyrosine	1
24	gut brain axis	3
25	blood brain barrier	3
26	long-chain fatty acids	3
27_18C	L-tyrosine	1
28	fatty acid ratio	3
29	heart rate variability	3
30	autonomic nervous system	3
31	respiratory sinus arrhythmia	3
32A	gut microbiome	2
33	prebiotics	1
34	central nervous system	3
35_07C	reward prediction error	3
36_32B	gut microbiome	2
37	neurotransmitters	1
38	circadian type fasting	3
39_32C	gut microbiome	2

Table 36. Potential problem triggers in Cycle III.

Appendix D Survey for Excel group in Cycle III

Please fill in your nickname in the study here:

□

1. How do you feel about your interpreting performance?

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

2. glossary task

statements	totally disagree	disagree	not sure	agree	totally agree
I like to use applications on my phone instead of a PC for term retrieval.	◎	◎	◎	◎	◎
While locating the term's translation, I would check its pronunciation.	◎	◎	◎	◎	◎
I did not verify the accuracy of the translation solution given by web resources (e.g., online dictionaries, term bank)	◎	◎	◎	◎	◎
I would rely on automatic term extraction from texts rather than human selection.	◎	◎	◎	◎	◎

3. booth task

statements	never	once in a while	sometimes	mostly	always
In a SI task, memorizing the term and its translation is more useful than term retrieval.	◎	◎	◎	◎	◎
I'd look for a suitable computer-assisted interpreting (CAI) tool for SI tasks	◎	◎	◎	◎	◎
I do not use the existing and shared glossary instead of creating my own.	◎	◎	◎	◎	◎
CAI training can influence my selection of CAI tools.	◎	◎	◎	◎	◎

4. Do you think that terminology management service will improve your term preparation efficiency?

☐ No ☐ Yes

5. Do you think that you will benefit from the advanced functions of CAI tools (e.g., speech recognition, and AI translation)?

☐ Yes ☐ No

Thanks for your participation and time.

Appendix E Surveys for InterpretBank group in Cycles II and III

Please fill in your nickname in the study here:

□

1. How do you feel about your interpreting performance?

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

2. glossary task

statements	totally disagree	disagree	not sure	agree	totally agree
Automatic extraction can spare me time at extracting technical terms from files	◎	◎	◎	◎	◎
Manual extraction is a must-have function	◎	◎	◎	◎	◎
InterpretBank is convenient to compile glossaries	◎	◎	◎	◎	◎

3. booth task

statements	never	once in a while	sometimes	mostly	always
I use Booth mode when noticing a technical term.	◎	◎	◎	◎	◎
Term retrieval in Booth mode can help locate the target terms correctly.	◎	◎	◎	◎	◎
Term retrieval in Booth mode can reduce my pressure when tackling technical terms.	◎	◎	◎	◎	◎

4. Do you think that you will continue to use InterpretBank to assist you in glossary preparation?

☐ No

☐ Yes

5. Do you think that you will continue to use InterpretBank to assist you in booth tasks?

☐ Yes

☐ No

Thanks for your participation and time.

Appendix F Follow-up survey

Please fill in your Nickname in the study here:

1. In the 12 months after our study in December 2023, how many times did you use InterpretBank?

☐ Never (zero times)
☐ Very rarely (1–2 times)
☐ Rarely (3–7 times)
☐ Occasionally (8–15 times)
☐ Frequently (16–30 times)
☐ Always (nearly every week)

2. How do you use InterpretBank? (Please rank them by frequency, from the most frequent to the least frequent). If you use InterpretBank in any other way, please explain it in the next question.

statements	most frequent	very frequently	less frequently	rarely	never
I use the automatic extraction of terms	◎	◎	◎	◎	◎
I use it to correct the entries of my glossary	◎	◎	◎	◎	◎
I use it to practice and memorize terms with flashcards	◎	◎	◎	◎	◎
I use it to find translations for the glossary entries	◎	◎	◎	◎	◎
I use it in the booth, to look up glossary entries	◎	◎	◎	◎	◎

3. If you use InterpretBank in any other way not covered above, please ex-plain here, and add how frequently you use it, using the above categories (i.e., from most frequent to never).

4. Did you pay anything to be able to use InterpretBank?

☐ Yes ☐ No

5. For InterpretBank, are there any improvements you think could be made? You may respond either in Chinese or in English.

6. Could you please share the reasons why you did not use InterpretBank? You may respond in either Chinese or English.

7. Thank you for your help. If you would like to add anything, please use the box below to do so. Your feedback is greatly appreciated.

Appendix G Information sheet for Phd raters

Intent to enroll as a quality rater in a research project

This is the registration form to volunteer to contribute to a cutting-edge PhD project. The task consists of assessing **45 audio files**, each one about **13** minutes long. That is, just listening to all audio files completely amounts to **a total of 10 hours**. The whole task may thus take **between 12 and 15 hours**. Luckily, you can break down the task into three to five sessions within 3 weeks, but you need to assess all audios in the exact order that you will be given because they will be intentionally randomized. However, you may listen to parts of an audio file again to compare it with the one you are currently assessing.

Further good news is that we are **NOT** interested in a detailed report or a rubric-based scoring analysis of any kind. We only want you to give us an ***intuitive ranking*** of each file into one of six quality groups, from best to worst. No further explanation needed!

You need, however, to be serious in your intuitive assessment, and to listen to all audios in full. Just remember that people tend to be stressed and get tired toward the end. So, the impression you get based on the first or the middle parts might not be the same when you reach the end of that audio. Please note that this registration does not constitute a guarantee of being enrolled. The final confirmation will be sent out through an email letter. This is also the best time for you to change your mind if you think this is too much work or that it will take too much time from you.

1. Please fill in your name here.
   Example: LI (first name), CHEN (last name) as in LI CHEN

2. Your gender?
   ☐ Woman ☐ Man ☐ Other/ Prefer not to say

3. Your age?

4. What's your affiliation? (For instance, the University of Oxford)

5. What is the prospective title of your current PhD dissertation or postdoctoral research project?

6. What year of your PhD program are you currently in?

☐ first-year
☐ second-year
☐ third-year
☐ further advanced
☐ completed
☐ post-doc

7. Please let me have your email address. (To be used only for matters strictly related to this study, then deleted)

8. Do you commit to thoroughly and completely listening to all audio files assigned to you?
☐ Yes
☐ No
☐ I changed my mind since I didn’t have enough time to finish the task.

9. Do you pledge to report any incidence, difference, anomaly, and the like? Accidents may happen and they may perhaps be solved, but we need to know them.
☐ Yes ☐ No

10. Would you be willing to be contacted for future research activities?
☐ Yes ☐ No ☐ Maybe

11. ID code: we need this code to avoid duplicate responses and will not be used in any other way or published.
Please write:
First TWO letters of grandmother’s Last Name?
Number of brothers or sisters (living and deceased)?
Number representing the month you were born.
Example:

- If the granddaughter (born in April) of CHEN LI, has two sisters, then she will write CH0204.
- If the grandson (born in June) of WANG YUE, has no brothers, then he will write WA0006.

Thanks for your interest.
I will contact you soon to tell you whether you are one of the few raters invited to participate. Please feel free to write whatever questions you may have: zhiqiang.du@studio.unibo.it

Appendix H Chunks and sentences of source speech texts

No.		chunks in Cycle I
sentences	chunks	
1	*1*	Today we are talking about ***time perception***.
2	*2*	Our perception of time is perhaps one of the most important factors in life.
3	*3*	and it is directly linked to the neurochemical states that control our sensation of mood, stress, and contentment.
4	*4*	Before we begin our discussion about the perception of time, I'd like to mention some issues related to the topic of ***fast time-restricted feeding***.
5	*5*	Fast time restricted feeding involves eating for a particular period of time in each 24 hour cycle in a fairly regular way
6	*6*	Some people observe shorter feeding windows but, regardless, each feeding window should occur at more or less the same time within each 24-hour day.
7	*7*	This affects ***gene expression*** that regulates a number of positive effects on different tissues in the body.
8	*8*	and for some people It makes weight loss easier because they are not eating for long periods in each 24 hour cycle.
9	*9*	It basically boils down to whether or not something you ingest, either liquid or solid food.
10	*10*	Now let's talk about the most fundamental aspect of perception of time, which is called ***entrainment***.
11	*11*	Entrainment is a way in which our internal processes, our biology and our psychology are linked to external phenomena, .
12	*12*	and the most basic form of entrainment that we are all a slave to all year round for our entire life is the socalled ***circannual rhythms*** we have.
13	*13*	We have neurons, nerve cells in our eyes, in our brain and in our body that mark off the passage of time throughout the year, in effect a calendar system in our brain and body.
14	*14*	and the way this works is beautifully simple.
15	*15*	Light seen by our eyes inhibits, meaning it reduces, the amount of a hormone released in our brain called ***melatonin***.
16	*16*	Melatonin has two major functions.
17	*17*	One function is to make us sleepy at night and the other is to regulate other hormones in the body.
18	*18*	There are other forms of entrainment, meaning the matching of our brain and body to phenomena in our external environment.
19	*19*	In one study, skin was exposed to sunlight for about two hours a day.
20	*20*	In this case, on the upper body, the subjects were not totally naked.
21	*21*	The study shows that we are entrained to the external ***light-dark cycle*** and as the day lengthens, our hormone levels change.
22	*22*	And we can override this effect through exposure to bright lights and **resting blood glucose**.

No.		chunks in Cycle I
sentences	chunks	
23	*23*	The next level of time that we are all entrained or matched to is the so-called ***circadian time cycle***, which is a 24 hour rhythm.
24	*24*	This is perhaps the most powerful rhythm that we all follow and that none of us can escape from。
25	*25*	The cells in the circadian clock fire, meaning they release chemicals into our brain and body following a very regular rhythm.
26	*26*	Our perception of time is also conscious, watching the clock tick down day after day and related to the length of the day, influencing hormones like ***melatonin***.
27	*27*	This is the circadian clock. It all happens on a 24 hour cycle.
28	*28*	We identify ***clock genes*** and these clock genesregulate a number of different functions.
29	*29*	Every cell in our body has a 24 hour cycle of gene and protein expression.
30	*30*	The Earth rotates once every 24 hours.
31	*31*	The processes in every cell of our body are linked within the cells of our brain and body and match out the outside ***light-dark cycle***.
32	*32*	Next I'd like to talk about the so-called ***ultradium rhythm***.
33	*33*	Ultradium rhythms are rhythms of about 90 minutes or so, and our whole existence is broken up into these 90 minute ultradium cycles.
34	*34*	When we go to sleep at night, that entire period of sleep is broken up into these 90-minute ultradian cycles.
35	*35*	Early in the night we tend to have more ***slow-wave sleep***.
36	*36*	Later in the night, we tend to have more ***REM sleep***.
37	*37*	Throughout, our sleep is broken up into these 90-minute cycles, increases your **resting blood glucose.**
38	*38*	And when we wake up in the morning, many of the things that we do are governed by these ultradian rhythms.
39	*39*	For instance, if you learn a language or do physical work, 90-minute blocks seem to be the units for the brain to enter a state of focus and alertness, for doing work and concentrating.
40	*40*	Well, what you're entraining to in this case is the release of particular neurochemicals, such as ***melatonin*** , ***acetylcholine*** and *dopamine*, which allow your brain to focus for particular periods of time.
41	*41*	The amount of these chemicals that can be released drops to very low levels, which is why our ability to focus becomes diminished.
42	*42*	To find out more about what is behind these ultradian rhythms, you should consider another phenomenon.
43	*43*	This was originally called the ***basic rest-activity cycle***.
44	*44*	The interesting thing about these basic rest-activity cycles, these ultradian rhythms, is that we can reset them whenever we want.
45	*45*	If you decide that you want to apply ultradian rhythms to work and performance, you can set a clock and decide, okay, now my period of focus begins.

No.		**chunks in Cycle I**
sentences	**chunks**	
46	*46*	Unlike the dark light side, the light dark cycle, the ***light-dark cycle*** , these 90-minute cycles are linked to the ability of neurons to release **dopamine** and ***norepinephrine***,
47	*47*	which give us narrow focus, motivation, and drive, and shape the way these 90-minute cycles evolve
48	*48*	After about 90 minutes, these circuits are far less willing to engage and therefore it's much harder to continue to focus to a high degree.
49	*49*	The beauty of time perception in the ***human nervous system*** is that it boils down to a couple of simple molecules.
50	*50*	We use our memory to reconstruct certain sets of events in the past and get a sense of their relative positioning in time.
51	*51*	Neurochemicals like dopamine and norepinephrine are called ***neuromodulators*** because they modulate the way that other ***neural circuits*** work.
52	*52*	Serotonin is also important. Serotonin is released from another site in the brain and has a different effect on perception of time.
53	*53*	Studies show that the more dopamine is released into our brain, the more we tend to overestimate the amount of time that has just passed.
54	*54*	Fine-slicing of time periods is like increasing the ***frame rate*** on your camera.
55	*55*	Slow motion is achieved in movies by increasing the frame rate.
56	*56*	So if you take a movie at 30 frames per second and watch it, the action will appear to have a certain speed, right?
57	*57*	Other factors can increase serotonin levels, including ***cannabinoid receptor activation.***
58	*58*	That's because serotonin and other related molecules in the brain tend to lead to slower frame rates and ***resting blood glucose***.
59	*59*	It is clear that dopamine and norepinephrine can impact our perception of the passage of time. The best example is trauma.
60	*60*	People who have been in car accidents or have gone through some other kind of major trauma often experience what is called overclocking.
61	*61*	Overclocking is when the frame rate is so high that a memory gets etched into the psyche and it is very hard time to shake that memory off, and the emotions associated with that memory.
62	*62*	Now, that might not seem like a bad thing overall but the problem with overclocking is the way in which that information gets etched into the memory system.
63	*63*	The memory system, which involves areas of the brain like the ***hippocampus*** and the ***neocortex***, is basically a space-time recorder.
64	*64*	The nervous system doesn't have a whole lot of information about the inside, about the outside world, except light coming in through the ice and so forth.
65	*65*	So it has to take all these neural signals into account to create a record, to create a record of what has happened.

No.		chunks in Cycle I
sentences	chunks	
66	*66*	Nowadays there are many treatments for trauma, like ***cognitive behavioral therapy***, involving not just trying to reduce the amount of emotion associated with a memory, but also a deliberate speeding up or slowing down of that memory.
67	*67*	In other words, trying to allow the person who experienced the trauma to take control of the rate of the experience in their memory, not just whether or not the memory happened at all.
68	*68*	Much of the information coming into the brain through our eyes has an impact on our attention.
69	*69*	Well, it turns out that dopamine, and increases in dopamine levels, are associated with increases in ***spontaneous eye-blink rate***.
70	*70*	And every time we blink, our perception of time shifts, leading to an overestimation of time.
71	*71*	It seems as though in some way blink rate is actually related to frame rate.
72	*72*	But dopamine, and the way that it relates to eye shuttering, seems to control the frame rate of our experience, similarly to cold exposure.
73	*73*	There are many positive effects of cold exposure provided it's done properly.
74	*74*	It can lead to increases in ***metabolism*** and ***brown fat stores***, which are the good fat stores that you want.
75	*75*	They're like a furnace that allows us to heat up, stay warm in cold environments, increase resilience and so forth.
76	*76*	For instance, cold water exposure can increase your dopamine levels, which will also change your perception of time.
77	*77*	Your frame rate is going up.
78	*78*	Up until now, I've been talking about how dopamine, and to some extent, serotonin can differentially impact our perception of how fast or how slowly things are happening in the moment.
79	*79*	We haven't talked a lot about the neural circuits and the various areas of the brain that underlie this.
80	*80*	Some studies suggest that researchers can measure surprise by the release of dopamine in two areas of the brain that are part of what is called the ***mesolimbic reward pathway***.
81	*81*	The two areas of the brain that are important here are the ***nucleus accumbens*** and the ***ventral tegmental area***.
82	*82*	These are areas that release dopamine as a kind of token reward any time something is surprising or a positive expectation is met.
83	*83*	So this speaks again to dopamine being something that's important not just for positive events, but for unexpected events.
84	*84*	Today we have covered many aspects concerning perception of time.
85	*85*	We certainly haven't covered everything about time perception but we have mentioned aspects like entrainment, the role of dopamine, habits and various routines that can adjust our sense of time for the sake of particular goals.
86	*86*	Thank you for your time and attention today, and last but certainly not least, thank you for your interest in science.

No.		chunks in Cycle II
sentences	chunks	
1	1	Today, we're going to discuss ***hormones*** and focus on how particular hormones can influence our energy levels and ***immune system***.
2	2	Let's get started talking about how to increase your energy and immunity.
3	3	I think it's fair to say that most people would like to feel energetic during the day, especially if they are daytime workers.
4	4	In other words, you want to have energy, and you want your immune system to function well, to ward off infections of various kinds, bacterial infections, viral infections, etc.
5	5	And it turns out that the two key hormones in the processes of having enough energy and having a healthy immune system are ***cortisol*** and ***epinephrine***.
6	6	First, cortisol is a hormone much like ***estrogen*** and it is derived from ***cholesterol***.
7	7	This cholesterol could be produced by the liver.
8	8	Some people claim that ***dietary cholesterol*** has zero impact on circulating cholesterol coming from the liver.
9	9	But there are also problems with the idea that cholesterol levels are determined entirely by dietary intake.
10	10	Whatever the case, if you are under stress, more cholesterol is devoted to creating cortisol, which is a ***stress hormone***.
11	11	However, the word stress shouldn't stress you out, because we all need cortisol: cortisol is vital.
12	12	You don't want your cortisol levels to be too low.
13	13	It's very important for immune system function and to stop you feeling depressed.
14	14	So now we're going to talk about how to control your release cycles and levels of cortisol.
15	15	But first we should mention epinephrine, also known as ***adrenaline***, which has been rather demonized.
16	16	We tend to think of it as the main stress hormone, the substance that makes us feel anxious, wanting to fight or flee.
17	17	The fact of the matter is that epinephrine is our best friend when it comes to immunity, for protecting ourselves from infection.
18	18	And it is our best friend when it comes to remembering things, learning and activating ***neuroplasticity***.
19	19	Once again, it's a question of how much and for how long, and the specific timing of release of cortisol and epinephrine, as opposed to simply thinking of them as good or bad.
20	20	They're great when they are properly regulated; but they are terrible when they are unregulated.
21	21	Our brain makes what we call ***corticotropin releasing hormone***, which is a substance made by neurons in our brain that triggers the ***pituitary*** gland.

No.		chunks in Cycle II
sentences	chunks	
22	22	This gland is about an inch above the roof of the mouth at the base of the brain.
23	23	When you sense stress in your mind or your body senses stress from a wound or something of that sort,
	24	A signal is sent to neurons down each side of the spinal cord.
24	25	They are called the ***sympathetic chain ganglia***.
25	26	They operate like a sprinkler system that hoses your body with epinephrine.
26	27	This will increase your pulse and will increase your breathing rate.
27	28	In some cases it will constrict your ***blood vessels***, and it will also expand your ***arteries***, allowing blood to flow to your vital organs.
28	29	This is why your extremities get cold when you're feeling stressed and your pulse gets faster.
29	30	At the same time, adrenaline (epinephrine) is released from your adrenal glands, situated above the kidneys.
30	31	Let's look at ***stress response***.
31	32	So, we have established that we have cortisol and we have epinephrine, and their ***net effect*** is to create a feeling of increased energy.
32	33	Now of course regulating cortisol levels will not cure mental illness on its own,
	34	but it can promote a healthy state of mind, and reduce unhealthy states of mind, including depression.
33	35	Make sure that your highest levels of cortisol are first thing in the morning when you wake up.
34	36	The best way to stimulate an increase in cortisol at the appropriate time is very soon after waking, within 30 minutes or so, to get outside, get some sunlight, get out even if it’s overcast.
35	37	Late-shifted cortisol increase, meaning cortisol that is released around 8 or 9 pm, is a signature feature of many depressive disorders, including deep depression and anxiety.
36	38	And of course it correlates with conditions like ***insomnia***.
37	39	In principle, it is best to have a high level of cortisol early in the day but you can also expect increases in cortisol and adrenaline throughout the day if you experience unpleasant events.
38	40	For me, the events that are most unpleasant are things like heavy traffic, emails asking me to fill out forms, links that I need but I can't find, and so on.
39	41	This kind of thing stresses me out; I'm a human being.
40	42	You shouldn't experience such periods so often or so long that you are in a constant state of ***chronic cortisol elevation***.
41	43	This system of stress control is designed to increase our alertness to things, and when we get frustrated, it provides an opportunity to change behavior.
42	44	So if you find yourself getting stressed and staying stressed, you could incorporate a ***non-sleep deep rest*** (NSDR) protocol into your routine.
43	45	But please understand that the energy you experience during stress, that sudden increase in alertness and attention that comes from a moment of difficulty,

No.		chunks in Cycle II
sentences	chunks	
	46	That is a healthy hormonal and neural system in operation.
44	47	And this is why I speak of ***neuroplasticity***, the brain's ability to change in response to experience.
45	48	If you take control of your cortisol level, you won't feel so troubled by small stress increases throughout the day.
46	49	Now, there are ways to leverage stress, epinephrine and cortisol in ways that serve us, and we can do it deliberately.
47	50	There are also ways to do it that increase our ***stress threshold***, meaning they make it less likely that epinephrine and cortisol will be released.
48	51	Things like ice baths. Things like ***high-intensity interval training***.
49	52	Such activities are useful.
50	53	Of course, we should also mention the negative consequences of too much cortisol, and too much epinephrine, over extended periods.
51	54	Consider ***abdominal fat accumulation***. And sleep disturbances.
52	55	These can cause an immediate increase in epinephrine in the brain and body.
53	56	And chances are they are going to increase levels of cortisol as well.
54	57	Your body is actually primed to resist infection, when you have high levels of epinephrine in it for short periods of time.
55	58	You'd think that maybe cold water exposure, or something that increases your levels of stress and adrenaline, would make the effects of an infection worse, but no, quite the opposite is the case.
56	59	Essentially it is a signal from the nervous system to immune system organs. And that’s how epinephrine works in the body and in the brain.
57	60	The immune system can recognize invading substances but the nervous system provides the signal.
58	61	The duration here is a really important factor because if stress remains high for too long then, yes indeed, stress can hinder your ***immune response***.
59	62	The reason we're talking about epinephrine and cortisol for increasing energy and immune system function is because they are largely independent of the ***blood vessel***.
60	63	Of course, we heard so much when we were growing up about the need to eat well for energy, but the energy that we're talking about today is actually a much more powerful kind than what derives from food. We could call it ***neural energy***.
61	64	This has an impact on the ability of the immune system to react in response to intruders.
62	65	There are ways to adjust cortisol levels, even if you're feeling stressed.
63	66	I don't have to list all the ways that stress and ***chronic stress*** are bad. I think you know.
64	67	For instance, as a result of ***sympathetic chain ganglia***, your immune system will get battered over time and you won't be able to fight infection off.
65	68	If cortisol levels get too high, if there's too much cortisol floating around in the bloodstream, a ***negative feedback loop*** is set up, and the brain and pituitary shut down, in a so-called negative feedback loop.

No.		chunks in Cycle II
sentences	chunks	
66	69	So you really don't want to undergo chronic stress, because it leads to a cause and effect cascade: stress leads to more stress leads to more stress.
67	70	This is why it's very important to learn how to turn off the stress response.
68	71	A further negative effect of stress is hair graying.
69	72	Hair pigmentation, like skin pigmentation, is controlled by ***melanocytes***.
70	73	Well, it turns out that activation of the so-called ***sympathetic nervous system***, which is really just another name for the system that liberates adrenaline from the adrenals and hormones in the brain, drives depletion of melanocytes in ***hair stem cells***.
71	74	In other words, there is aging that we undergo based on our genetic makeup, but stress will also make us go gray.
72	75	A recent study showed that stress in various forms will deplete the ***melanocyte stem cells***.
73	76	So if chronic stress is so bad because of its effects on epinephrine and cortisol being elevated for too long then the question becomes And how do I keep chronic stress at bay?
74	77	People can take supplements, prescription drugs, et cetera. Supplements can modestly adjust the levels of cortisol like ***low-density lipoprotein cholesterol***, so-called "bad cholesterol" quote unquote.
75	78	It has been shown to have a profound effect on anxiety and cortisol itself.
76	79	There are some consequences due to reducing cortisol, for instance lowered heart rate, lowered rates of insomnia .
77	80	So we should think about stress mechanistically, in terms of epinephrine and cortisol.
78	81	And then if we do that we can think about how to regulate its ***sympathetic chain ganglia***.
79	82	Most ***psychological stress***, but also the release of substances like ghrelin that make you hungry, has an effect on our food consumption clock.
80	83	But we all eat to suppress cortisol and epinephrine. When we're hungry cortisol and epinephrine create an agitation so we go seek food.
81	84	When we ingest food typically if it includes carbohydrate, there's a blunting of epinephrine in the ***blood vessel***.
82	85	There are any number of ways to increase your adrenaline and stay calm.
83	86	We tend to focus on things like exercise as a way of raising our energy levels.
84	87	But today, again I'm talking about ***neuroplasticity***, deliberately increasing adrenaline while staying calm mentally, pulling back on adrenaline and cortisol, training the nervous system.

No.		chunks in Cycle III
sentences	chunks	
1	1	Today we're talking about ***emotions***, which are central to our entire experience of life.
2	2	And today we're going talk a lot about how the brain and body interact to create these things called emotions in the context of food and nutrition.
3	3	Nutrients like proteins, fats, and ***carbohydrates***, as well as ***micronutrients***, can impact the chemicals in our brain that give rise to the feelings of being happy or sad or sleepy or alert.
4	4	Delight or happiness or excitement are feelings of attraction to certain ideas, songs, people, places, and so on.
5	5	There's an action there: you're either moving toward or you're moving away from something.
6	6	And any time you're talking about action in the nervous system,
	7	You're talking about literally the contraction of muscles to move you toward or away from things.
7	8	And when you're talking about nerve-to-muscle phenomena and action, you're talking about the brain and the body, because the brain can't move around by itself. The brain has a body so that the organism can move it around.
8	9	There are circuits in the brain for aversion and attraction to things. And the body governs a lot of that.
9	10	so I’m going to introduce you to the nerve pathways connecting brain and body. One key component is the ***vagus nerve***.
10	11	So, the vagus nerve is one, not the only one, but one way in which our brain and body are connected, regulating our emotional states.
11	12	Basically the vagus is the ***10th cranial nerve***, which means that the control center of each of the ***neurons*** in the vagus is near the neck.
12	13	And a branch of the vagus goes into the brain. It is connected into the brain.
13	14	It is also connected into the stomach, the ***intestines***, the heart, and the lungs.
14	15	This vagus nerve is incredible because it's taking information from the body and it operates in two directions.
15	16	Reward prediction error senses things that are happening in the gut, in the lungs, everything, and sends that information up into the brain.
16	17	It also senses things in the gut like how distended or empty your stomach is. It can sense your pulse, your ***heart rate***.
17	18	It can sense your immune system, whether or not you have bacteria or other intruders in your body.
18	19	But the vagus is not just for sensing things. It's actually for controlling things too.
19	20	But you certainly don't want to just stimulate the vagus.
20	21	There was a fairly pioneering theory about the vagus which is called ***polyvagal theory***.
21	22	Polyvagal: "poly" means many. That is appropriate because it acknowledges that the vagus has many branches.

No.		chunks in Cycle III
sentences	chunks	
22	23	The idea is that there's a ***dorsal vagus***, which kind of runs down the back of the ***spinal cord***, which is involved in alertness and activation and fight or flight type stuff.
23	24	The problem with the polyvagal theory is that people often say that if your dorsal vagus is too active then you tend to be someone who's a little too keyed up; and when people are kind of in a state of freeze, or flacid and lethargic, it is because that pathway is hypoactive.
24	25	Let's keep things simple. The vagus nerve analyzes many features within the body and informs the brain of how to feel and what to do.
25	26	When you eat something sweet, you have sensor cells within your stomach, neurons, that sense the presence of sugary foods, and signal this to the brain.
26	27	This is a particular set of neurons detecting that something in your body has a particular feature, in this case the presence of sugars, sending information to the brain, essentially to control your behavior.
27	28	What most people don't know about is an area of the ***hypothalamus***, deep in the brain, kind of in the middle part of the brain, called the ***lateral hypothalamus***.
28	29	And the lateral hypothalamus is really interesting because it controls eating, and it inhibits eating. It stops us from eating.
29	30	And there's another area in the brain. It is called the ***locus coeruleus***.
30	31	Now the locus coeruleus is further back in the brainstem. It makes us feel alert.
31	32	But what's interesting is that as we approach food the locus coeruleus releases all these molecules that make us feel more anxious and alert. Sometimes it's felt as excitement.
32	33	Many people aren't aware that this interaction between the locus coeruleus and the lateral hypothalamus is a basic mechanism making us more alert and anxious around meal times.
33	34	And there are accelerators, things that make us want to eat more, like sugar and fats.
34	35	And they help generally, at least in the short term, to support the survival of animals. And also there are ***amino acids***.
35	36	Amino acids of course are important because they are the building blocks of muscle tissue, and other parts of our body that may need repair.
36	37	And it's fair to say that people will basically eat, not until their stomach is full but until the brain perceives that they have an adequate intake of amino acids.
37	38	But what most people don't realize is that amino acids are what the ***neurochemicals*** in the brain are made from.
38	39	And this feels good, when it is caused by events that you're looking forward to, and production of these neurochemicals is inhibited by events you're looking forward to that don't work out.
39	40	This is called ***reward prediction error***.
40	41	And as I mentioned, these amino acid sensors in our gut are detecting how many amino acids are there but they're also detecting which amino acids.
41	42	And there's a particular amino acid called ***L-tyrosine*** which comes from food.

No.		chunks in Cycle III
sentences	chunks	
42	43	It is in meat, nuts, and some ***plant-based foods***.
43	44	Now, hopefully you don't have Parkinson's disease. It's clear that dietary L-tyrosine supports the healthy production of things like dopamine, as well as other factors within the brain.
44	45	We have a brain body connection. There are many connections, but one of the main ones is the vagus nerve.
45	46	The vagus collects information about a lot of things, breathing, heart rate stuff that's happening in the gut, the intestines, and sends that information up to the brain.
46	47	Most of this information is in the neurons of the brain in an area called the ***raphae nucleus***. There are a few other locations too,
	48	and these are the neurons that control whether we feel satiated or not. Whether or not we feel happy and calm.
47	49	Now, you can't discuss the vagus without mentioning serotonin and ***antidepressants***.
48	50	They can be quite useful for many people. Not everyone responds well to them as I'm sure you've all heard.
49	51	They can do all sorts of things, and can work really well.
50	52	We ingest these foods that are rich in ***L-tyrosine***, and these supplements are things people take, they don't put them directly into the brain.
51	53	So yes, there's a ***gut brain axis***. Certain things to do with our experience of life and our emotions are happening in our body.
52	54	What you have to do is ingest things that are metabolized in certain ways that communicate to the brain or they pass into the brain themselves across what's called the ***blood brain barrier***.
53	55	These barriers exist. Just because you eat something, just because you ingest it, doesn't mean it's going to cross the blood brain barrier.
54	56	There are also nerves in the gut that detect the nutrient content of food.
55	57	And this has many effects. They are involved in mitochondrial activation of ***long-chain fatty acids***, which is a big mouthful, but it has some interesting effects on the neuro side.
56	58	So, the ovaries and the brain are the organs of the body that nature has gone out of its way to protect, with this additional layer of the blood brain barrier.
57	59	So, that's the effects of foods that are rich in ***L-tyrosine***.
58	60	Let's talk a little bit more about things that we ingest in our body, allowing our body to inform our brain and shift our mood.
59	61	I don't think most people know the omega-3 ***fatty acid ratio*** has a profound effect on depression.
60	62	People who are clinically depressed, suffering major depression, are found to be equally able to reduce depressive symptoms when they have high levels of omega-3.
61	63	It’s especially interesting for effects on the heart, because we know that omega-3 fatty acids can come from other sources too.
62	64	We know that having a heart rate that's really high or heart rate, that's really low. Neither of those are good.

No.		chunks in Cycle III
sentences	chunks	
63	65	We call it ***heart rate variability***.
64	66	A lot of people think, oh you just want a low heart rate, and a big stroke volume.
65	67	This has a lot to do with the tone of the ***autonomic nervous system***.
66	68	when you inhale, it speeds up the heart rate. When you exhale, it decreases the heart rate.
67	69	That's called ***respiratory sinus arrhythmia***. That's the basis of heart rate variability.
68	70	And we should mention another aspect of the gut-brain relationship that may surprise you; in some cases it might shock you.
69	71	And that's ***gut microbiome*** and ***prebiotics***. I think these compounds are powerful.
70	72	They carry risks for some people, but not for others. We'll look at this in more detail another time.
71	73	So that is the so-called gut brain axis.
72	74	Today we've actually been talking a lot already about the gut brain axis.
73	75	But let's just take a step back and think about our body plan.
74	76	We are actually a series of tubes.
75	77	There's the ***central nervous system*** that all started out as a tube.
76	78	It starts with our mouth, also our nose. And then we have all ***reward prediction error***.
77	79	They go down through our throat and then into our stomach and then into our various intestines.
78	80	Some of them make us feel better and they do that mainly by changing the conditions of our gut environment.
79	81	Maintaining a healthy ***gut microbiome*** is good for mood, great for digestion. And foods and fermented foods are going to be the best source.
80	82	In addition to that, they impact the ***neurotransmitters*** and the neurons that live in the gut.
81	83	So find the diet that's right for you.
82	84	Circadian type fasting is when I push out my first meal by a few hours. My first meal is generally around lunchtime or so but the longer periods of fasting go on for a day or two or three days.
83	85	Today we've talked mainly about how the body and things that we put inside this tube, that runs from our mouth to our rectum
84	86	Also what you believe about certain substances like ***gut microbiome***, and certain foods that have a profound effect on our bodies.

Appendix I Other applications / services in glossary tasks in Cycles II &III

Cycle	group	environment	Applications / services	duration (s)	percentage (%)
II	XL	online	chat.openai.com	1705.431	3.61
			termbox.lingosail.com	1245.872	2.63
			reverso.net	1171.258	2.48
			dict.cnki.net	878.053	1.86
			cn.bing.com	352.828	0.75
			eng.ichacha.net	341.012	0.72
			linguee.com	325.027	0.69
			translate.google.com	152.78	0.32
			deepl.com	89.053	0.19
			termonline.cn	56.614	0.12
			merriam-webster.com	46.74	0.10
			dictionary.cambridge.org	41.219	0.09
			baike.baidu.com	19.383	0.04
			fanyi.so.com	17.517	0.04
			english-corpora.org	15.944	0.03
			sum	***6458.731***	***13.65***
	IB	local	MS Excel	823.116	1.51
			Saladict App	460.927	0.84
			Zhiyun App	147.537	0.27
			WPS Word	20.692	0.04
			Adobe Acrobat Reader	13.731	0.03
			sum	***1466.003***	***2.68***
		online	baidu.com	1480.516	2.71
			fanyi.so.com	1198.68	2.19
			cn.bing.com	968.094	1.77
			cnki.net	535.394	0.98
			deepl.com	337.212	0.62
			zhihu.com	206.717	0.38
			fanyi.baidu.com	194.174	0.36
			dict.cnki.net	193.577	0.35
			linguee.com	141.83	0.26
			reverso.net	69.851	0.13
			letpub.com	69.199	0.13
			zhidao.baidu.com	51.871	0.09
			wikipedia.org	33.475	0.06
			dict.cn	31.427	0.06
			baike.baidu.com	31.086	0.06
			merriam-webster.com	26.234	0.05

Cycle	group	environment	Applications / services	duration (s)	percentage (%)
			termonline.cn	25.033	0.05
			scholar.google.com	15.467	0.03
			sum	***5609.837***	***10.26***
III	**XL**	**local**	Youdao app	1368.518	3.02
			Lingoes App	51.47	0.11
			Oulu App	10.212	0.02
			sum	***1430.2***	***3.16***
		online	reverso.net	718.406	1.59
			tmxmall.com	284.785	0.63
			cn.bing.com	281.773	0.62
			dict.cnki.net	168.998	0.37
			linguee.com	143.608	0.32
			Youdao Web	136.512	0.30
			translate.google.com	130.965	0.29
			baike.baidu.com	128.074	0.28
			zhihu.com	122.843	0.27
			eng.ichacha.net	35.039	0.08
			termonline.cn	32.93	0.07
			fanyi.baidu.com	27.752	0.06
			dictionary.cambridge.org	14.889	0.03
			deepl.com	12.161	0.03
			sum	***2238.735***	***4.94***
	IB	**local**	MS Word	1837.611	3.73
			MS Excel	659.494	1.34
			Saladict App	320.483	0.65
			Lingoes App	180.067	0.37
			Adobe Acrobat Reader	46.723	0.09
			MS word	17.952	0.04
			sum	***3062.33***	***6.21***
		online	baidu.com	690.315	1.40
			tmxmall.com	655.791	1.33
			translate.google.com	328.261	0.67
			zhihu.com	153.689	0.31
			baike.baidu.com	50.385	0.10
			termonline.cn	37.817	0.08
			deepl.com	36.458	0.07
			linguee.com	20.564	0.04
			reverso.net	18.585	0.04
			eng.ichacha.net	15.533	0.03
			sum	***2007.398***	***4.07***

Appendix J Individual rating scores per text

group	name	text 1	text 2	text 3
InterpretBank	Alex	4.6	2.3	2.3
	Blake	3.3	2.3	3.6
	Casey	4	3.7	4.3
	Dana	3.3	2.3	1.8
	Erin	3	3	2.7
	Frankie	3.7	2	2.7
	Gale	5	3	3.3
	Harley	2.7	2.3	3
	Ira	3.7	3.3	2.8
	Jordan	4	4	3.8
	Kelly	4.4	3.7	4
	Lee	3.7	3.7	3.3
Excel	Morgan	4.2	3	3.3
	Noel	5.7	4.8	4.3
	Oakley	4.3	4.3	4.7
	Peyton	4	3	2.3
	Quinn	3	3.3	3.2
	Riley	4	2.3	2.7
	Sidney	4.3	3.4	3.7
	Taylor	3.8	5.3	3.3
	Uli	5	4	4
	Val	5	3.8	5

Table 37. Informants' rating scores per text.