The goal of this doctoral research is to automatically predict difficult words in a text for non-native speakers. This prediction is crucial because good text comprehension is strongly determined by vocabulary. If a text contains too high a percentage of unknown words, the reader is likely to struggle to understand it. In order to provide good support to the non-native reader, we must first be able to predict the number of difficult words. Usually we do this manually based on expertise or prior vocabulary tests. However, such methods are not practical when we are reading in a computer-based environment such as a tablet or an online learning platform. In these cases, we need to properly automate the predictions.

The thesis is divided into three parts. The first part contains a systematic review of the relevant scientific literature. The synthesis includes 50 years of research and 140 peer-reviewed publications. The analyses highlight some crucial limitations, including that the scientific scope is divided into two little connected fields of research. The second part looks at two measures of lexical difficulty for non-native readers. On the one hand, the results show that there are important inconsistencies in how words are introduced in reading materials for Dutch and French labeled with CEFR levels. Therefore, this difficulty measure does not yet seem valid as a basis for an automated system. On the other hand, data was collected on how non-native speakers themselves perceive difficult words during reading. This difficulty measure is appropriate to develop a personalized and contextualized system. The final part looks at two types of predictive models developed on this data, namely mixed-effects models and artificial neural networks. On the one hand, the results clearly show that a personalized model makes significantly better predictions than a non-personalized model. On the other hand, the results show that a contextualized model can better discriminate difficulty, although these improvements are not always significant for each learner.

**Anaïs Tack** obtained a Master of Arts in English and French Linguistics and Literature from KU Leuven. Subsequently, she studied language engineering and computer science at UCLouvain, where she graduated as Master in Computational Linguistics.

**Faculté de philosophie, arts et lettres • Institut Langage et Communication • UCLouvain**
Place Blaise Pascal 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgium www.uclouvain.be/fial

**Faculty of Arts • Department of Linguistics • KU Leuven**
Blijde Inkomststraat 21 box 3301, 3000 Leuven, Belgium www.arts.kuleuven.be

ON THE AUTOMATED PREDICTION OF LEXICAL DIFFICULTY FOR FOREIGN LANGUAGE READERS Anaïs Tack




# MARK MY WORDS!

# ON THE AUTOMATED PREDICTION OF LEXICAL DIFFICULTY FOR FOREIGN LANGUAGE READERS

**Anaïs Tack**

June 2021

UCLouvain
KU LEUVEN

Thesis presented for the joint degree of

**Docteur en langues, lettres et traductologie**
**Doctor in de Taalkunde**

# MARK MY WORDS!

## ON THE AUTOMATED PREDICTION OF LEXICAL DIFFICULTY FOR FOREIGN LANGUAGE READERS

**UCLouvain**
Faculté de philosophie, arts et lettres
Institut Langage & Communication
Pôle de recherche en linguistique
Centre de traitement automatique du langage

**KU Leuven**
Faculty of Arts
Department of Linguistics
itec, imec research group




# MARK MY WORDS!

## ON THE AUTOMATED PREDICTION OF LEXICAL DIFFICULTY FOR FOREIGN LANGUAGE READERS

Dissertation submitted for the joint degree of
Docteur en langues, lettres et traductologie from UCLouvain
Doctor in de Taalkunde from KU Leuven

by

ANAÏS TACK

Louvain-la-Neuve
2020-2021

**Examination Committee**

*Chair*

| | |
|---|---|
| Prof. Dr. Magali Paquot | UCLouvain |

*Supervisors*

| | |
|---|---|
| Prof. Dr. Cédrick Fairon | supervisor UCLouvain |
| Prof. Dr. Piet Desmet | supervisor KU Leuven |
| Prof. Dr. Thomas François | co-supervisor UCLouvain |

*Jury Members*

| | |
|---|---|
| Prof. Dr. Núria Gala | Aix-Marseille Université |
| Prof. Dr. Detmar Meurers | Universität Tübingen |
| Prof. Dr. Maribel Montero Perez | representative KU Leuven; UGent |

Copyright © 2021 by Anaïs Tack.

Front cover design by Anaïs Tack.
This dissertation was typeset with LaTeX, *classicthesis* by André Miede and Ivo Pletikosić, and Linux Libertine by Philipp H. Poll.

Printed by Ciaco Imprimerie, Louvain-la-Neuve, Belgium.

In dedication to the loving memory of my grandfather, Henri.

What do we live for, if it is not to make life less difficult to each other?

— George Eliot, *Middlemarch* (1871), Book VIII. Chapter LXXII

# ACKNOWLEDGMENTS

As many have experienced before me, writing a doctoral thesis can be quite a daunting adventure. Someone dear once told me it is like embarking into a vast ocean and trying to make the journey overseas. Steering clear of a sea of troubles and landing on a new safe haven would not have been possible without the help, listening ears, and happy thoughts from all of you who have stood by me.

A main thank you goes to my three thesis supervisors: Cédrick, for his cordiality, moral support, and infallible trust in my work; Piet, for his kindness, strategic advice, and astute guidance; and Thomas, for his thorough suggestions, relevant pointers, and detailed reading of the manuscript.

A grateful thank you goes to my thesis advisors, Maribel Montero Perez and Lucia Specia, for their useful insights and to all researchers who have given me an opportunity to grow. First of all, I would like to thank the F.R.S.-FNRS (National Fund For Scientific Research) for having enabled me to conduct this project as a research fellow. I also thank all members of the examination committee for having accepted to review the thesis. I am particularly grateful to Núria Gala for having welcomed me as a visiting researcher in her lab and whose hospitality made me feel instantly at ease in Aix-en-Provence. I also thank Detmar Meurers for having invited me to participate in the SLA summer school in Tübingen. Furthermore, I would like to thank Chris Biemann for his enthusiasm in taking us aboard the CWI shared task venture. I thank Pierre Dupont for having allowed me to participate as a TA in his *Computational Linguistics* course. Finally, I am humbled to have met so many interesting and (com)passionate people during past conferences, seminars, and summer schools.

I am indebted to all students at UCLouvain and KU Leuven who gave me a helping hand. My humblest gratitude goes to Anne-Sophie Desmet for her most precious time. I thank Ruben Vanhauwaert for his contribution to the data collection, Dorian Ricci for his development skills, Brayan Delmée for his creative designs, and Louis Escouflaire for his meticulous annotation. I thank Carmen Eggermont for having helped me during data collection with her students.

A friendly thank you goes to my colleagues at CENTAL and itec. My most trustworthy PhD travel companion has been Serge, who was my desk buddy over the years and also, quite literally, my *compagnon de route* in the commute between both offices. I am much appreciative for his inspiring ideas, incessant kind-heartedness, and for keenly sharing his meticulous know-how of systematic literature reviews. I also thank Dirk De Hertog for his useful insights and recommendations for the deep learning models. Moreover, being part of two research groups also meant to be blessed with an incredible troop of colleagues: Adeline, Agathe, Amélie, André, Ann-Sophie, Bert, Damien, Frederik, Hans, Hubert, Isabeau, Jonas, Leonardo, Magali N., Olga, Patrick, Paul, Rodrigo, Romane, Trang, and Violeta. I also thank my colleagues who organized the 'PhD Lunches' at UCLouvain and the 'IICK and Friends' events at KU Leuven, for having brought many pleasant diversions into the workplace. Also, I would like to send a heartfelt thank you to my Hatha and Ashtanga Yoga instructors, Vincent, Stéphane, Philippe, Gemma, Elise, who helped me stay grounded throughout the past eight years studying and working in Louvain-la-Neuve. Namasté.

I count my lucky pennies to have had the presence and support of my closest friends and family. I thank my group of gal pals from Kortrijk – *de hippe coole trendy bende* as our WhatsApp group is called – for their quirkiness, pep talk, and almost 14 years of friendship. I'm also grateful for the support and encouragements I received from my uncle Luc, *il mio padrino*, my aunts Dominique and Titi, my cousins Matthis and Julian, and my brother-in-law Stijn. As the youngest of two, I very much look up to my sister Laura, who always set the example with much determination. I hope she can be at least half as proud of me as I was of her during her thesis defense. Moreover, the past burdensome months would not have been endurable without the many

reassuring and heart-warming text messages from my two grandmothers, who went in full virtual mode over the past pandemic year and of whom I'm proud to say they've mastered the Art of the iPad to a tee.

Last but not least, my keen interest in linguistics and technology was instilled first and foremost by my parents. From an early age, my interest in technology was ardently encouraged by my father. I am most obliged for his blind trust in my juvenile excavations on his Windows 95 computer, which set the stage for years to come. My mother, on the other hand, eagerly integrated historical and French language content into our childhood development long before CLIL became a hot topic in Flemish education. There are not enough words to express how grateful and blessed I am of her life-long, unwavering, unassuming love and support.

Finally, I dedicate this work to my grandfather, Henri. He sadly missed my very first graduation by a breath. Still, I found his influence and courage to be the most significant driving force for completing the dissertation. In his memory, I would like to invoke some concluding words of wisdom: *What do we live for, if it is not to make life less difficult to each other?*

# ABSTRACT

The goal of this doctoral research is to automatically predict difficult words in a text for non-native speakers. This prediction is crucial because good text comprehension is strongly determined by vocabulary. If a text contains too high a percentage of unknown words, the reader is likely to struggle to understand it. In order to provide good support to the non-native reader, we must first be able to predict the number of difficult words. Usually we do this manually based on expertise or prior vocabulary tests. However, such methods are not practical when we are reading in a computer-based environment such as a tablet or an online learning platform. In these cases, we need to properly automate the predictions.

The thesis is divided into three parts. The first part contains a systematic review of the relevant scientific literature. The synthesis includes 50 years of research and 140 peer-reviewed publications on the statistical prediction of lexical competence in non-native readers. Among other things, the analyses show that the scientific scope is divided into two fields of research that have little connection with each other. On the one hand, there is a long tradition of experimental research in foreign language acquisition (SLA) and computer-assisted language learning (CALL). These experimental studies mainly test the effect of certain factors (e.g., repeating difficult words or adding electronic glosses) on learning unrecognized words during reading. On the other hand, recent studies in natural language processing (NLP) rely on artificial intelligence to automatically predict difficult words.

Moreover, the literature review points out some limitations that were further studied in this doctoral research. The first limitation is the lack of contextualized measures and predictions. Although we know from research that the context in which a word occurs is an important factor, predictions are often made based on isolated vocabulary tests, among other things. The second limitation is the lack of personalized measures and predictions. Although

research in foreign language acquisition has shown that there are many differences among non-native readers, recent studies in artificial intelligence make predictions based on aggregate data. The final limitation is that the majority of studies (74%) focus on English as a foreign language. Consequently, the goal of this doctoral research is a contextualized and personalized approach and a focus on Dutch and French as foreign languages.

The second part looks at two measures of lexical difficulty for non-native readers. On the one hand, it investigates how words are introduced in didactic reading materials labeled with CEFR levels. This study introduces a new graded lexical database for Dutch, namely NT2Lex (Tack et al., 2018b). The innovative feature of this database is that the frequency per difficulty level was calculated for the meaning of each word, disambiguated based on the sentence context. However, the results show that there are important inconsistencies in how etymologically related translations occur in the Dutch and French databases. Therefore, this difficulty measure does not yet seem valid as a basis for an automated system.

On the other hand, it is investigated how non-native speakers themselves perceive difficult words during reading. The perception of difficulty is important to predict because the learner's attention is a determining factor in the learning process (Schmidt, 2001). The study introduces new data for readers of French. An important goal of these data is to make correct predictions for all words in the text, which contrasts with studies in foreign language acquisition that focus on a limited number ($Mdn = 22$) of target words in the text. Moreover, the analyses show that the data can be used to develop a personalized and contextualized system.

The final section looks at two types of predictive models developed on the aforementioned data, namely mixed-effects models and artificial neural networks. The results validate the idea that perceptions of lexical difficulty can be predicted primarily on the basis of "word surprisal", a central concept in information theory. Furthermore, the analyses show that commonly used performance statistics (such as accuracy and F-score) are sensitive to individual differences in rates of difficulty. Because these are therefore not appropriate for comparing predictions correctly for different learners, the D and Phi coefficients are used. Moreover, the results clearly show that a person-

alized model makes significantly better predictions than a non-personalized model. On the other hand, the results show that a contextualized model can better discriminate difficulty, although these improvements are not always significant for each learner.

# CONTENTS


Abstract xi
List of Tables xxi
List of Figures xxv
List of Abbreviations xxxi
List of Symbols xxxv
Publications xxxvii

Introduction 1

1 RESEARCH SCOPE AND AIMS 3
1.1 Study Object . . . 6
1.1.1 Reading and Vocabulary in a Foreign Language . . . 6
1.1.2 Complexity and Difficulty of the Input . . . 8
1.1.3 Technological Development . . . 12
1.2 Scientific Framework . . . 14
1.3 Aims and Contributions . . . 16
1.3.1 Research Objectives . . . 17
1.3.2 Research Studies . . . 21
1.4 Structure of the Thesis . . . 23

I STATUS QUAESTIONIS 27

2 THE PREDICTION OF LEXICAL COMPETENCE IN FOREIGN LANGUAGE READING: A SYSTEMATIC SCOPING REVIEW 29
2.1 Previous Literature Reviews . . . 31
2.1.1 The Effect of Reading on Vocabulary Development . . . 31
2.1.2 The Effect of Vocabulary on Reading Comprehension . . . 33
2.1.3 Research Questions . . . 33
2.2 Method . . . 34
2.2.1 Identification . . . 36
2.2.2 Selection . . . 38

2.2.3 Extraction . . . 41
2.3 Delineating the Research Scope . . . 41
2.3.1 Publications . . . 41
2.3.2 Studies . . . 47
2.4 Lexical Competence as a Criterion Variable . . . 50
2.4.1 Measurements . . . 51
2.4.2 Predictors . . . 55
2.5 Discussion . . . 67
2.6 Conclusion . . . 69
2.A Appendix . . . 71
2.A.1 Search Documentation . . . 71
2.A.2 Exclusion Keywords . . . 80
2.A.3 Publications Included in the Synthesis . . . 81
2.A.4 Data Coding Variables . . . 94
2.A.5 Tools and Software . . . 99

3 METHODS FOR IDENTIFYING COMPLEX AND DIFFICULT WORDS IN READING: THEORETICAL, EMPIRICAL, COMPUTATIONAL 101
3.1 Theoretical Methods . . . 102
3.1.1 Complexity Thresholds . . . 102
3.1.2 Expert Simplifications . . . 105
3.2 Empirical Methods . . . 111
3.2.1 Implicit and Indirect Measures . . . 113
3.2.2 Explicit and Direct Measures . . . 118
3.3 Computational Methods . . . 123
3.3.1 Probabilistic Language Models . . . 124
3.3.2 Statistical Machine Learning . . . 126
3.4 Conclusion . . . 131

II MEASURING LEXICAL DIFFICULTY 135

4 A PRIORI KNOWLEDGE OF DIFFICULTY: WORD OCCURRENCE IN CEFR-GRADED READING MATERIALS 137
4.1 Background . . . 139
4.1.1 Lexical Frequencies Linked to a Difficulty Scale . . . 140
4.1.2 Semantic Complexity and Lexico-Semantic Networks . . . 142
4.1.3 Cognateness . . . 145

4.1.4 Research Hypotheses . . . 148
4.2 NT2Lex: A CEFR-Graded Lexicon for Dutch as a Foreign Language Linked to Open Dutch WordNet . . . 149
4.2.1 Resource Development . . . 149
4.2.2 Resource Description . . . 161
4.2.3 Web-Based Tools for Cross-lingual Search and Lexical Complexity Analysis . . . 164
4.3 A Complexity Analysis with Lexical Features and Norms . . . 166
4.3.1 Lexical Features . . . 167
4.3.2 Psycholinguistic Norms . . . 171
4.4 Averaged and Disambiguated Semantic Complexity . . . 177
4.4.1 Rank in the WordNet Hypernymy Tree . . . 177
4.4.2 Semantic Complexity Across Levels in NT2Lex . . . 181
4.5 A Cross-Lingual Comparison of Cognates . . . 188
4.5.1 Etymological Relatedness and Translation Equivalence 188
4.5.2 Cognates in NT2Lex and FLELex . . . 193
4.6 Discussion . . . 197
4.7 Conclusion . . . 201
4.A Appendix . . . 202
4.A.1 RESTful API . . . 202
4.A.2 Hypernymy Ranks . . . 204
5 A POSTERIORI KNOWLEDGE OF DIFFICULTY: PERCEIVED LEXICAL DIFFICULTY IN A NATURAL READING TASK 205
5.1 Learner Data Collection . . . 207
5.1.1 Trial 1 . . . 208
5.1.2 Trial 2 . . . 213
5.2 Learner Data Analysis . . . 218
5.2.1 Learner-Specific Distributions of Difficulty . . . 219
5.2.2 High and Low Semantically Constrained Contexts . . . 223
5.3 Conclusion . . . 229
5.A Appendix . . . 233
5.A.1 Reading Ease . . . 233
5.A.2 Reading Materials . . . 233
5.A.3 Power Analyses . . . 234

III PREDICTING LEXICAL DIFFICULTY 247

6 EXPLANATORY FACTORS OF DIFFICULTY: A GENERALIZED LINEAR MIXED MODEL WITH FIXED LEXICAL EFFECTS AND RANDOM LEARNER EFFECTS 249

6.1 Lexical Features . . . 252
6.1.1 Form . . . 253
6.1.2 Meaning . . . 257
6.1.3 Use . . . 258
6.1.4 Exposure . . . 264
6.1.5 Etymology . . . 265
6.2 Generalized Linear Mixed-Effects Models . . . 266
6.2.1 Definition . . . 266
6.2.2 Selection . . . 270
6.2.3 Analysis . . . 275
6.3 Conclusion . . . 287
6.A Appendix . . . 291
6.A.1 Computing Infrastructure . . . 291
6.A.2 Dependencies . . . 291
6.A.3 Commands . . . 292

7 DEEP LEARNING OF DIFFICULTY: A COMPARISON OF NEURAL NETWORKS WITH NON-SENSITIVE PERFORMANCE METRICS 293

7.1 Deep Learning Architectures . . . 296
7.1.1 Word and Character Embeddings . . . 297
7.1.2 Contextualized and Personalized Neural Networks . . . 302
7.2 Model Evaluation . . . 308
7.2.1 Sensitivity Analysis of Performance Metrics . . . 309
7.2.2 Repeated Measures Performance Comparisons . . . 315
7.3 Conclusion . . . 329
7.A Appendix . . . 333
7.A.1 Computing Infrastructure . . . 333
7.A.2 Dependencies . . . 334
7.A.3 Parameters and Hyperparameters . . . 334
7.A.4 Cross-Validation Setup . . . 338
7.A.5 Evaluation Metrics . . . 339

Conclusion 343

8 CONCLUSION 345

8.1 Main Conclusions and Implications . . . . . 346

8.1.1 Implications for Educational NLP and the CWI Shared Tasks . . . . . 348

8.1.2 Implications for SLA and CALL . . . . . 350

8.1.3 Implications for Foreign Language Teaching . . . . . 351

8.2 Limitations and Perspectives . . . . . 353

Bibliography 355

Summaries 407

Summary in French 409

Summary in Dutch 413

# LIST OF TABLES


CHAPTER 1 3

Table 1.1 The Common Reference Levels of the Global CEFR Scale for Reading . . . . . 10

CHAPTER 2 29

Table 2.1 Definition of the Criteria for the Scoping Review . . . 36

Table 2.2 List of Keywords Defined per Search Facet . . . . . 37

Table 2.3 Coding Scheme for Data Extraction . . . . . 95

CHAPTER 3 101

Table 3.1 Post-Hoc Comparisons of the Probability of Making Reading Errors on Words Targeted or Not Targeted for Simplification . . . . . 108

Table 3.2 Classification of Empirical Measures of Lexical Difficulty in Reading . . . . . 113

Table 3.3 Frequently Used Machine Learning Algorithms in the CWI Tasks . . . . . 128

Table 3.4 Frequently Used Features of Lexical Complexity in the CWI Tasks . . . . . 129

CHAPTER 4 137

Table 4.1 Conceptions of Depth of Vocabulary Knowledge . . . 143

Table 4.2 Number of Books, Document, and Tokens in the Corpus of CEFR-Graded Reading Materials for Dutch L2 152

Table 4.3 List of Simplified CGN Tags . . . . . 159

Table 4.4 Number of Lexical Entries in NT2Lex . . . . . 161

Table 4.5 Number of Lexical Entries per Level in NT2Lex . . . 163

Table 4.6 The Number of Word Senses, Polysemes, and Unique Synsets in NT2Lex . . . . . 164

Table 4.7 Median Values of Hypernymy Ranks for Novel Entries at Each Level in NT2Lex . . . 183
Table 4.8 Kruskal-Wallis Tests of Differences in Hypernymy Ranks Across CEFR Levels . . . 184
Table 4.9 DSCF Tests for Multiple Comparisons of Hypernymy Ranks Across CEFR Levels . . . 185
Table 4.10 Chi-Squared Tests of Trends in Proportions of Dutch-French Cognates Among New Entries per Level in NT2Lex and FLELex . . . 194
Table 4.11 Spearman and Kendall Correlations of the Difficulty Levels at which Dutch-French Cognates are First Introduced in FLELex and NT2Lex . . . 195

CHAPTER 5 205

Table 5.1 Reading Materials in Trial 1 . . . 209
Table 5.2 FLELex Vocabulary in Trial 1 . . . 210
Table 5.3 Cloze Procedure for Identifying High and Low Constraint Contexts in Trial 1 . . . 211
Table 5.4 Participants in Trial 1 . . . 212
Table 5.5 Reading Materials Trial 2 . . . 215
Table 5.6 FLELex Vocabulary in Trial 2 . . . 216
Table 5.7 Cloze Procedure for Identifying High and Low Constraint Contexts in Trial 2 . . . 217
Table 5.8 Participants in Trial 2 . . . 218
Table 5.9 Number of Observations and Percentages of Difficulty in Trials 1 & 2 . . . 220
Table 5.10 Agreement Between Participants in Trials 1 & 2 . . . 222
Table 5.11 Strength of Association Between High & Low Contextual Constraints and Difficult & Non-Difficult Words 224
Table 5.12 Percentages of Low Semantic Constraints Per Difficulty & Non-Difficulty . . . 225
Table 5.14 Reading Materials for Trial 1 . . . 236
Table 5.15 Reading Materials for Trial 2 . . . 240
Table 5.16 List of Words Found Difficult by All Learners Participating in Trial 1 . . . 242

Table 5.17 List of Words Found Difficult by All Learners Participating in Trial 2 . . . 243
CHAPTER 6 249
Table 6.1 What is Involved in Knowing a Word Receptively? . 252
Table 6.17 Feature and Model Selection . . . 273
Table 6.18 Generalized Linear Mixed-Effects Model of Perceived Lexical Difficulty . . . 277
Table 6.19 Probability of Difficulty and Non-Difficulty Estimated by the GLMM Model . . . 286
CHAPTER 7 293
Table 7.1 Deep Learning Architectures . . . 297
Table 7.2 $R^2$ and Pseudo-$R^2$ Values for Features of Complexity Explained by FastText Word Embeddings . . . 299
Table 7.3 Average Probability of Difficulty on Difficult and Non-Difficult Words . . . 318
Table 7.4 Mean and Median Performance on Tenfold Cross-Validation . . . 319
Table 7.5 Model Ablation on Tenfold Cross-Validation . . . 320
Table 7.6 Mixed-Effects Analysis of the Average Predicted Probability on Difficult Words . . . 322
Table 7.7 Mixed-Effects Analysis of Model Performance on Discriminatory Power . . . 323
Table 7.8 Mixed-Effects Analysis of Model Performance on Correlation . . . 324
Table 7.9 Performance on Tenfold Cross-Validation of a Fine-Tuned BERT Transformer . . . 328
Table 7.10 Summary of Results . . . 328

# LIST OF FIGURES


CHAPTER 1 3

Figure 1.1 An Illustration of a Reading Task Where the Non-Native Reader's Attention is Drawn to Boldfaced and Hyperlinked Difficult Words . . . . . 5

Figure 1.2 Normalized Frequencies of Word Occurrence in Reading Activities at Various Levels of Difficulty . . . . . 11

Figure 1.3 Identification of Complex Words from Subjective Judgments by Non-Native Speakers of English . . . . 13

Figure 1.4 A Computerized Dynamic Assessment System that Assists Lexical Inferencing During Reading . . . . . 15

Figure 1.5 Pipeline for Automated Lexical Simplification . . . . 16

Figure 1.6 Structure of the Thesis . . . . . 24

CHAPTER 2 29

Figure 2.1 PRISMA Flow Chart of Literature Identification, Selection, and Inclusion . . . . . 35

Figure 2.2 Number of Records per Year, Type, and Top 15 Sources of Publication . . . . . 42

Figure 2.3 Power Laws of Author Productivity and Citation Similarity . . . . . 43

Figure 2.4 Bibliographic Coupling Network of Publications Included in the Scoping Review . . . . . 45

Figure 2.5 Co-Citation Network of Authors Cited in the Publications Included in the Scoping Review . . . . . 46

Figure 2.6 Distribution of Target Languages Across Studies . . . 48

Figure 2.7 Heatmap of Native Languages in Learners of a Target Language . . . . . 49

Figure 2.8 Distributions of Proficiency and Education Levels Across Studies . . . . . 50

Figure 2.9 Criteria Used to Select Target Vocabulary . . . . . 51
Figure 2.10 Proportions of Stimulus and Response Values per Type of Procedure . . . . . 52
Figure 2.11 Measurements with a Selected Response . . . . . 56
Figure 2.12 Measurements with a Constructed Response . . . . . 58
Figure 2.13 Measurements with a Selected & Constructed Response 60
Figure 2.14 Self-Report Categories in the Vocabulary Knowledge Elicitation Scale . . . . . 62
Figure 2.15 Taxonomic Tree of Predictors per Type of Response Measurement . . . . . 63
Figure 2.16 UML Class Diagram of the Database Structure for Data Extraction . . . . . 94
CHAPTER 3 101
Figure 3.1 Illustration of a Simple English Wikipedia Edit History from the CW Corpus . . . . . 106
Figure 3.2 Alignment of Reading Errors on a Lexically Simplified Text with a Substitution . . . . . 109
Figure 3.3 Alignment of Reading Errors on a Lexically Simplified Text with a Deletion . . . . . 110
Figure 3.4 Artificial Illustration of Eye-Movement Measures and Metrics . . . . . 115
Figure 3.5 Artificial Illustration of the N400 Effect in EEG Brain Potentials . . . . . 117
Figure 3.6 Distribution of the Probability to Misread a Word Token in the Sample of Poor-Reading and Dyslexic Children . . . . . 121
Figure 3.7 Example of Highlighted Difficult Words in the CWI Shared Task 2018 . . . . . 123
Figure 3.8 Difficulties in Processing Garden-Path Sentences Predicted from Word Surprisal Values Computed with an Early Parser on a Simple PCFG . . . . . 126
CHAPTER 4 137
Figure 4.1 Flowchart of the NT2Lex Resource Development Process . . . . . 150

Figure 4.2 Illustration of the Analyses Performed by the Frog Tagger . . . 153
Figure 4.3 Illustration of the Output of the Word-Sense Disambiguation Tool . . . 154
Figure 4.4 Example of a Lexical Entry in Open Dutch WordNet . 155
Figure 4.5 Example of a Synset Entry in Open Dutch WordNet . 156
Figure 4.6 Fallback to Two Types of Sense Identifiers . . . 157
Figure 4.7 Illustration of Entries in NT2Lex with Adjusted Frequencies (CGN+ODWN Version) . . . 163
Figure 4.8 Screenshot of an Online Crosslingual Search Query in NT2Lex . . . 165
Figure 4.9 Screenshot of a Lexical Complexity Analysis with NT2Lex . . . 166
Figure 4.10 Comparison of NT2Lex-CGN and SUBTLEX-NL Frequencies . . . 167
Figure 4.11 The Interplay Between Frequency and Dispersion in NT2Lex . . . 168
Figure 4.12 Mean SUBTLEX-NL Frequencies per Level in NT2Lex-CGN . . . 170
Figure 4.13 Lexical Dispersion per Level in NT2Lex . . . 171
Figure 4.14 The Degree of Lexical Sophistication Across Levels in NT2Lex . . . 172
Figure 4.15 Polysemy and Synonymy in NT2Lex . . . 173
Figure 4.16 Age of Acquisition in NT2Lex . . . 174
Figure 4.17 Density of Age of Acquisition per CEFR Level in NT2Lex 175
Figure 4.18 Density of Word Concreteness per CEFR Level in NT2Lex . . . 176
Figure 4.19 Computation of Absolute, Relative, and Averaged WordNet Hypernymy Ranks . . . 178
Figure 4.20 Wilcoxon Pair-Wise Comparisons of Hypernymy Ranks in NT2Lex Before and After Disambiguation . 182
Figure 4.21 Identification of Cognates in NT2Lex and FLELex . . 190
Figure 4.22 Number of Dutch-French Cognates per Level in FLELex and NT2Lex . . . 193

Figure 4.23 Confusion Matrix of the Difficulty Levels at which Dutch-French Cognates are First Introduced in FLELex and NT2Lex . . . . . 196
CHAPTER 5 205
Figure 5.1 Percentages of Participants Judging a Word as Difficult in Trials 1 & 2 . . . . . 221
Figure 5.2 Predictions of Difficulty from a Logistic Mixed-Effects Model With Fixed Effects of Low (0) and High (1) Contextual Constraints . . . . . 228
Figure 5.3 Predictions of Difficulty from a Logistic Mixed-Effects Model With Fixed Effects of Contextual Surprisal . . 230
Figure 5.4 Monte Carlo Power Simulations of GLMM 5.1 on Trial 1 as the Number of Participants Increases . . . . . 244
Figure 5.5 Monte Carlo Power Simulations of GLMM 5.2 on Trial 2 as the Number of Participants Increases . . . . . 245
CHAPTER 6 249
Figure 6.1 Generalized Linear Mixed-Effects Model of Perceived Lexical Difficulty . . . . . 251
Figure 6.2 Null and Full Generalized Linear Mixed-Effects Models 270
Figure 6.3 Random Effects for Learners in the Prediction of Perceived Lexical Difficulty . . . . . 279
Figure 6.4 The Effects of Isolated and Contextual Word Surprisal for the Prediction of Perceived Lexical Difficulty . . . 281
Figure 6.5 Probability of Difficulty on Non-Difficult and Difficult Words Estimated by the GLMM Model . . . . . 287
CHAPTER 7 293
Figure 7.1 An Illustration of Word and Character Embeddings . 298
Figure 7.2 Convolutional Neural Network with Character Embeddings . . . . . 301
Figure 7.3 Deep Learning Architecture for CWI (De Hertog & Tack, 2018) . . . . . 303
Figure 7.4 Feedforward Neural Network Architecture . . . . . 305
Figure 7.5 Bidirectional Long-Short Term Memory Neural Network Architecture . . . . . 306

Figure 7.6 Model Selection on Held-out Tenfold Cross-Validation 308
Figure 7.7 Tenfold Cross-Validation with Stratified Sampling . . 309
Figure 7.8 Spider Plot of Changes in Constant Baseline Performance . . . . . 314
Figure 7.9 Probability of Difficulty on Difficult and Non-Difficult Words . . . . . 317
Figure 7.10 Discriminatory Power per Trial and per Proficiency Level . . . . . 326
Figure 7.11 Correlation per Trial and per Proficiency Level . . . . 327

# LIST OF ABBREVIATIONS

**ACL** Association for Computational Linguistics 38
**AGHQ** Adaptive Gauss-Hermite Quadrature 274, 275
**AI** artificial intelligence 4, 6, 13
**AIC** Akaike information criterion 272, 274
**ANN** artificial neural network 23
**AoA** age of acquisition 171, 174
**AOI** area of interest 115
**BERT** Bidirectional Encoder Representations from Transformers xxiii, 205, 207, 325, 328
**BIC** Bayesian information criterion 272
**BiLSTM** bidirectional long-short term memory 23, 296, 297, 302, 304, 307, 308, 315–324, 328–331, 346, 349
**CAF** complexity, accuracy, and fluency 10
**CALL** computer-assisted language learning xix, 14–16, 21, 41, 68, 69, 345, 350, 351
**CEFR** Common European Framework of Reference xxi, xxii, 10, 11, 21, 25, 116, 137–139, 141, 148, 151, 152, 159, 162, 168, 169, 177, 183, 185, 186, 196, 198–201, 209, 211, 214, 282, 347, 350–352
**CGN** Corpus Gesproken Nederlands 152, 158
**CI** confidence interval 277
**CLIL** Content and Language Integrated Learning ix
**CNN** convolutional neural network 300, 301, 320, 329, 330, 332, 350
**CWI** complex word identification vii, xix, xxviii, 13, 16, 19, 20, 22, 45, 103, 121, 127, 131, 254, 265, 288, 289, 294–296, 303, 345, 347–349
**DPC** Dutch Parallel Corpus 192
**DSCF** Dwass-Steel-Critchlow-Fligner xxii, 108, 183–186
**EEG** electroencephalography xxvi, 18, 117, 118, 125, 134, 231, 288, 346
**EFL** English as a foreign language 17, 20, 122, 295

**ERP** event-related potential 117, 125, 288

**FFNN** feedforward neural network 297, 302, 304, 308, 315–319, 322–324, 328–331

**GLM** generalized linear model 251, 266–269, 272–274

**GLMM** generalized linear mixed model xxiii, xxviii, 23, 249–251, 266, 267, 269, 273–276, 285–287

**HR** hypernymy rank 177, 179–181

**ICC** intra-class correlation coefficient 222, 223, 250, 277

**IG** information gain 283

**IP** information processing 8

**L1** native language 4, 20, 48, 53, 54, 64, 68, 69, 102, 122, 125, 132, 133, 137, 141, 171, 172, 188, 195, 196, 206, 221, 266, 282, 294, 321, 348, 350, 353

**L2** foreign language xxi, 3, 4, 6, 7, 9–11, 17, 20–23, 25, 29–31, 33, 34, 41, 48, 53, 54, 64, 65, 68, 70, 101, 104, 115–118, 122, 132–134, 137–145, 149, 151, 152, 166, 172, 174, 194, 195, 197, 205–209, 214, 250, 258, 259, 282, 284, 288, 289, 294, 295, 329, 330, 332, 345–347, 351–354

**LCP** lexical complexity prediction 127

**LCS** longest common subsequence 191

**LRT** likelihood ratio test 322–324

**LSTM** long short-term memory 304

**MCC** Matthews correlation coefficient 311

**MCQ** multiple-choice questionnaire 54, 64

**NED** normalized edit distance 191

**NLP** natural language processing xix, 6, 14–16, 21, 42, 68, 69, 141, 294, 345, 346, 348, 351

**NNS** non-native speaker 10, 349

**NT2** Dutch as a foreign language 21, 139, 152

**OCR** optical character recognition 151

**ODWN** Open Dutch WordNet 153, 154, 156, 157, 170

**OFAT** one-factor-at-a-time 313

**OLD20** orthographic Levenshtein distance 20 255, 256, 285

**OLS** ordinary least squares 299

**OMW** Open Multilingual WordNet 154, 156, 157, 192, 257

**OOV** out-of-vocabulary 164, 298, 299

**PCFG** probabilistic context-free grammar xxvi, 124–126
**PDP** parallel distributed processing 294
**ReLU** rectified linear unit 293, 302
**RLD** reference level descriptor 138, 198, 200, 201
**RNN** recurrent neural network 124–126, 296
**RQ** research question 34
**SD** standard deviation 95, 122, 171, 173, 209, 210, 215, 216, 220, 318–320, 328
**SE** standard error 227, 229, 277, 284, 322–324
**SFI** standard frequency index 140, 160, 167, 168, 209, 210, 214, 216, 259, 273, 276, 277, 282, 299, 300
**SLA** second language acquisition vii, xix, 14, 21, 41, 55, 67, 69, 345, 349–351
**SVM** support vector machine 153, 154
**TED** tree edit distance 191
**UD** universal dependencies 264, 283
**VIF** variance inflation factor 271, 277, 278, 299
**VKS** Vocabulary Knowledge Scale 55
**WSD** word-sense disambiguation 21, 137, 152–154, 156, 157, 167, 168, 177, 181, 187, 346

# LIST OF SYMBOLS

**beta** ($\beta$) standardized regression coefficient 277

**Carroll's D** ($D$) dispersion index 140, 160, 168, 259

**Carroll's U** ($U$) normalized frequency 140, 160, 167, 259

**frequency** ($F$) raw frequency of occurrence 140, 159, 163

**L** ($\mathcal{L}$) likelihood 272

**LL** ($\ell$) log-likelihood 255, 261, 277

**Mean** ($M$) arithmetic mean 51, 95, 171, 173, 209, 210, 214–216, 220, 221, 286, 318–320, 328, 349, 350

**Median** ($Mdn$) xii, 184, 186, 209, 210, 214–216, 286, 318–320, 328, 410, 414

**OR** ($e^{\beta}$) odds ratio 277

**Phi** ($\phi$) binomial correlation coefficient 223–225, 293, 311, 313–315, 319–321, 324, 328, 329, 331, 332, 339, 349, 350, *see* MCC

**Q1** ($Q_1$) first quartile 209, 210, 215, 216, 286

**Q3** ($Q_3$) third quartile 209, 210, 215, 216, 286

**surprisal** ($S$) the information content, or the reciprocal of the probability of a random event 124, 254, 255, 261, 262, 273, 276, 277, 280, 282, 285, 299, 300

**Tjur's D** ($D$) coefficient of discrimination 293, 312–315, 319–323, 325, 328, 339, 349

# PUBLICATIONS

The following publications were issued in relation to the present doctoral dissertation.

## PAPERS


De Hertog, D., & Tack, A. (2018). Deep learning architecture for complex word identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 328–334. https://www.aclweb.org/anthology/W18-0539

François, T., Volodina, E., Pilán, I., & Tack, A. (2016–May 28). SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 213–219. http://www.lrec-conf.org/proceedings/lrec2016/pdf/275_Paper.pdf

Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 1346–1354. https://www.aclweb.org/anthology/2020.lrec-1.169/

Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016a–May 28). Evaluating lexical simplification and vocabulary knowledge for learners of French: Possibilities of using the FLELex resource. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 230–236. http://www.lrec-conf.org/proceedings/lrec2016/pdf/544_Paper.pdf

Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016b–July 8). Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. *Actes de La 23ème Conférence*

*Sur Le Traitement Automatique Des Langues Naturelles (TALN'16)*, 221–234. https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/Papers/T22.pdf

Tack, A., François, T., Roekhaut, S., & Fairon, C. (2017). Human and automated CEFR-based grading of short answers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 169–179. https://www.aclweb.org/anthology/W/W17/W17-5018.pdf

Volodina, E., Borin, L., Pilán, I., François, T., & Tack, A. (2017, April). SVALex. En andraspråksordlista med CEFR-nivåer. In E. Sköldberg, M. Andréasson, H. Adamsson Eryd, F. Lindahl, J. Prentice, S. Lindström, & M. Sandberg (Eds.), *Svenskans beskrivning* (pp. 369–382). Göteborgs Universitet. https://gupea.ub.gu.se/bitstream/2077/52211/1/gupea_2077_52211_1.pdf

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A report on the complex word identification shared task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78. https://www.aclweb.org/anthology/W18-0507

## ABSTRACTS

Gala, N., David, C., Tack, A., & François, T. (2019, August 29). *Assessing vocabulary knowledge for learners of French as a foreign language: Accounting for L1 variability to go beyond the CEFR scale.* The 27th Conference of the European Association for Computer-Assisted Language Learning (EUROCALL 2019), Louvain-la-Neuve, Belgium.

Tack, A. (2019, November 15). *Towards a contextualization of complexity to better the prediction of lexical competence in L2 reading.* Colloquium "Broadening the Scope of L2 Complexity Research", VUB Campus Etterbeek, Brussels, Belgium.

Tack, A., François, T., Desmet, P., & Fairon, C. (2017, February 10). *Introducing NT2Lex: A machine-readable CEFR-graded lexical resource for Dutch as a foreign language.* Computational Linguistics in the Netherlands 27 (CLIN 2017), Leuven, Belgium.

Tack, A., François, T., Desmet, P., & Fairon, C. (2018a, May 18). *CEFR-based complex word identification for French and Dutch L2*. PLIN Linguistic Day 2018 on "Technological innovation in language learning and teaching", Louvain-la-Neuve, Belgium.

Tack, A., François, T., Desmet, P., & Fairon, C. (2018b, July 4). *Making sense of L2 lexical complexity with NT2Lex, a CEFR-graded lexicon linked to Open Dutch WordNet*. The XIXth International Computer Assisted Language Learning (CALL) Research Conference, Bruges, Belgium.

Tack, A., François, T., Desmet, P., & Fairon, C. (2019a, August 29). *The prediction of lexical competence in foreign language reading: A systematic synthesis* (Presentation). The 27th Conference of the European Association for Computer-Assisted Language Learning (EUROCALL 2019), Louvain-la-Neuve, Belgium.

Tack, A., François, T., Desmet, P., & Fairon, C. (2019b, July 2). *The role of cognate vocabulary in CEFR-based word-level readability assessment*. Vocab@Leuven International Conference, Leuven, Belgium.

## AWARDS

- OUTSTANDING REVIEWER AT ACL 2020
  Received as PC member of the track "Cognitive Modeling and Psycholinguistics" at the 58th Annual Meeting of the Association for Computational Linguistics.

- BEST POSTER PRESENTATION AT PLIN DAY 2018
  Received for the study *CEFR-based Complex Word Identification for French and Dutch L2* (See Chapter 4), which was presented at the symposium on "Technological innovation in language learning and teaching" (18 May 2018, Louvain-la-Neuve, Belgium).

- BEST PAPER AT TALN 2016
  Received for the precursory study that enabled this doctoral dissertation *Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de FLE*, which was presented at the 23ème Conférence

sur le Traitement Automatique des Langues Naturelles (08 July 2016, Paris, France).

# INTRODUCTION

Ardua molimur, sed nulla, nisi ardua, virtus:
Difficilis nostra poscitur arte labor.

— Ovidius, *Ars Amatoria II, vv. 537-38*

CHAPTER 1

# RESEARCH SCOPE AND AIMS

Reading is a central skill when learning any foreign language (L2). In an instructional setting, learners may be encouraged to improve their target language proficiency by reading texts and books as part of their curriculum. In a day-to-day setting, learners may often want to read texts written in a foreign language for information and pleasure. Furthermore, having fluent reading skills is critical considering the 'Information Age' in which we live today. Because we recurrently receive an extensive flow of information, much of which is text-based (e.g., news articles, e-mails, and administrative documents), the reading process has become widespread and fast-paced. When this information is in a foreign language, it may be challenging to keep up the pace as non-native readers are not always well-prepared for such extensive reading.

A first reason why non-native readers are ill-prepared is that there is an apparent reluctance to capitalize on knowledge in L2 reading instruction. When consulting the official curricula for language education in Belgium[1], we notice that the reading curriculum targets are rather vague. They require determining the subject, main idea, and main points and extracting relevant information from written texts. There are no particular directives on helping students understand unfamiliar words, which may hamper their overall reading comprehension. There seems to be a sense of optimism that the student

1 The official curricula can be found on https://onderwijs.vlaanderen.be/leerplannen for Dutch-speaking education and on http://www.enseignement.be/index.php?page=24737&navi=295 for French-speaking education.

can single-handedly overcome such difficulties. However, this should not be overlooked, as meta-analytic research has shown that successful reading comprehension is strongly associated with the non-native's grammatical and lexical knowledge (see Jeon & Yamashita, 2014).

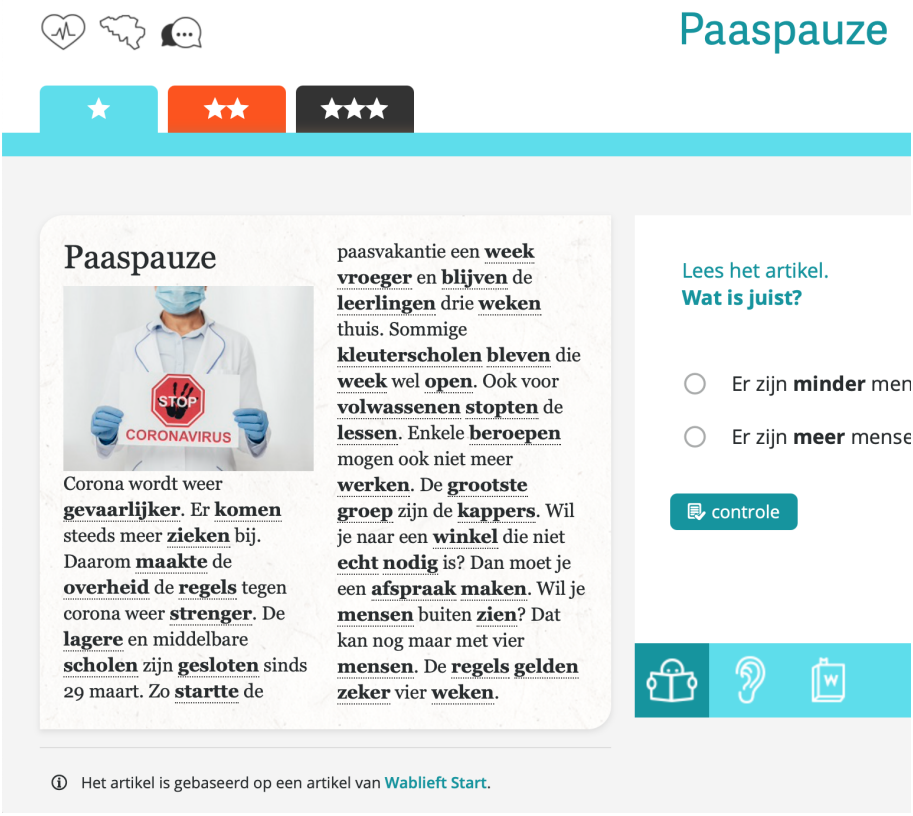
Another reason why non-native readers are ill-prepared is that it is less evident to identify difficulties in the reading process than, for instance, the writing process. In the development of written language fluency, much effort is put on enhancing the quality of texts written by non-natives in terms of accuracy (e.g., grammar errors and lexical adequacy) and complexity (e.g., syntactic structures and lexical variation). To this end, we currently have at our disposal several easy-to-use and AI-powered digital writing assistants, such as *Grammarly*[2], *Writefull*[3], and *Write & Improve*[4] (Yannakoudakis et al., 2018), to name but a few. Although these three applications focus on English, similar technologies are being developed for other languages too, for instance the writing assistant for Dutch L1 and L2[5] released by the Belgian public broadcasting service. Notably, a common functionality in all these digital assistants is that they automatically highlight language errors and, consequently, draw the non-native writer's attention to difficulties and common mistakes. As such, they mainly provide micro-level support and revisions (cf. Strobl et al., 2019).

In contrast, it is much less evident to support the reading process, at least in a similar analytical way. In computer-assisted language learning, there is much interest in developing digital reading assistants that provide access to electronic glosses and dictionaries (Abraham, 2008; Taylor, 2006; Yun, 2011). In Belgium, there exist several systems for Dutch L2, such as the CoBRA (Corpus-Based Reading Assistant)[6] plugin for Moodle (Deville et al., 2019) and the NedBox[7] platform. However, while it is (relatively) straightforward for a digital writing assistant to automatically highlight word choice errors for a specific learner, a digital reading assistant cannot simply highlight personal word decoding problems, as reading comprehension is a fundamentally implicit process. The NedBox platform, for example, draws the non-native reader's

2 https://www.grammarly.com
3 https://www.writefull.com
4 https://writeandimprove.com
5 https://schrijfassistent.be and https://nt2.schrijfassistent.be
6 https://researchportal.unamur.be/en/projects/cobra-corpus-based-reading-assistant
7 https://www.nedbox.be

**Figure 1.1**

*An Illustration of a Reading Task Where the Non-Native Reader's Attention is Drawn to Boldfaced and Hyperlinked Difficult Words*




*Note.* Source: https://www.nedbox.be/teaser/paaspauze

attention to difficult words (i.e., boldfaced and underlined; Figure 1.1), but these markings do not necessarily reflect personal word decoding problems. In computer-assisted reading support, difficult words are often marked in a manual and static way (cf. *infra*).

In conclusion, because current teaching practices do not prepare a non-native for extensive extramural reading, there is an avenue for technologies that provide better, more micro-level reading support. In particular, there is a critical need to foresee potential difficulties at the micro-level (e.g., predict word decoding problems) and nudge the non-native to use efficient help options (e.g., marginal glosses, online dictionary, and multi-media annotations).

In developing such intelligent computer-assisted reading support, machine learning (AI) and language engineering (NLP) may be the pieces that solve the puzzle.

A technology that can automatically alleviate potential language comprehension difficulties in L2 reading must solve a crucial issue:

> How can words in a reading text be automatically predicted as difficult for a foreign language (L2) learner?

The task of automatically predicting lexical difficulty in L2 reading is the subject of this doctoral research. The road map for the doctoral research project is set out in the remainder of this introductory chapter. The first section introduces the study object (Section 1.1), which revolves around three aspects: (a) the interplay between reading and vocabulary in a foreign language, (b) the complexity of the input, and (c) how this difficulty can be analyzed and predicted technologically. The second section gives a bird's eye view of the scientific framework (Section 1.2). The third section introduces two research objectives and three research studies (Section 1.3). The fourth and final section outlines the structure of the thesis (Section 1.4).

## 1.1 STUDY OBJECT

### 1.1.1 *Reading and Vocabulary in a Foreign Language*

Throughout the various stages of literacy development, one can distinguish different types and functions of reading. These range from storybook reading to preliterate children and emergent readers (Carger, 1993; Roberts, 2008) to isolated word reading (Wang & Koda, 2005), individual vocal or silent reading, intensive and pleasure reading (Beglar et al., 2012), extensive reading (Coady, 1996), narrow reading (Krashen, 1981, 2004), and academic reading (Parry, 1991). What is more, reading skills also serve an essential function in other tasks such as, for instance, when one reads subtitles or captions while watching videos in a foreign language (Montero Perez et al., 2014; Montero Perez et al., 2013). While this variety of purposes underscores the broader relevance and applicability of researching foreign language reading, it also makes a doctoral study that encompasses all possible forms and functions of reading practically

impossible. This dissertation focuses on the individual silent reading of texts in a foreign language because it is perhaps the most general and pervasive form of reading.

From a simple view, reading ability is composed of word decoding and language comprehension, $R = D \times C$ (Hoover & Gough, 1990). The *word decoding* component, as defined by Hoover and Gough (1990), is "the ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon, and thus, the retrieval of semantic information at the word level" (p. 130). This first component, also commonly referred to as *lexical access* or *word recognition*, is a critical part of the reading process in a foreign language. While we may think of decoding disfluencies occurring typically in readers with dyslexia, non-natives readers can also face word recognition problems, especially given the potential cross-linguistic interference with their native language (Hamada & Koda, 2011; Koda, 1996). The *language comprehension* component, on the other hand, is "the ability to take lexical information (i.e., semantic information at the word level) and derive sentence and discourse interpretations" (Hoover & Gough, 1990, p. 131). This second component is also sometimes referred to as *word-to-(con)text integration*.

Even in this simple view of reading, it is clear that vocabulary plays a fundamental role. As Stanovich (1986) seminally noted, "vocabulary knowledge is involved in a reciprocal relationship with reading ability (...) that continues throughout reading development and remains in force for even the most fluent adult readers" (p. 379). Various findings in L2 research corroborated Stanovich's observation of a "reciprocal causation" in the development of reading and vocabulary, although most evidence has been of a correlational rather than experimental nature (cf. Jeon & Yamashita, 2014). As a general rule, there appears to be a positive correlation between vocabulary knowledge and reading comprehension: the larger the percentage of known words in a text, the better the text will be understood (Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Some studies advanced the existence of lexical thresholds of 95% to 98% of known words necessary for adequate and optimal reading comprehension (Laufer & Ravenhorst-Kalovski, 2010), whereas other studies called into question the existence of such thresholds (Schmitt et al., 2011).

Although we may view reading as a combination of decoding and comprehension, with vocabulary playing an essential part, this view of the reading process does not allow us to understand its learning potential. From the perspective of one of the most influential learning theories, namely information processing (IP) (McLaughlin, 1987; Skehan, 1998), the learning process takes place in three stages. During the first stage (viz., input), a learner receives new information. In reading, this new information refers to the textual input, which may include either a small or large proportion of unknown words. During the second stage (viz., central processing), the learner processes the input in either a controlled or automatic way and may restructure his/her knowledge. In reading, it is evident that this central processing largely depends on word decoding and language comprehension. During the third and final stage (viz., output), the learner produces a meaningful production. In reading, this output may refer to the learner's ability to recall some words after reading. An essential step in this entire process is, however, the perception of the input. Without the learner's attention towards the input (Schmidt, 2001), there is no learning potential. Therefore, we could reappraise the simple view of reading to include three, instead of two, pivotal components: perception (with attention), decoding, and comprehension.

### 1.1.2 *Complexity and Difficulty of the Input*

In foreign language development, much depends on the type of input to which a non-native is exposed. Krashen's (1989) influential Input Hypothesis, for instance, states that language is best acquired incidentally through meaning-focused activities (i.e., reading) rather than through explicit instruction (i.e., rote learning and corrective feedback). Although subsequent research has disproved Krashen's claim that we acquire vocabulary best through reading (e.g., see Laufer, 2003; Raptis, 1997), current research is still very much guided by his idea that language acquisition requires comprehensible input. From a Krashenian point of view, the input is sufficiently comprehensible if it is at the level of $i + 1$, that is, only slightly exceeding the learner's level of competence $i$ (Krashen, 1978). In reading, the comprehensibility of the input is related to the notion of readability (i.e., the degree to which a text is easy to

read). Readability can be established either by looking at the input's complexity from a linguistic perspective or observing how the reader processes the input from a psychological perspective.

In their extensive synthesis of the literature, Kortmann and Szmrecsanyi (2012) provide several definitions of linguistic complexity. First of all, the complexity of a language can be located either at the level of the entire linguistic system (i.e., global complexity) or at the phonological, morphological, syntactic, lexical, semantic, or pragmatic systems (i.e., local complexity). The readability of the textual input has been mainly established based on the latter. More specifically, almost all classical formulae integrate a measure of syntactic and lexical complexity: a text is easier to read if it contains many small frequent words embedded in short sentences (Coleman & Liau, 1975; Dale & Chall, 1948; Flesch, 1948; Gunning, 1952; Mc Laughlin, 1969).

In these classical readability formulae, lexical complexity is assessed either at the level of the text or at the individual word level. The Flesch Reading Ease formula (Flesch, 1951) and the Coleman-Liau index (Coleman & Liau, 1975), for instance, compute the average number of syllables or letters per word. As such, these formulae measure the degree of lexical complexity for the text as a whole. Conversely, although the Gunning Fog index (Gunning, 1952), the SMOG formula (Mc Laughlin, 1969), and the Dale-Chall formula (Dale & Chall, 1948) also compute the proportion of complex words for the text as a whole, this ratio is based on a heuristic which classifies a word as complex based on either a length-based or frequency-based decision threshold. As such, these formulae incorporate a measure of complexity at the level of the individual word. Now, in L2 reading, achieving an adequate assessment of complexity at the word level is crucial. As mentioned previously, it unquestionable that the incomprehensibility of a reading text is contingent on the number of unknown words it contains. However, it is questionable whether the simple thresholds used in the classical formulae (e.g., a word is complex if it counts three or more syllables) can adequately identify the number of complex words in this respect.

In addition to the global–local dichotomy, Kortmann and Szmrecsanyi (2012) also distinguish between developmental and acquisitional complexity. On the one hand, acquisitional complexity refers to the various aspects of a language

**Table 1.1**

*The Common Reference Levels of the Global CEFR Scale for Reading*

| User | Level | Description |
|---|---|---|
| Proficient | C2 | Can understand with ease virtually everything heard or read. |
| | C1 | Can understand a wide range of demanding, longer texts, and recognise implicit meaning. |
| Independent | B2 | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. |
| | B1 | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. |
| Basic | A2 | Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). |
| | A1 | Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. |

*Note.* This table only includes the descriptions that pertain to language understanding. For a complete description, see https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale.

that remain non-acquirable to an adult non-native speaker (NNS). This type of complexity relates to the Fossilization and Selective Fossilization Hypotheses in L2 research (Han, 2009, 2013; Selinker, 1972), aiming to apprehend why some NNSs display persistent errors, even after years of practice. On the other hand, developmental complexity refers to the moment at which more advanced aspects of the target language can be acquired and used. However, researchers have studied this type of complexity in light of L2 performance and proficiency (cf. CAF research; Housen & Kuiken, 2009) rather than reading.

Nevertheless, several readability measures link the complexity of a text to different stages of L2 development. These measures are established based on materials intended for reading instruction and included in textbooks for specific levels in the L2 language curriculum. An educational scale commonly

**Figure 1.2**

*Normalized Frequencies of Word Occurrence in Reading Activities at Various Levels of Difficulty*

| word | tag | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|---|
| épais | ADJ | 0.83 | 15 | 22.0 | 13.6 | 6.00 | 0 |
| épaisseur | NOM | 0 | 0 | 3.6 | 6.1 | 6.9 | 0 |
| épanchement | NOM | 0 | 0 | 0 | 0 | 0 | 19.3 |
| épandre | VER | 0 | 0 | 0 | 2.28 | 6.00 | 0 |
| épanouir | VER | 0 | 0 | 0.64 | 8.40 | 6.00 | 12.8 |
| épanouissement | NOM | 0 | 0 | 0 | 2.28 | 19.9 | 19.3 |
| épargne | NOM | 0 | 0 | 0.64 | 0 | 0 | 0 |

*Note.* This excerpt is taken from FLELex, a graded lexicon for French L2 based on the six-point CEFR scale (Francois et al., 2014).

used in the L2 curriculum is the Common European Framework of Reference (CEFR) scale (see Table 1.1; Council of Europe, 2001). Based on a corpus of materials calibrated to the CEFR scale, several L2-specific readability formulae exist for a variety of languages, including French (François & Fairon, 2012), Dutch (Velleman & van der Geest, 2014), Chinese (Sung et al., 2015), Swedish (Pilán, Vajjala, et al., 2016), and English (Uchida & Negishi, 2018). Through the integration of various features of syntactic and lexical complexity (among others), these formulae correlate specific learning stages in the L2 with a degree of readability at the text level. Moreover, researchers have also used the same graded materials to examine the distribution of words in reading texts intended for various stages of L2 learning (see Figure 1.2; Francois et al., 2014).

Lastly, Kortmann and Szmrecsanyi (2012) also distinguish between absolute and relative complexity. Absolute complexity refers to the complexity of a language understood in theoretical terms. In contrast, relative complexity refers to the degree of linguistic complexity that one can only understand regarding a particular language user's experience. In this way, the research focus shifts from a purely linguistic perspective on complexity towards a more psycholinguistic and psychological perspective. Instead of defining the complexity of a particular word in absolute terms, the idea is to derive a measure of readability based on our observations of the complexities typical of

a target reader population or even idiosyncratic for a given reader. Researchers have also referred to this type of complexity as *cognitive complexity* or *difficulty* (Housen & Kuiken, 2009). In the remainder of this dissertation, the term *complexity* will refer to the linguistic aspects of the textual input, whereas the term *difficulty* will refer to the input's cognitive impact on a particular reader.

### 1.1.3 *Technological Development*

Because the input's complexity and difficulty may inhibit a successful reading comprehension, it is essential to provide adequate learning support. A conventional way the learning-through-reading process could be made more efficient is through instructional intervention, by supplementing a reading task with exercises or explicit instruction (e.g., see Paribakht & Wesche, 1996). Another way is to develop intelligent assistive technologies. These technologies include systems that make the input more comprehensible through:

- input selection (e.g., adaptive content sequencing; Wang, 2016),
- input modification (e.g., lexical simplification; Bingel, Paetzold, et al., 2018), and
- input enhancement (e.g., access to electronic glosses; Abraham, 2008; Yun, 2011).

For these technologies to function correctly, it is necessary to accurately analyze and predict potential complexities and difficulties.

For the automatic analysis of lexical complexity, on the one hand, various tools have been proposed in the literature. The most well-known tool is the Coh-Metrix tool (Graesser et al., 2004), which includes 108 features of syntactic, lexical, and discursive complexity. Another oft-used tool is the Lexical Complexity Analyzer (Lu, 2012), which includes 25 features of lexical density, sophistication, and variation. More recently, Chen and Meurers (2016) developed the web-based Common Text Analysis Platform (CTAP). An advantage of these tools is that they aim to cover many relevant features of complexity. However, a downside is that they generally provide measures of lexical complexity for the text as a whole instead of measuring complexity at the word level.

**Figure 1.3**

*Identification of Complex Words from Subjective Judgments by Non-Native Speakers of English*

Leo, on December 23, took an **oath** of **purgation** concerning the **charges** brought against him, and his opponents were **exiled**.

*Note.* This example is taken from Paetzold and Specia (2016a, p. 564). A word is complex (boldface) if at least one non-native finds the word difficult.

The automatic prediction of lexical difficulty, on the other hand, can be achieved by adopting a supervised machine learning approach. The idea is to develop a system that performs either one of two tasks: a binary classification task, classifying each word as either difficult ($y = 1$) or non-difficult ($y = 0$); or a logistic regression task, emitting a probability of difficulty $P(y = 1) \in [0, 1]$ for each word in the text. These tasks adopt a supervised machine learning approach because the predictions are optimized based on empirical, gold-standard data where each word in a reading text has a measurement of difficulty. Based on this data, either one or an ensemble of learning algorithms are optimized to predict difficulty automatically. Finally, the predictions' accuracy is assessed with standard performance metrics (e.g., the $F_1$ score).

The task of developing such artificial intelligence (AI) systems that can automatically identify complex words in a text (Shardlow, 2013a) is called complex word identification (CWI). During past CWI shared tasks (Paetzold & Specia, 2016a; Yimam et al., 2018), various systems were trained on data with subjective judgments of complexity (see Figure 1.3). However, because the ultimate performance that these systems could achieve depended on how the data were collected, some researchers stressed the need for further research (Finnimore et al., 2019; Zampieri et al., 2017). Moreover, it should be noted that the term "complex word identification" leads to a certain degree of terminological confusion. The term may refer either to the identification of morphologically complex words from a linguistic perspective (e.g., see Meunier et al., 2008) or words that are complex for a specific reader. Because the latter indicates *relative* complexity, *cognitive* complexity, or *difficulty*, this task will hitherto be referred to as the identification of difficult words. However, the acronym CWI will still be used to refer to the shared tasks.

## 1.2 SCIENTIFIC FRAMEWORK

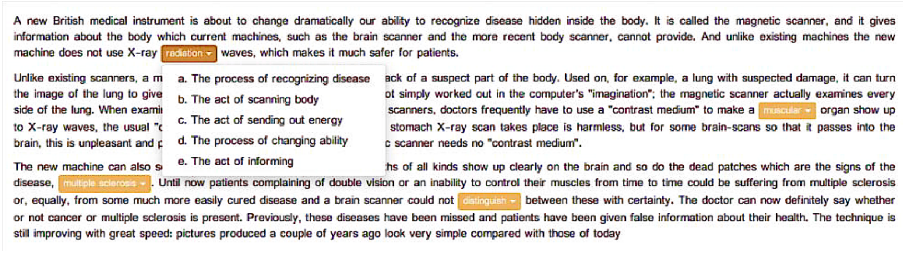
The study object introduced in the previous section calls for an interdisciplinary research focus. On the one hand, as the study relates to foreign language reading, it is essential to look for relevant research in second language acquisition (SLA) and computer-assisted language learning (CALL). On the other hand, as the study pertains to the automated prediction of difficult words, it is essential to look for recent advances in natural language processing (NLP) on this topic.

In the field of SLA, most relevant studies examine various predictors of incidental vocabulary acquisition through reading (i.e., the factors that explain whether non-natives will acquire the form and meaning of unknown words while reading; inter al. Adams, 1982; Ender, 2016; Pigada and Schmitt, 2006; Pulido, 2003; Pulido and Hambrick, 2008; Rott, 1999; Waring and Takaki, 2003; Watanabe, 1997; Webb, 2007, 2008; Zhao et al., 2016). To this end, researchers conventionally adopt an experimental pre-post test design. Before the reading task, they administer a pretest to establish a set of unknown words manually. After the reading task, they administer one or more immediate and delayed posttests to assess vocabulary acquisition and retention. Finally, they compare the learning effect of several experimental conditions (e.g., number of repeated exposures to the word) with inferential statistics.

In addition to measuring lexical knowledge before and after reading, other SLA studies examine contextual word learning while reading (Elgort et al., 2018). Many of these studies focus on lexical inferencing tasks (inter al. Bensoussan & Laufer, 1984; Hamada, 2015; Paribakht, 2005; Pulido, 2007; Shen, 2008). Because this procedure explicitly measures the ability to recognize or recall words' meaning, it elicits evidence for the word decoding component of the reading process. Regarding the perception of the input, other studies measure how non-natives notice and pay attention to unknown words while reading through eye-tracking procedures (Godfroid et al., 2013). However, as will be concluded in the systematic literature review in Chapter 2, the use of such *online* procedures remains relatively scant compared to the conventional use of *offline* pretests and posttests.

**Figure 1.4**

*A Computerized Dynamic Assessment System that Assists Lexical Inferencing During Reading*



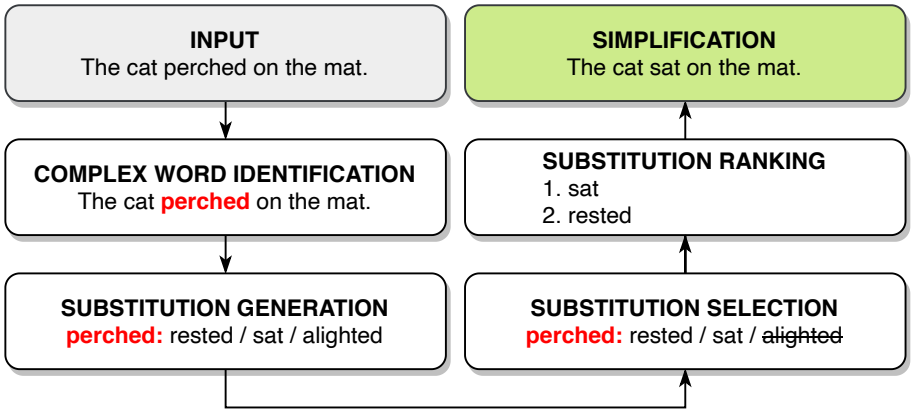

*Note.* This screenshot is taken from Ebadi et al. (2018, p. 799, Figure 1). The researchers manually identified four target words (highlighted) as unfamiliar based on frequency bands and a vocabulary pretest. The learners were asked to use the system to infer the meaning of the highlighted words with a multiple-choice question.

In CALL, various studies examine the learning effectiveness of computer-assisted reading support to enhance the noticing and decoding of unknown words while reading. Some researchers examine the effectiveness of salience (i.e., highlighting) as a technique to draw the reader's attention to electronic glosses (De Ridder, 2000, 2002) or of a computerized dynamic assessment system that gives customized hints and cues for lexical inferencing (see Figure 1.4; Ebadi et al., 2018). The vast majority of CALL research focus, however, on how incidental vocabulary acquisition can be enhanced by giving access to electronic glosses and dictionaries (AbuSeileek, 2011; AbuSeileek, 2008; Chen, 2016; Chen & Yen, 2013; Hu et al., 2014; Lee & Lee, 2015; Lee, Warschauer, et al., 2017) as well as by providing multimodal help options (Rouhi & Mohebbi, 2013; Tabatabaei & Shams, 2011; Türk & Erçetin, 2014; Yanguas, 2009; Yoshii, 2006). Several meta-analyses show medium to large positive effects on vocabulary acquisition and reading comprehension, although these effects depend on learner and task variables (Abraham, 2008; Taylor, 2006; Vahedi et al., 2016; Yun, 2011).

Finally, when it comes to investigating how technology can effectively assist the reading process, there are some notable parallels and differences between the CALL and NLP fields. In both fields, researchers take an interest in how technology can make difficult words more comprehensible, either by enhancing the input with help options in the case of the former (cf. *supra*) or

**Figure 1.5**

*Pipeline for Automated Lexical Simplification*




*Note.* This illustration is taken and adapted from Paetzold and Specia (2017, p. 551).

by automatically simplifying difficult words with more accessible synonyms in the case of the latter (see Figure 1.5; Paetzold and Specia, 2017). In automated lexical simplification, it is crucial to develop a system that can accurately identify difficult words (cf. the CWI task; Section 1.1.3). Significant errors in the simplification pipeline may result from an unnecessary identification of difficult words (Shardlow, 2014). However, research in the fields of CALL and NLP significantly differ in how they identify difficult words. Whereas this task is done automatically in NLP research, the sets of difficult target words are often manually selected in CALL research. Nevertheless, the automated identification of complicated words is also relevant for CALL if the ambition is to have fully automated computer-assisted learning support.

## 1.3 AIMS AND CONTRIBUTIONS

While the automated identification of difficult words may have originated as an initial step in text simplification research, it seems that the scope, relevance, and purpose of this task are much broader. Therefore, this doctoral study aims to examine this task as a separate research topic and a fully-fledged task relevant for any form of computer-assisted learning support. The dissertation contributes to this topic with several research objectives and studies.

### 1.3.1 *Research Objectives*

The doctoral study's core research objectives are twofold: adopting both a contextualized and personalized approach to the prediction of lexical difficulty in L2 reading. Additionally, a secondary objective is to research other target language populations than EFL learners.

*A Contextualized Approach to Measuring and Predicting Difficulty*

Reading is, at its core, a meaning-focused activity (cf. Krashen, 1989), and an essential ingredient for constructing meaning is the context that surrounds a word. We know this, of course, from seminal theoreticians in general linguistics, including Firth's (1957) oft-cited adage "You shall know a word by the company it keeps" (p. 11) and Harris (1954) viewing meaning as a function of language distribution. Consequently, the process of reading articles, books, or any other type of text entails the construction of meaning from context, as individual words appear in sentences and paragraphs.

There are two approaches to measure how difficult words are processed while reading: contextualized or decontextualized. These can be perhaps best exemplified by citing Weaver's (1955) influential observation from his memorandum:

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which.
>
> But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word. (pp. 20–21)

Similarly, when measuring difficulty at the word level in reading, one may adopt either one approach: one may look only at the target word in isolation – as through a hole in an opaque mask covering the rest of the text – or one may opt to account for the context surrounding the word.

Recent studies on the automated prediction of lexical difficulty in reading have adopted either one approach. On the one hand, several studies modeled lexical difficulty in isolation. These studies used a variety of methods, including a list of words graded with a Likert scale (Maddela & Xu, 2018), vocabulary tests (Ehara, 2019, 2020; Lee & Yeung, 2018a), comparative judgments (Gooding et al., 2019), and a lexicon of word frequencies in graded reading materials (Alfter & Volodina, 2018; Tack et al., 2016a). Although we may use such decontextualized measurements to predict lexical difficulty, this prediction is made for each word in isolation, as if looking through a hole in an opaque mask covering the rest of the text. On the other hand, some studies accounted for the context surrounding the individual word. These studies resorted to measures of lexical difficulty in context (e.g., eye movements, brain signals, and subjective perceptions while reading; inter al. Elgort et al., 2018; Paetzold and Specia, 2016a; Shardlow, 2013a; Yimam et al., 2018) and examined the effect of contextual cues (e.g., Godfroid et al., 2013), contextual constraints (e.g., Chen et al., 2017), and the degree to which the use of a word is surprising in a given context (e.g., Frank et al., 2013, 2015).

Given the reading task's contextualized nature, it seems more appropriate to adopt a contextualized approach to predicting lexical difficulty. However, the existence of both contextualized and decontextualized approaches in current research raises a crucial question:

RQ1.1 Does a contextualized approach lead to more accurate measurements and predictions of lexical difficulty for foreign language readers than a decontextualized one?

There is some evidence in the earlier cited eye-tracking and EEG studies on contextual cues, constraints, and word surprisal that supports the contextualized approach (see Chapter 3). However, there are still other types of data for which the added value of a contextualized approach is worth investigating. The first type pertains to word frequency measured in reading materials labeled with difficulty levels. To establish such a frequency measure, we generally count the number of times an inflected or canonical word form occurs in a series of texts. However, this is a decontextualized measure of difficulty as it does not disambiguate the meaning of the word from the surrounding context. The second type pertains to the subjective perceptions of lexical difficulty by

non-natives while they are reading. Although this is a contextualized measure of difficulty, most predictive models trained on this type of data integrate decontextualized features of complexity (e.g., word length and frequency) and use learning algorithms that do not consider the surrounding context (e.g., random forests and feedforward neural networks). A first objective will therefore be to contextualize these measurements and predictions.

*A Personalized Approach to Measuring and Predicting Difficulty*

Reading is, in essence, a cognitive process. Much of our understanding of the reading process therefore derives from the field of psychology, which is nevertheless characterized by two divergent scientific approaches, as noted by Dörnyei (2005):

> Ever since the early days of its existence, the field of psychology has been trying to achieve two different and somewhat contradictory objectives: to understand the *general principles* of the human mind and to explore the *uniqueness* of the individual mind. The latter direction has formed an independent subdiscipline within the field that has traditionally been termed *differential psychology* but recently more frequently referred to as *individual difference research*. (p. 1)

These two contradictory forces are also manifest in how non-natives perceive and attend to the input (Schmidt, 2001). While attention is contingent on several general principles, such as the fact that all learners have a limited and selective attention span, attention is also subject to individual differences between learners, such as motivation, learning strategies, and aptitude.

Recent studies on the automated prediction of lexical difficulty in reading have also adopted either a generalized or personalized approach. On the one hand, most CWI studies used data aggregated from subjective perceptions of difficulty by a group of readers (e.g., see Shardlow, 2013b; Yimam et al., 2017a), whereas other studies used personalized data consisting of decontextualized vocabulary lists and tests (Ehara, 2019, 2020; Lee & Yeung, 2018b; My et al., 2017). Some studies, such as Paetzold and Specia (2016a), adopted a hybrid approach. They trained a predictive model for non-natives in general and

evaluated its predictions for individual learners. Their data consisted of a training set of aggregated difficulty judgments and a test set of personal difficulty judgments.

Because individual differences may influence how a learner perceives the input, it seems more appropriate to adopt a personalized approach to predicting difficulty perceived by L2 readers. However, the fact that most CWI papers currently adopt a generalized approach raises a crucial question:

RQ1.2 Does a personalized approach lead to a better prediction of lexical difficulty perceived by non-native readers?

As a second research objective, this study aims to adopt a personalized approach to measuring and predicting lexical difficulty in L2 reading.

*French and Dutch L2*

A secondary research objective is concerned with the language of study. As will be evident from the systematic review of the literature (Chapter 2), most previous research (72%) has targeted English as a foreign language (EFL), while a more limited number of studies targeted learners of Spanish, German, French, Chinese, and Japanese as a foreign language. This abundance of studies on English is also prevalent in other topics in L2 teaching research. In their review, Gurzynski-Weiss and Plonsky (2017) found that most interaction research (78%) focused on either learners of English or on native speakers of English learning a foreign language. Similarly, most recent work on the identification of complex words focused on English, with difficulty labeled either in context (Paetzold & Specia, 2016a; Shardlow, 2013b) or in isolation (Ehara, 2019, 2020; Gooding et al., 2019; Lee & Yeung, 2018b; Maddela & Xu, 2018; My et al., 2017). Besides English, some studies collected training data for German and Spanish (Yimam et al., 2017a, 2017b), test data for French (Yimam et al., 2018), as well as data for Chinese (Lee & Yeung, 2018a; Lee, Liu, et al., 2017) and Korean (Yancey & Lepage, 2018), albeit with a mix of L2 and L1 participants. Because of the prevalence of studies conducted on English, this dissertation will study other foreign languages, and more specifically, French and Dutch.

### 1.3.2 *Research Studies*

The doctoral thesis comprises three research studies: (a) a systematic literature review, (b) an exploratory study of word occurrence in reading materials labeled with difficulty levels; (b) a machine learning study on the prediction of difficult words in a text for a foreign language reader.

*Study 1 – A Systematic Scoping Review on The Prediction of Lexical Competence in L2 Reading*

The doctoral investigation's starting point is a systematic inquiry into previous studies relevant to the research topic. Although there exist several literature reviews on the interplay between vocabulary and (computer-assisted) reading (inter al. Abraham, 2008; Huckin & Coady, 1999; Hunt & Beglar, 2005; Melby-Lervag & Lervag, 2014; Raptis, 1997; Taylor, 2006; Tsai, 2017; Yun, 2011), there does not exist a systematic review that comprehensively addresses the prediction of lexical difficulty in L2 reading. The first contribution is therefore a systematic literature review that brings together the fields of SLA, CALL, and NLP. The review includes empirical studies on the construct of lexical competence predicted either with inferential statistics or supervised machine learning.

*Study 2 – Measuring Lexical Difficulty from Reading Materials Graded Along the CEFR Scale*

The second contribution is an exploratory study of L2 reading materials calibrated to different proficiency levels. The study looks at newly introduced words in reading activities at a specific CEFR level. As such, the study contributes to previous work on the creation of lexical resources from CEFR-graded reading materials (cf. *supra*; Francois et al., 2014). In particular, this study's outcome is a new resource for Dutch as a foreign language (NT2), namely NT2Lex (Tack et al., 2018b). Through the development of this new resource, the study seeks to enhance the existing methodology. Instead of tallying the occurrence of words by their form (more specifically, by their lemma), words are tallied by distinguishing their meaning from the surrounding sentence context with automatic word-sense disambiguation (WSD). Each sense

identifier is subsequently linked to its concept in the Open Dutch WordNet semantic network, which allows for a representation of lexical difficulty at the conceptual level. Lastly, to determine the potential use of this data as a measure of lexical difficulty, several characteristics of lexical complexity are investigated, including the degree of conceptual specificity and genericity in the hypernymy tree as well as cognate status.

*Study 3 – Predicting Perceived Lexical Difficulty in Foreign Language Reading*

The third and final contribution is a study on the development of a predictive model of perceived lexical difficulty. A supervised machine learning procedure is adopted, which is composed of two key steps: (a) to establish a reference measure of lexical difficulty for foreign language readers and (b) to develop a predictive model on the basis of this reference measure.

The first stage of this study is the collection of data with measurements of lexical difficulty in reading for French L2. The adopted methodology is similar to the previous CWI tasks in that a sample of non-native readers are asked to identify (highlight) difficult words while they are reading. In total, the data counts 261,942 observations from a sample of 56 learners. An important feature of this data set is that each individual learner is uniquely identified instead of having aggregated judgments of difficulty for the group of learners.

It should be noted that this measure accounts for how non-natives themselves *perceive* difficult words. It does not strictly account for what learners genuinely have difficulty with in terms of word decoding and comprehension. As noted by Laufer and Yano (2001), learners do not always take notice of words that are unfamiliar to them while reading. It could be that they have difficulty decoding a word while reading, but that they do not necessarily perceive this difficulty. Consequently, when predicting lexical difficulty in foreign language reading, there are two steps: (a) to predict whether a word will be perceived as difficult by a given learner and (b) to predict whether a word will be difficult based on the learner's state of knowledge. When a system is able to perform both steps, this technology can then automatically direct the learners' attention towards difficult words they may not perceive as difficult. Because it is key to predict the perception of the input as it prompts

the learning process (cf. *supra*), this study focuses on the first step, that is on the development of a predictive model of perceived difficulty.

The second stage of this study is the development of a generalized linear mixed model (GLMM). A stepwise regression analysis is performed with a view to identifying significant predictors of perceived lexical difficulty. First, various indicators of lexical complexity are automatically extracted at the word level. These features pertain to the form, meaning, and use of the word, as well as to the learner's exposure to the word during the task. Second, a two-level GLMM model is contructed, with fixed effects at the level of the individual word (Level 1) and random effects at the level of the individual learner (Level 2). A mixed-effects analysis is conducted with a view to determining the potential between-learner variance in the effect of lexical complexity.

The third and last stage of this study is the development of an artificial neural network (ANN) model. The model is based on a deep learning architecture composed of word and character embeddings, which capture the same information as most standard features of lexical complexity (De Hertog & Tack, 2018). The model is trained to predict lexical difficulty with various deep learning layers, including convolutional and bidirectional long-short term memory (BiLSTM) layers. Finally, a standard cross-validation procedure is used to evaluate the predictive power of the model.

## 1.4 STRUCTURE OF THE THESIS

The dissertation comprises three main parts, which coincide with the doctoral study's methodological steps. An illustration of this structure and these three steps is given in Figure 1.6.

The first part of the thesis (Part i) deals with the review of the scientific literature. The first chapter of this part (Chapter 2) describes the first research study (Study 1), a systematic review of 140 publications on the prediction of lexical competence in L2 reading. Because this extensive review looks more generally at lexical competence as a superordinate of lexical difficulty, the next chapter (Chapter 3) zooms in on lexical difficulty in particular. The chapter summarizes three methods to identify difficult words: theoretical, empirical,

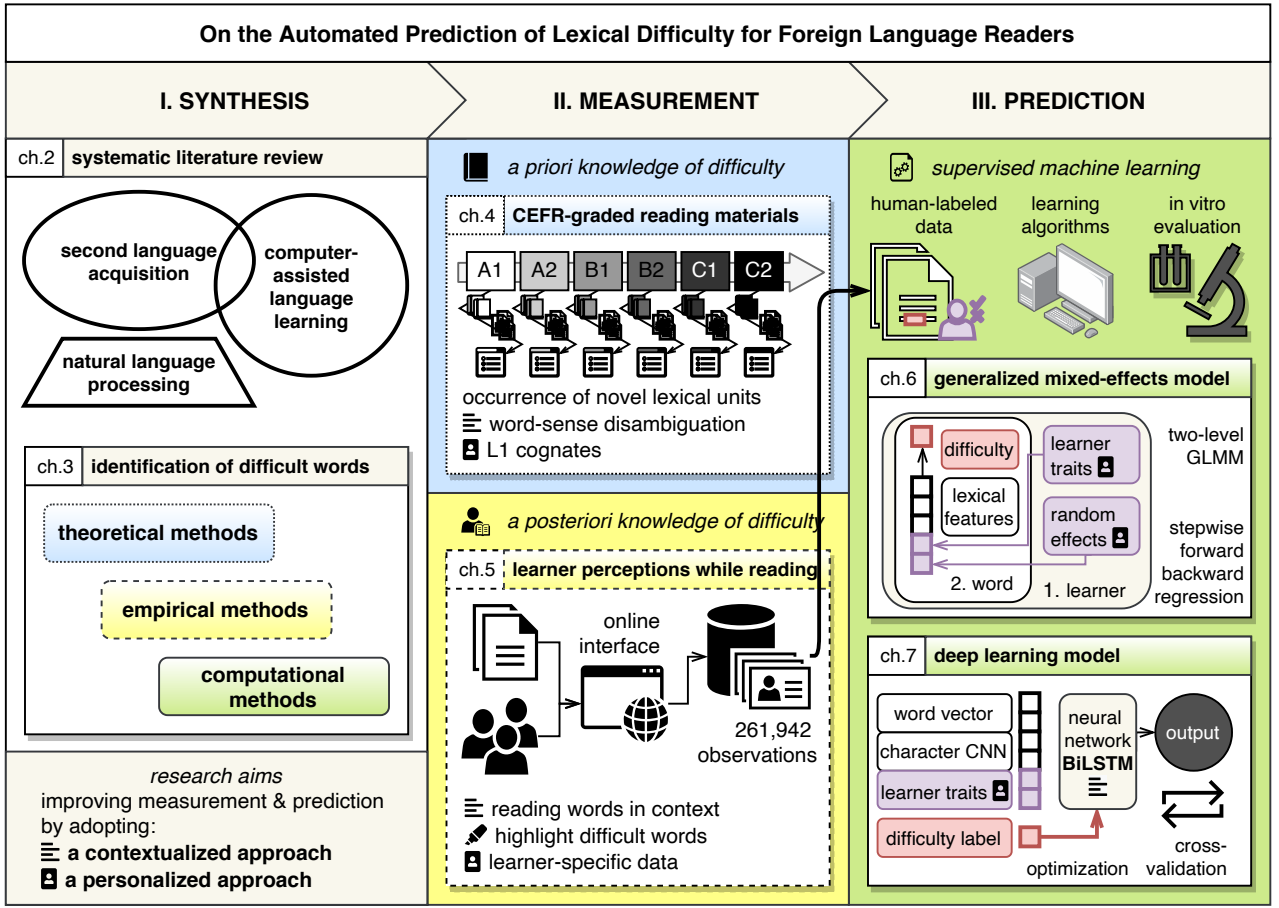



**Figure 1.6**
*Structure of the Thesis*

and computational. This distinction will be an important one to make as each of these three methods will be further looked into in the subsequent chapters.

The second part of the thesis (Part ii) deals with measuring lexical difficulty in L2 reading. The first chapter of this part (Chapter 4) explores whether it is possible to measure lexical difficulty from a theoretical (a priori) perspective by looking at the occurrence of new words in reading materials labeled with CEFR levels. As such, the chapter describes the second research study (Study 2) and the development of a new lexical resource for Dutch L2 (viz., NT2Lex). The next chapter (Chapter 5) presents an empirical (a posteriori) perspective to measuring lexical difficulty. The chapter describes two data collection trials during which learners of French L2 identify difficult words while reading.

The final part of the thesis (Part iii) deals with predicting lexical difficulty in L2 reading. This part covers the second and last phases of the final research study (Study 3) and describes two computational models trained on the data described in Chapter 5. These models explain why and predict when learners are triggered to perceive difficult words while reading. The first chapter (Chapter 6) presents a generalized linear mixed-effects model that predicts difficulty at two levels: the level of the individual word (lexical complexity) and the level of the individual learner (random effects). The second chapter (Chapter 7) presents a comparison of several neural network models, which make (non-)contextualized and (non-)personalized predictions of difficulty. The analyses compare the models' predictive power for the group of learners collectively and individually.

The last chapter of this dissertation (Chapter 8) presents the main conclusions and implications regarding the added value of contextualizing and personalizing the predictions of lexical difficulty. The chapter also suggests some limitations and perspectives for future research.

PART I

# STATUS QUAESTIONIS

The greatest difficulties lie where we do not look for them.

— Johann Wolfgang von Goethe, *Maximen und Reflexionen* (1833)

CHAPTER 2

# THE PREDICTION OF LEXICAL COMPETENCE IN FOREIGN LANGUAGE READING

## *A Systematic Scoping Review*

**Abstract** This chapter presents a systematic literature review of 140 publications on the prediction of lexical competence in L2 reading. The literature study addresses two research questions. What is the scope of studies that have statistically assessed vocabulary as a criterion/dependent variable in reading? How has the construct of lexical competence been measured and predicted? The chapter presents several citation and descriptive analyses highlighting main trends as well as some key limitations.

When browsing through the literature dedicated to L2 vocabulary and reading, it is clear that the body of research is vast and long-standing. One of the earliest traces of modern experimental research on this subject can be found in Grinstead's (1915) *An experiment in the learning of foreign words*. The study compared learning the meaning of unknown words either in isolation (i.e., in a list) or in context (i.e., in a text) and showed an immediate superior effect for contextual word learning. However, the use of a single-subject design mitigated the study's impact. Throughout the subsequent century, further evidence on what has since been defined as incidental vocabulary acquisition (Saragi et al., 1978), either from exposure to written context (Nagy et al., 1985) or through extensive reading (Coady, 1996), provided more impactful additions to these precursory insights.

Besides the effect reading has on vocabulary acquisition, another research focus is the role vocabulary plays as a reading variable. Reading in a foreign language is a process of simultaneously interacting variables in which sufficient language competence and vocabulary mastery are essential for success (Swaffar, 1988). This research perspective views vocabulary as a 'causal' or predictor variable of reading comprehension. In this interactive view of reading, researchers have mainly conceptualized vocabulary mastery in terms of *lexical coverage*, that is, the number of words in the reading text that are known to the reader; and *vocabulary size*, that is, the total number of words in the reader's mental lexicon (Tsai, 2017). In contrast, an opposite research perspective sees lexical competence as a criterion variable in reading. If we can define *language competence* as the ability to process and understand a language (from a receptive point of view), we may similarly define *lexical competence* as processing and understanding this language's lexical system. In this view, researchers may be interested in uncovering the factors that predict, among others, the abilities to decode the word's form, recall its meanings, and infer its precise meaning and function from the surrounding context.

This brings us to a fundamental question: If lexical competence is necessary for successful L2 reading, how can we predict this competence as a criterion variable? The first step in answering this question is to review the relevant scope of scientific literature. However, given that most previous literature reviews focused either on the effect of reading on vocabulary acquisition or the role vocabulary plays as a predictor variable (cf. *infra*), there is need for a survey that systematically delineates this research scope. Therefore, this literature study aims to provide a systematic scoping review of empirical studies in L2 reading that examined lexical competence as a criterion variable.

The chapter is structured as follows. Section 2.1 provides a summary of previous literature reviews on the interplay between vocabulary and reading. The summary addresses two research perspectives: the effect of reading on vocabulary and the effect of vocabulary on reading. The section ends with an introduction to the study's two research questions. Next, Section 2.2 describes the method adopted to conduct the systematic scoping review. The method comprises three steps: identifying potentially relevant research, selecting publications based on precise inclusion/exclusion criteria, and extracting data

from these publications. Sections 2.3 and 2.4 present the results of the data analysis regarding the two research questions. Finally, Section 2.5 discusses the main conclusions and implications.

## 2.1 PREVIOUS LITERATURE REVIEWS

### 2.1.1 *The Effect of Reading on Vocabulary Development*

In his seminal synthesis of the literature, Krashen (1989) made an influential claim for the effectiveness of incidental vocabulary acquisition that has shaped research over the last thirty years. According to Krashen's Input Hypothesis, a non-native acquires vocabulary best through meaning-focused activities providing comprehensible input (i.e., reading) than through rote learning and corrective feedback. The appeal of such a hypothesis may stem from the fact that incidental acquisition has shown some degree of universality; its evidence has been underscored in studies on both native (Herman, 1987) and non-native (Day et al., 1991) development, as well as in the case of spoken input (Ellis, 1999).

Nevertheless, incidental learning does not have a monopoly on effective lexical development. In her review on L2 vocabulary learning through reading, Raptis (1997) viewed incidental learning as the "default hypothesis" against which researchers should compare other approaches. Consequently, the assumed efficacy of unconscious and incidental learning over conscious and intentional vocabulary learning has been long-disputed (Hulstijn, 2001; Hunt & Beglar, 2005). For instance, Laufer (2003) found that learners achieved higher vocabulary gains by engaging in word-focused activities than through reading. Moreover, since incidental learning is a slow process and often requires extensive reading[8], some studies examined "whether instructional intervention could support the process and make it more directed and efficient" (Paribakht & Wesche, 1996, p. 174). Advocates of such a reading+ approach found that reading supplemented with exercises or explicit instruction lead to more effi-

8 Saragi et al. (1978), for instance, studied the incidental learning of unknown *nadsat* words by native English speakers reading Anthony Burgess' *A Clockwork Orange*. The results showed that participants sometimes failed to learn the correct meaning from context, even after hundreds of exposures to the word.

cient learning (Lin et al., 2018; Peters, 2012a; Zimmerman, 1997). For instance, Sonbul and Schmitt (2010) observed that directed vocabulary instruction after reading increased meaning recognition and lead to a better recall of meaning and form.

The question then remains as to what factors contribute to the acquisition of vocabulary through reading. In their narrative review, Huckin and Coady (1999) surveyed different aspects of incidental vocabulary learning: "There is reason to believe (...) that extensive reading for meaning does not lead automatically to the acquisition of vocabulary. Much depends on the context surrounding each word, the nature of the learner's attention, the task demands, and other factors" (p. 183). In particular, reading effectiveness depends on the amount of involvement the task induces in the learner (see Laufer and Hulstijn, 2001). This construct of task-induced involvement encompasses:

- motivational factors, characterizing the need to engage in new word learning; and
- cognitive factors, characterizing the allocation of attention to searching for possible form-meaning mappings and evaluating their adequacy from the surrounding context.

However, few meta-analytic reviews provided evidence for the overall effectiveness of these different variables. Huang et al. (2012) showed that incidental vocabulary learning was bettered with output tasks than through reading alone, thus supporting the involvement load hypothesis. Similarly, Uchihara et al. (2019) examined the effect of the repeated number of encounters with a word on subsequent lexical gains. The results showed a significant positive correlation between word repetition and word learning, but the overall effect size was moderate. Consequently, the study suggested that other learner and task factors affect lexical growth through reading.

Other meta-analyses focused on the effect of input enhancement on incidental vocabulary acquisition. Input enhancement provides marginal glosses, hypertext information, and other multimedia content to help the reader understand a text better. Because this reading support is mainly computer-based, most meta-analytic research on this topic stem from the field of computer-assisted language learning. Abraham (2008) showed that computer-mediated

glosses had a large effect on the immediate and delayed vocabulary retention after reading. Vahedi et al. (2016) and Yun (2011) showed that providing dual-modality glosses (i.e., text & visuals) increased vocabulary learning significantly, although the overall effect size was weak due to small sample sizes. These meta-analyses suggested that reading with input enhancement had a significant positive effect on incidental vocabulary acquisition, but this effect depended on other factors such as instruction, proficiency, and sample size.

### 2.1.2 *The Effect of Vocabulary on Reading Comprehension*

The previously cited reviews and meta-analyses showed that reading, possibly enhanced with learning-conducive tasks, positively affects vocabulary development. Contrariwise, vocabulary knowledge is also a significant determinant of successful reading comprehension (Swaffar, 1988; Tsai, 2017). According to Jeon and Yamashita's (2014) meta-analysis, grammar and vocabulary predominantly regulate reading comprehension rather than pure (meta)cognition. Thus, achieving skilled reading in the L2 ultimately "poses a language problem rather than a reading problem" (p. 196). Similarly, Melby-Lervag and Lervag's (2014, March) meta-analysis suggested that non-natives must develop strong language skills to attain reading skills on par with native speakers. Furthermore, other meta-analyses in computer-assisted language learning demonstrated that enhancing the comprehension of unknown words positively affects reading comprehension as well. Taylor (2006) observed that glossed reading achieved a medium effect size compared to unglossed reading, with a large effect on computer-assisted glosses compared to traditional glosses. Similarly, Abraham (2008) observed large effect sizes for computer-mediated glosses in intermediate learners and narrative texts.

### 2.1.3 *Research Questions*

Some critical observations emerge from the previous overview. Firstly, although past reviews provided essential insights into the interplay between vocabulary and reading, they did not explicitly focus on predicting lexical competence. Some meta-analyses did focus on vocabulary as a criterion variable

in reading, but these were limited to a single predictor and competence: the effect of either word repetition or input enhancement (glosses) on incidental vocabulary learning. Based on these reviews, we cannot determine the scope of all possible empirical studies having predicted any form of lexical competence in reading. Therefore, a more comprehensive review is needed.

Secondly, most previous reviews were either narrative syntheses or meta-analyses. In contrast, few scoping reviews have been conducted in the field to date (see Gurzynski-Weiss and Plonsky, 2017 and more generally Pham et al., 2014). Scoping reviews are similar to meta-analyses in that they are both systematic reviews. However, their research objectives are fundamentally different. Scoping reviews do not aim to aggregate evidence for a specific hypothesis. Instead, they "are commonly used for 'reconnaissance' – to clarify working definitions and conceptual boundaries of a topic or field (...) [and] are therefore of particular use when a body of literature has not yet been comprehensively reviewed" (Peters et al., 2015, p. 141). As such, the scoping review seems to be the type of systematic review that is particularly suitable for this study.

The present study aims to provide a systematic scoping review of empirical studies that predict vocabulary as a core aspect of L2 reading. The review aims to answer the following research questions (RQs):

RQ2.1 How is the scope of research structured? What types of studies have been conducted to date?

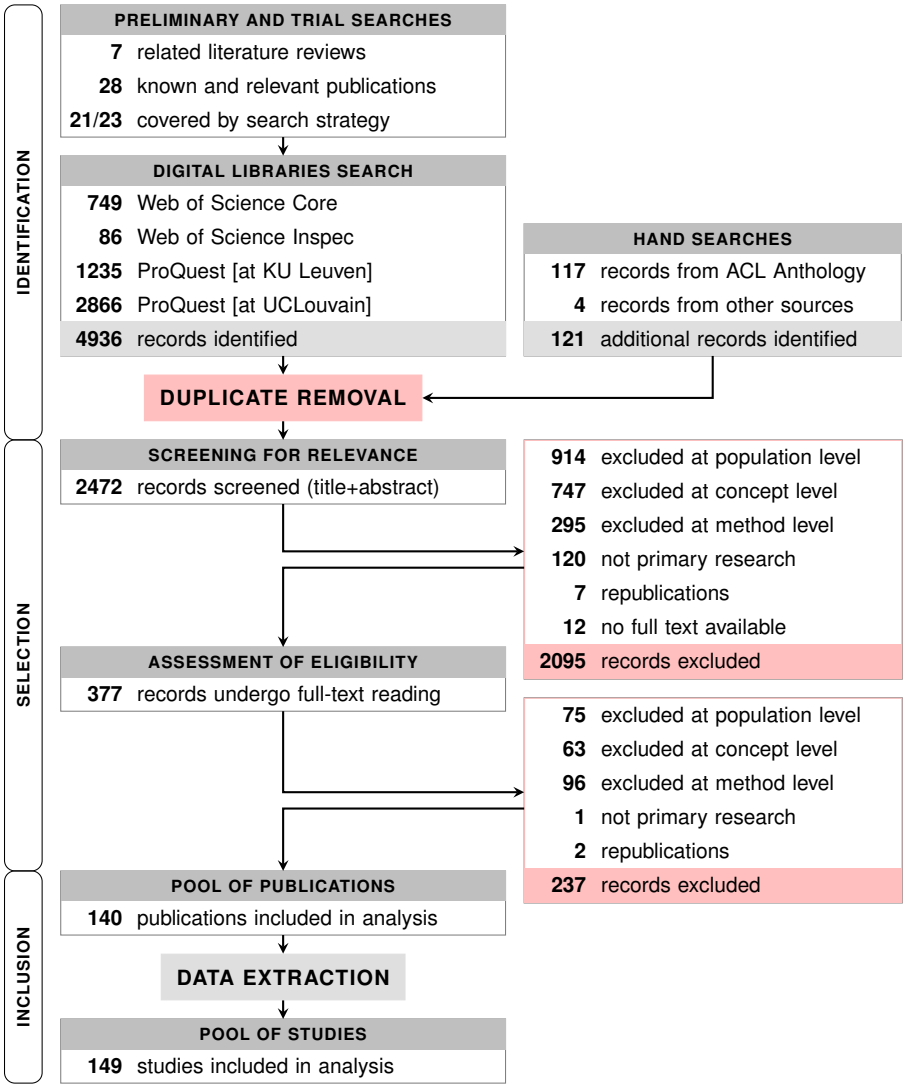
RQ2.2 What measurements and predictors have been used to statistically model lexical competence in L2 reading?

## 2.2 METHOD

Figure 2.1 summarizes the methodology of the systematic scoping review. Although the precise methodology may differ between research areas, the general methodology for conducting a systematic literature review always involves the following three steps: (a) the identification of potentially relevant literature in the form of a search strategy, (b) the selection of primary studies based on inclusion/exclusion criteria, and (c) the extraction of numerical and

**Figure 2.1**

*PRISMA Flow Chart of Literature Identification, Selection, and Inclusion*




categorical data from the included studies. The following sections further describe these three steps in more detail. More detailed documentation can also be found in the Appendix to this chapter (Section 2.A).

**Table 2.1**
*Definition of the Criteria for the Scoping Review*

| Criteria | Description | Definition |
|---|---|---|
| **Population** | foreign language learners | learners of a non-native language in an instructional and non-naturalistic setting |
| **Concept** | lexical competence in reading | processing, decoding, comprehension, and acquisition of vocabulary elicited while reading words in context |
| **Method** | prediction | empirical studies which statistically assess the defined concept as a dependent/criterion variable |

### 2.2.1 *Identification*

*Preliminary and Trial Searches*

The search for literature started with topically related syntheses and 28 potentially relevant publications. From these bibliographic entries, all keywords were extracted and structured according to three search criteria: Population, Concept, and Method (Table 2.1). Supplementary keywords were also semi-automatically extracted through a collocation retrieval on all titles and abstracts. All bigram and trigram collocations were retrieved with Dunning's likelihood ratio test and a significance level of $\alpha = .001$ (Dunning 1993, as cited in Manning & Schütze, 1999, pp. 161–4).

The initial search strategy was incrementally revised to keep the most relevant keywords and increase recall on retrieving the set of available publications. Given that five out of 28 records were absent from Web of Science or ProQuest, these trial searches only targeted the 23 records indexed in either database. In Web of Science, the search strategy covered 14 out of 17 publications. In ProQuest, the search strategy covered 20 out of 22 publications. Missing records mainly lacked abstract/keyword fields (Knight, 1994) or their abstracts/keywords lacked a clear link to written receptive vocabulary knowl-

**Table 2.2**

*List of Keywords Defined per Search Facet*

| Search set | Keywords |
|---|---|
| **Population** | foreign language(s), second language(s), FL, L2, L3, non-native(s), nonnative(s), acquisition, education, instruction, learning, learner(s) |
| **Concept** | (passage, reading, text) comprehension, incidental learning, reading, written receptive |
| | lexical (competence, inferencing, knowledge), lexicon, lexis, vocabulary, vocabulary (breadth, depth, development, difficulty, familiarity, growth, identification, knowledge, level, perception, recognition, recall, retention, representation, size), (passive, sight, targeted) vocabulary, word identification, (complex, key, target, unknown) word(s) |
| **Method** | annotation, complex word identification, controlled study, dictionary lookups, dictionary use, empirical, experiment, eye-tracking, glosses, models, modeling, outcome, prediction, pupil judgment, read-aloud protocols, self-assessment, think-aloud protocols, treatment, vocabulary tests, (feature, target, independent, predictor, dependent, criterion) variable |

edge (Webb, 2007; Yoshii, 2006). The final revised search strategy (see Table 2.2 and Section 2.A.1) covered 21 out of 23 known publications.

*Digital Library Searches*

Trial and database searches were conducted from September to December 2018. Web of Science and ProQuest were searched through a combination of three search sets targeting one of the facets described above. The search syntax was applied on all fields (i.e., title, abstract, keywords) except for the full text. Section 2.A.1 provides detailed documentation of the final search syntax used in ProQuest and Web of Science. Importantly, ProQuest was searched twice. First, the database was searched for relevant peer-reviewed records. Second, the collection of book chapters was searched without a peer-reviewed filter. Because these generally lack a peer-reviewed label, they would have been ignored during the first search.

*Searches in Peer-Reviewed Proceedings*

As noted above, five out of 28 known publications were absent from the digital libraries. These publications were from peer-reviewed conference proceedings in the field of computational linguistics. The ACL Anthology was searched to retrieve relevant publications in this field.

*Bibliography Management*

All matched records were imported into Zotero. After removing duplicate entries with Zotero's built-in tool, a collection of 2,472 publications remained.

### 2.2.2 *Selection*

Some publications were excluded for several practical reasons: they were not accessible in full-text, not written in English, without a peer review, mere republications or duplicate studies, or not primary research. Moreover, although the search strategy achieved a high recall on available publications, it was practically impossible to create a precise strategy that would, without reducing recall, filter out studies that matched the keywords only peripherally. The following criteria were defined in order to delineate a more precise and coherent scope of research.

*Inclusion/Exclusion Criteria*

The Population criterion focused on non-native subjects acquiring a human language (i.e., which naturally developed within a community over time) in an instructional and non-naturalistic setting. Consequently, studies were excluded with the following criteria:

- The sample was an indistinguishable mix of natives and non-natives.
- The sample included second-language learners, that is non-natives learning the societal language while residing in the host country (language minorities, migrants, or other) or learning the language of schooling in immersion, plurilingual heritage language, or target language medium education.

- The reading materials were a mix of native and foreign languages, such as code-switched or macaronic texts.
- The target language was a "foreign" artificial, formal, or programming language.
- The primary subjects were teachers, not learners.
- The study focused on language-related impairments or any other disability, disorder, or syndrome.

The Concept criterion focused on lexical competence in reading, which was not easy to define because different competencies and various reading modes exist. On the one hand, the criterion excluded studies when they focused on any competence falling beyond the lexical paradigm; the definition of lexical competence was, therefore, maximalist. On the other hand, the reading task was defined more restrictively since various tasks could trigger some form of reading. Studies were excluded with the following criteria:

- The primary subjects did not read themselves (e.g., storybook reading to preliterate children).
- There was no link between lexical competence and the reading activity (e.g., unrelated vocabulary exercises before, after, or in parallel with a reading activity).
- The reading process was completely decontextualized, with isolated prompts, stimuli, test items, or word lists. Contextualized self-paced word-by-word reading was not excluded.
- The core reading material was not in a single textual modality. Studies were not excluded when, additionally, multi-modal glosses or test measurements were used.
- Reading was a by-product of another task (e.g., data-driven learning) or combined with dialogic writing or speaking.

The Method criterion focused on studies empirically gauging the concept defined above, both during reading (i.e., online) and before/after reading (i.e.,

offline) (see Godfroid et al., 2010). Any type of analysis was relevant as long as the researchers statistically tested the concept as a criterion variable. Studies were excluded with the following criteria:

- The researchers conducted the study for action research.
- The researchers compared a relevant task to other non-relevant tasks (e.g., no reading, listening, reading+, and explicit vocabulary instruction). Such studies addressed reading as one possible factor of vocabulary development, which falls beyond the review's aims.
- The researchers used tests that did not target vocabulary attested in the reading materials (e.g., vocabulary size tests) or did not explicitly state whether the tested vocabulary was attested in the reading materials.
- The researchers extracted only average measurements (e.g., average word reading rate).
- The researchers only analyzed the measured lexical competence as an independent variable or a mere correlation of any other factor (e.g., reading comprehension).

*Screening for Relevance*

After a title and abstract review, 377 publications were screened as relevant for a full-text review. Zotero's advanced search was used to rule out records whose title, abstract, or tags matched keywords that were excluded per definition from the inclusion/exclusion criteria (see Section 2.A.2 in the Appendix for more details). Other non-relevant records were excluded during a more detailed abstract review.

*Assessment of Eligibility*

After a full-text review, 140 publications were eligible for inclusion in the synthesis. No studies were excluded because of unsatisfactory data quality or methodological validity. As pointed out by Peters et al. (2015), since "scoping reviews are designed to provide an overview of the existing evidence base regardless of quality (...), a formal assessment of [the] methodological quality

of the included studies is generally not performed" (p. 142). Although study quality assessment is a *sine qua non* for meta-analyses, it is unnecessary for a scoping review.

### 2.2.3 *Extraction*

During full-text reviewing, a preliminary version of the data coding scheme was piloted in Excel 2016 and was further enhanced based on the retained studies. Because this revised scheme (see Table 2.3 in the Appendix) required an object-relational mapping too tricky to handle in Excel, an SQL database (see Figure 2.16 in the Appendix) was used to collect all data during the last full-text reading. The data were analyzed to answer the study's two main research questions. In what follows, the terms "record" and "publication" will refer to the articles included in the synthesis. The term "study" will refer to the separate experiments reported in a publication.

## 2.3 DELINEATING THE RESEARCH SCOPE

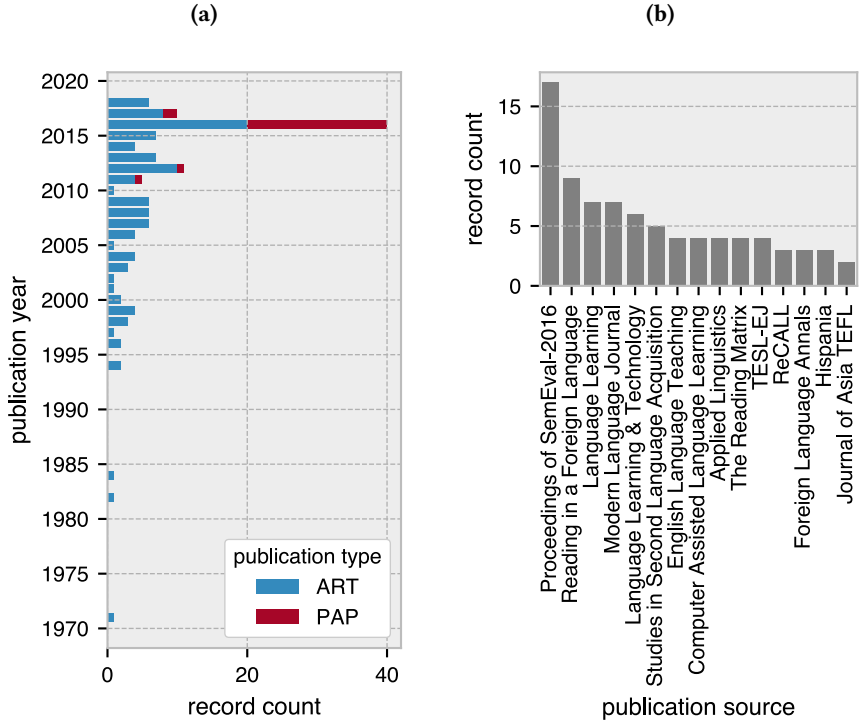
The first research question (RQ2.1) focused on delineating the research scope on predicting lexical competence in L2 reading. To delineate this scope, citation and descriptive analyses were conducted at two levels: (a) the network of scientific publications (Section 2.3.1) and (b) the studies reported per publication (Section 2.3.2). The following sections describe the results of these analyses.

### 2.3.1 *Publications*

Figure 2.2 shows the number of publications per year, type (i.e., journal article or conference paper), and top sources of publication. The figure shows that the collection of 140 publications spans nearly 50 years, with almost half (48%) of the publications issued in the last five years (2014–2018). The majority of publications (83%) appeared in major SLA and CALL journals such as *Reading in a Foreign Language*, *Language Learning*, *Modern Language Journal* (MLJ), *Language Learning & Technology* (LL&T), and *Studies in Second Language*

**Figure 2.2**

*Number of Records per Year, Type, and Top 15 Sources of Publication*



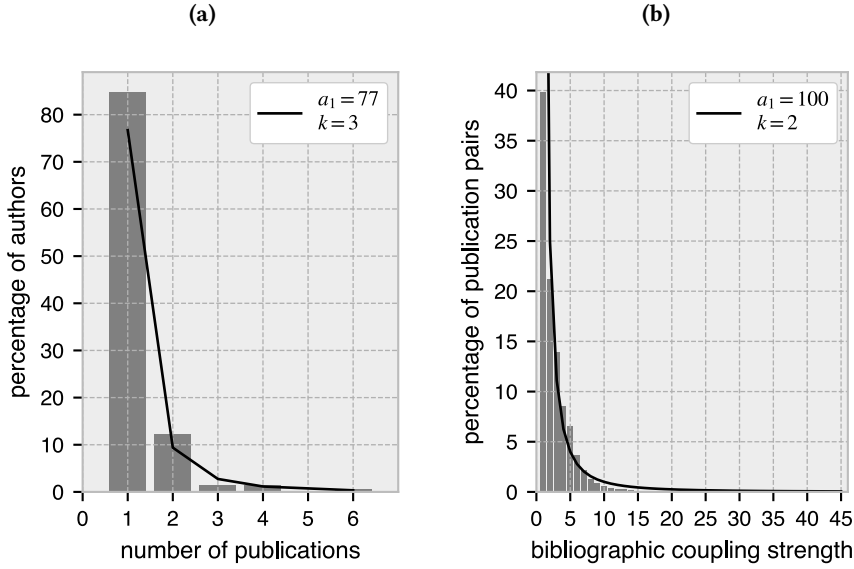

*Acquisition* (SSLA). Furthermore, the topic also recently sparked interest in the field of NLP, with numerous peer-reviewed conference papers published during the *International Workshop on Semantic Evaluation* (SemEval; Paetzold and Specia, 2016a). Several bibliometric analyses were performed to determine the collection's overall structure, comparing the publications in terms of author productivity and citation similarity.

*Author Productivity*

Figure 2.3a shows the percentage of authors (out of 230) contributing to one or more publications in the collection. The distribution shown in the figure suggests the application of Lotka's (1926) law (2.1). This power law states that author productivity $a_n$ (i.e., the number of authors $a$ writing $n$ publica-

**Figure 2.3**

*Power Laws of Author Productivity and Citation Similarity*




tions) is inversely proportional to $a_1$, or the number of authors making one contribution.

$$\forall n \in \mathbb{Z}^+ : a_n = a_1 n^{-k} = a_1 / n^k \tag{2.1}$$

The distribution parameters in Figure 2.3a, $a_1$ (i.e., the percentage of authors making one contribution) and $k$ (i.e., the power law's exponent), were determined by an optimal least-squares fit. The distribution verged on Pareto's rule: approximately 80% of authors made only one contribution, whereas 20% contributed to more than one publication. The most productive authors included Diana Pulido ($n$ = 6), Dorothy M. Chung ($n$ = 4), Hansol Lee, Jang Ho Lee, Jan L. Plass, and Marcos Zampieri ($n$ = 3). This small proportion of productive authors was the first indication of a dispersed research field.

*Bibliographic Similarity*

This dispersed research scope was also observed when publications were compared in terms of bibliographic coupling strength (i.e., the number of references shared by two publications; Definition 2.1). Only 38% of all pairs of publications in the collection had shared references. Moreover, Figure 2.3b[9] shows that most of these pairs (40%) cited only one reference in common ($BC = 1$). Unsurprisingly, the most similar publications ($BC = 45$) were mainly written by the same authors (e.g., Pulido, 2003, 2004a).

**Definition 2.1: bibliographic coupling strength**

Let $C_x$ be the set of references cited in publication $x$ and $C_y$ the set of references cited in publication $y$.

$$C_x = \{c : c \text{ is cited in } x\}$$
$$C_y = \{c : c \text{ is cited in } y\}$$

Then, bibliographic coupling similarity (Kessler, 1963)

$$BC(x, y) = |C_x \cap C_y| \tag{2.2}$$

measures the absolute number of references $c$ that are cited in common by $x$ and $y$.

A bibliographic coupling network of publications included in the synthesis was constructed with *metaknowledge* (McLevey & McIlroy-Young, 2017) and *NetworkX* (Hagberg et al., 2008). The nodes were clustered with label propagation. For each cluster, the most salient keywords were extracted from abstracts with TextRank (Mihalcea & Tarau, 2004; Nathan, 2016). This network is visualized in Figure 2.4. The figure shows that the clustering analysis identified two distinct subject areas. The largest cluster ($c_{02} = 117$) corresponded to the area of applied linguistics, with publications on vocabulary learning. The smallest

9 Like Figure 2.3a, the distribution parameters in Figure 2.3b, $a_1$ (i.e., the percentage of publication pairs citing one reference in common) and $k$ (i.e., the power law's exponent), were determined by an optimal least-squares fit.

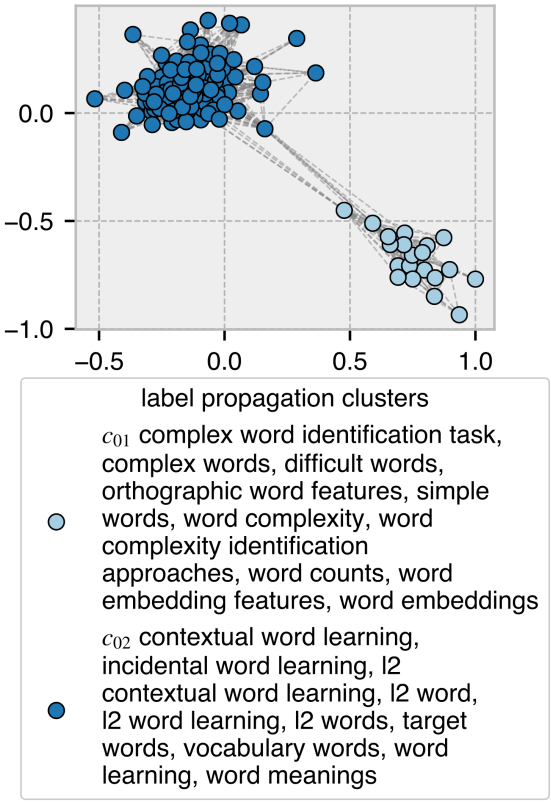



**Figure 2.4**

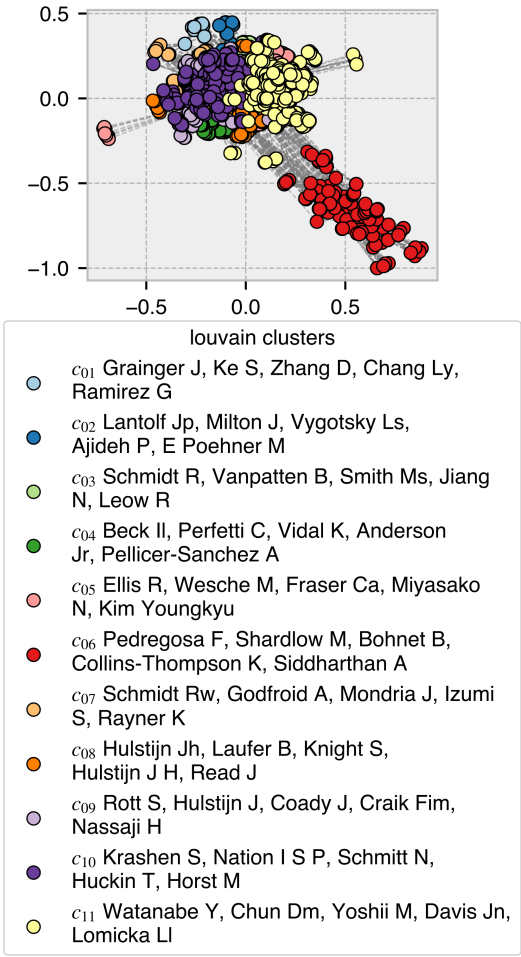
*Bibliographic Coupling Network of Publications Included in the Scoping Review*

*Note.* The network is a graph $G$ of publications linked by bibliographic coupling strength. The network includes 137 vertices $v$, with a minimal degree $\delta(G)$=4 and a maximal degree $\Delta(G)$=97. Each cluster was identified via label propagation and annotated with the top 15 most salient keywords extracted from abstracts with TextRank. Outliers with deg($v$)=0 ($n$=1) and deg($v$)=1 ($n$=2) were removed. The Fruchterman-Reingold algorithm was used for two-dimensional rendering.

cluster ($c_{01}$ = 20) corresponded to the area of computational linguistics, with papers published on the CWI shared task. Furthermore, the publications displayed a tendency to form local triangular clusters, as shown by the high value for the transitive clustering coefficient $C(G)$ = 0.69. Given a triple combination of records $x$, $y$, and $z$ in graph $G$, there was at least a two-thirds probability of all three sharing cited references, thus forming a bibliographic coupling triangle. These results combined suggested that the research scope comprised two distinct scientific areas that were weakly interconnected but intrinsically densely clustered.

**Figure 2.5**

*Co-Citation Network of Authors Cited in the Publications Included in the Scoping Review*




*Note.* The network is a graph $G$ of the first authors cited in the publications linked by co-citation strength. The network includes 1,850 vertices $v$, with a minimal degree $\delta(G)=4$ and a maximal degree $\Delta(G)=1216$. Each cluster was identified with Louvain community detection and annotated with the five most central reference authors regarding eigenvector centrality. Outliers with $\deg(v)=0$ or $\deg(v)=1$ were removed from the networks. The Fruchterman-Reingold algorithm was used for two-dimensional rendering.

### *Communities of Reference Authors*

Besides comparing publications in terms of bibliographic similarity, the first authors cited in the publications were also compared in co-citation strength[10] (Definition 2.2). A co-citation network was constructed with *metaknowledge* and *NetworkX*. In the network, 11 communities of reference authors were detected with the Louvain algorithm (Blondel et al., 2008). Per community, the

10 Typically, co-citation strength measures the similarity between two publications in terms of shared forward citations. Contrary to bibliographic similarity, a static and retrospective measure, co-citation strength is a prospective and evolving measure. The co-citation similarity of two publications can increase over time depending on whether they will be co-cited in future publications. Importantly, this "prospective" co-citation strength was not computed.

five most central authors were determined based on eigenvector centrality. This network is visualized in Figure 2.5. The most visually distinct cluster ($c_{06}$) is computational linguistics, with reference authors in machine learning, computational readability, and automatic simplification. The other communities covered partly overlapping domains in applied linguistics, including, among others, authors specialized in theories of attention and awareness ($c_{03}$), incidental vocabulary acquisition ($c_{10}$), and (electronic) glosses and dictionaries ($c_{11}$). This result suggested that the publications in applied linguistics addressed a variety of topics, whereas the publications in computational linguistics were focused on a single topic.

**Definition 2.2: co-citation strength**

Let $C_x$ be the set of publications that cite reference $x$ and $C_y$ the set of publications that cite reference $y$.

$$C_x = \{c : c \text{ cites } x\}$$
$$C_y = \{c : c \text{ cites } y\}$$

Then, co-citation similarity (Small, 1973)

$$CC(x, y) = |C_x \cap C_y| \tag{2.3}$$

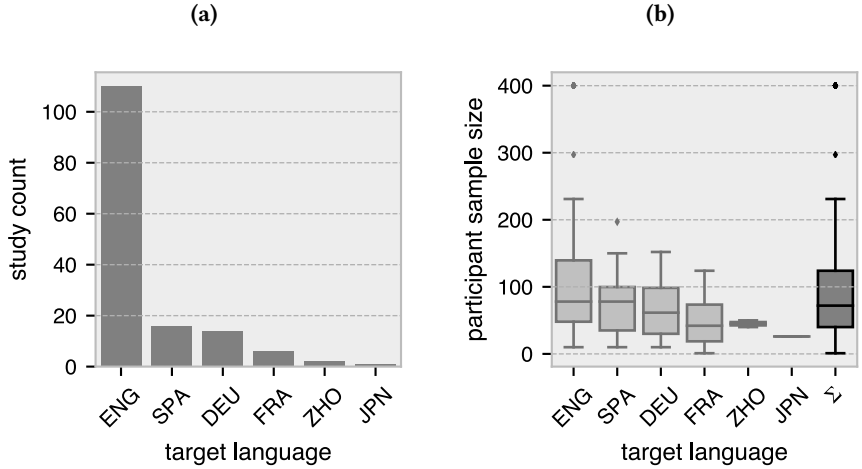
measures the number of times $x$ and $y$ are cited together in another publication $c$.

### 2.3.2 *Studies*

The collection of 140 publications reported at least one study, which referred to an experiment or trial with a different participant sample and reading task. In total, 149 studies were identified during data analysis. Each study was characterized by a target foreign language, a participant sample learning this target language, and a reading task comprising of a sample of texts and a set of target words. Two series of descriptive analyses were conducted to gain an overall picture of the studies' methodology.

**Figure 2.6**

*Distribution of Target Languages Across Studies*



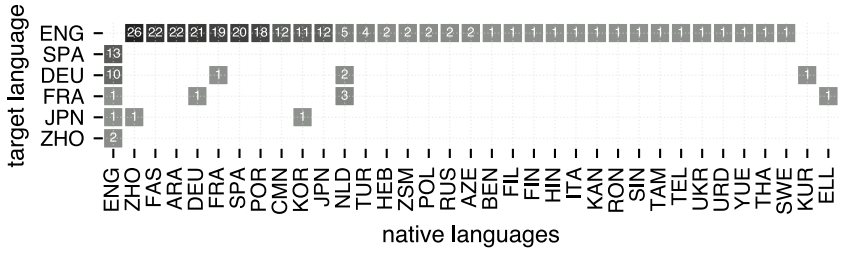

*Population*

The first series of descriptive analyses targeted the target learner population. Figure 2.6a shows that all studies examined the acquisition of an international language, such as English, Spanish, German, or French. It is clear from this figure that the vast majority, almost three-fourths (74%) of studies targeted English learners. However, this result was not surprising as it reflected findings of other scoping reviews conducted in L2 research (Gurzynski-Weiss & Plonsky, 2017). Regarding the participants' native language (L1)[11], English natives were tested on all other target languages, whereas speakers of all other native languages were tested mainly in English. This sparsity in the combination of L1s and L2s can be seen more clearly in Figure 2.7. Furthermore, most studies relied on convenience samples, which was manifested in the fact that most participants were (under)graduate students or university staff with a low to high-intermediate proficiency (Figure 2.8). Additionally, the number of participants tested across studies was highly variant (Figure 2.6b), although a

11 The results should be taken with caution since not all studies reported the participants' native languages.

**Figure 2.7**

*Heatmap of Native Languages in Learners of a Target Language*




Pearson product-moment correlation showed a small but significant increase in sample size over the years, $r(147) = .19$, $p = .02$.

*Vocabulary*

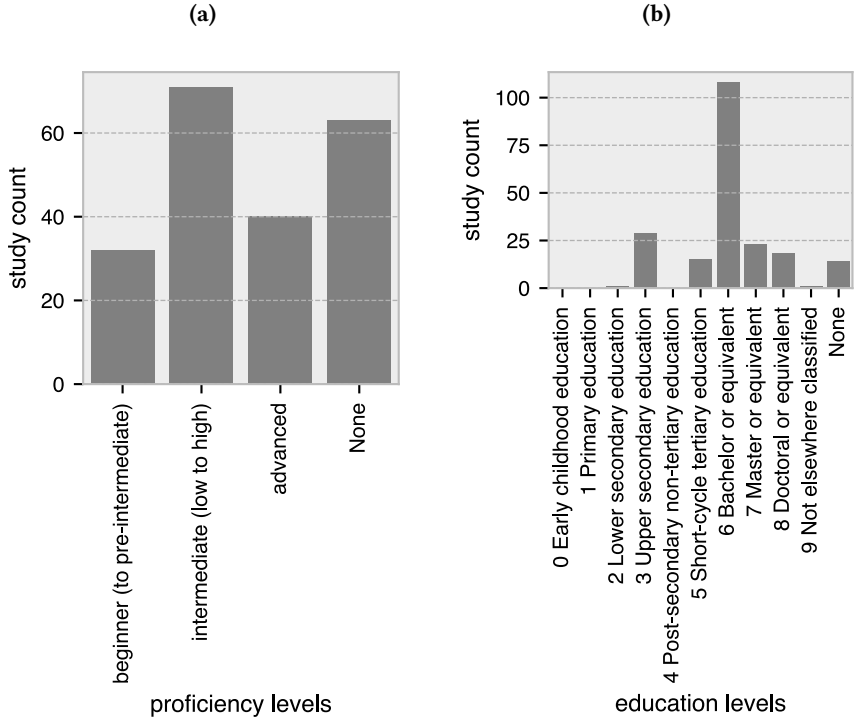
The second series of descriptive analyses focused on the vocabulary that was targeted in the studies. Most studies focused on incidental learning and therefore targeted unfamiliar words (Figure 2.9a). Unfamiliarity was commonly assessed through pilot ratings and pretests, while other studies used non-existent[12] words, either created explicitly for the experiment (e.g., *nonsense words*, *pseudowords*, *substitute words*, *invented forms*, and *disguised forms*) or invented by an author in a particular work of fiction[13] (i.e., *nonce words*). Some studies did not give a precise number of words tested because they targeted either all words or those that were clicked on. Other studies focused on a set of carefully chosen words, although the number of words tested was not large on average ($M = 34.9$, $Mdn = 22$) (Figure 2.9b). Furthermore, the reading materials in which the target vocabulary was attested also varied greatly across studies. In half (51%) of the tasks, learners were presented with a narrative or expository text or reader in its entirety, while other tasks resorted

12 Pilot or pretest ratings cannot guarantee that the targeted words will be entirely unfamiliar. This can be controlled by using non-existent words. However, this experimental control comes at the expense of ecological validity. Webb (2007) made a critical observation: "While it may be the goal of studies using nonsense words to investigate authentic language learning, it may be more accurate to state that they are simulating language learning" (p. 50).

13 Well-known examples are the *nadsat words* in Anthony Burgess' *A Clockwork Orange* (Saragi et al., 1978) and the *giant words* in Roald Dahl's *The BFG* (Reynolds, 2016).

**Figure 2.8**

*Distributions of Proficiency and Education Levels Across Studies*



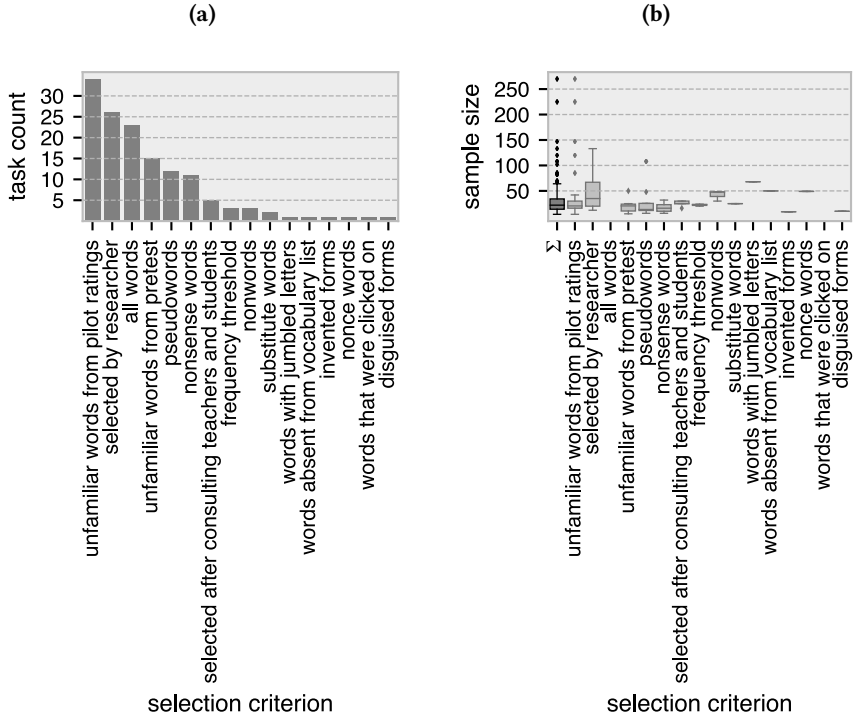

to text excerpts (passages, paragraphs, sentences, or contexts). However, no notable difference was observed in the number of target words, although one might argue that a more varied lexical field could be examined if the materials consisted of shorter but more texts.

## 2.4 LEXICAL COMPETENCE AS A CRITERION VARIABLE

The second research question (RQ2.2) focused on the target vocabulary and addressed how the learners' competence of these target words was measured and predicted. Each study included at least one lexical competence measurement and at least one predictive analysis (see the data coding schema in Section 2.A.4). A *measurement* was the outcome measured by a single instrument. An *analysis* was a statistical test or model that integrated one

**Figure 2.9**

*Criteria Used to Select Target Vocabulary*



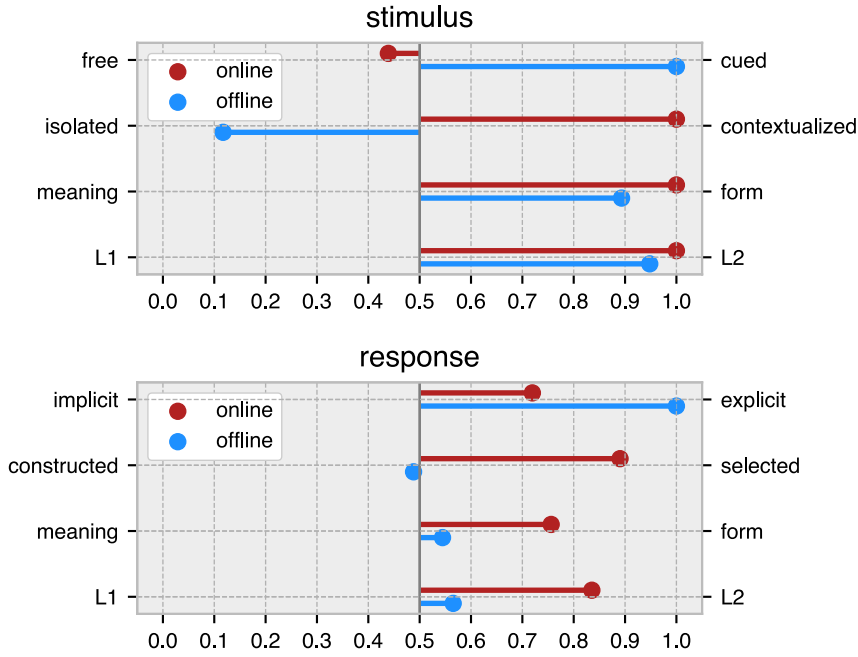

or more independent variables (predictors) and which targeted one or more measurements. Because of the numerous predictive analyses performed across studies ($N$ = 475, $\min_{study}$ = 1, $\max_{study}$ = 32, $M_{study}$ = 3.19), the following sections will outline only the most prevalent measurements and predictors.

### 2.4.1 *Measurements*

More than three hundred measurements were identified across studies ($N$ = 407, $M_{study}$ = 2.73, $M_{analysis}$ = 1.35). Most measurements (83%) were taken before or after the reading task, while few measurements (17%) were made during the reading task. These offline and online measurements used different types of stimuli, responses, and instruments.

**Figure 2.10**

*Proportions of Stimulus and Response Values per Type of Procedure*




*Note.* This lollipop chart shows the proportional tendency toward a specific stimulus and response attribute. The theoretical proportion is equal to 0.5 (gray vertical line). Each attribute was represented as a binary variable set to 1 (*True*) or 0 (*False*) during data extraction. Unspecified attributes were set to 0.5 (*None*).

*Stimuli & Responses*

To understand how stimuli and responses were coded during data analysis, consider Example 2.1 showing two test items for the pseudoword *verzettern*.

(2.1) a. verzettern ❑ yes ❑ no

b. Immigranten neigen dazu, sich **verzettert**, ausgebeutet und [...] zu fühlen.

verzettern: _____

*Example of two vocabulary test items from Peters (2007, p. 42)*

Each stimulus was coded with four binary variables:

1. The stimulus *verzettern* was coded as *cued* in both 2.1a and 2.1b because the researcher explicitly signaled the target word. A *free* stimulus would be a word in a text not explicitly marked by the researcher.

2. The stimulus *verzettern* was *isolated* in 2.1a and *contextualized* in 2.1b.

3. The stimulus *verzettern* was a *form* in both 2.1a and 2.1b. A *meaning* stimulus would be an item where a definition is given.

4. The stimulus *verzettern* was given in the L2 (target language) in both 2.1a and 2.1b. An L1 stimulus would be a test item where a definition is given in the native language.

Each response was coded with four binary variables:

1. The response was *explicit* in both 2.1a and 2.1b because the participant explicitly responded. An *implicit* response would be the learner's eye movements or brain signals while reading.

2. The response was *selected* in 2.1a because it required the learner to select whether the word was recognized. The response was coded as *constructed* in 2.1b because the learner had to construct a definition or translation.

3. The response was a *form* or a *meaning* (cf. stimulus).

4. The response was given in the L2 or L1 (cf. stimulus).

Figure 2.10 shows the proportional preference for specific stimuli and responses between offline and online measurements. Most offline procedures (88%) presented the target word in isolation rather than in the context in which it appeared. This preference for decontextualized over contextualized (12%) stimuli was significantly different from the theoretical proportion (50%) according to a one-proportion z-test, $z = -22.10$, $p < .001$. By contrast, all (100%) online procedures used contextualized measurements, which was significantly different from offline procedures according to a two-proportions z-test, $z = -24.80$, $p < .001$. This was unsurprising as it stemmed from the fact that decontextualized reading tasks were excluded from the synthesis. Since fewer online procedures were administered overall, these results suggested

that lexical competence was mainly operationalized in a decontextualized manner.

As for the types of responses, there was an almost equal proportion of offline measurements requiring either constructed (49%, $z = -0.49$, $p = .62$) or selected responses, providing either a form (54%, $z = 1.59$, $p = .11$) or a meaning, and given either in the L2 (57%, $z = 2.36$, $p = .018$) or L1. For online measurements, in contrast, there was a significant preference for selected (89%, $z = 11.30$, $p < .001$) form (76%, $z = 5.40$, $p < .001$) responses.

*Instruments*

Figures 2.11 to 2.13 on pages 56–60 provide a Sankey diagram illustrating the flow from a defined competence to the instrument with which it was measured. The three diagrams represent measurements requiring a selected response, a constructed response, and both a selected and constructed response. Each color in the diagram represents a different instrument.

Figure 2.11 on page 56 shows the list of instruments requiring a selected response. To measure the retention of form and meaning recognition after reading, the most often used instrument was the multiple-choice questionnaire (MCQ). Other offline instruments included matching tests (i.e., where a word form was linked to its definition), matching cloze tests (i.e., a fill-in-the-gap exercise with a list of possibilities), multiple-choice cloze tests (i.e., a fill-in-the-gap exercise with an MCQ), checklists (i.e., the learners check whether they remember the word after reading or were familiar with the word before reading), and lexical decision tasks. Although a small number of online procedures were also based on an MCQ to measure the learner's ability to infer the meaning of highlighted words, most online procedures used eye-tracking (i.e., tracking the eye movements on word forms), gloss-tracking (i.e., clicking on a word form to access its gloss or dictionary entry), or annotation and self-assessment tasks (i.e., marking a difficult or unknown word form).

Figure 2.12 on page 58 shows the list of instruments requiring a constructed response. Most offline procedures were based on vocabulary tests where the target word had to be translated or its definition supplied. As was already highlighted above, not many online procedures measured a constructed while reading. There were a few cases where either noticing or the ability to pro-

duce the target word's meaning was measured with think-aloud protocols or inferencing tasks.

Finally, Figure 2.13 on page 60 shows the list of instruments requiring both a selected and constructed response and measuring word reception and production at the same time. However, few researchers used such "composite" measurements. The Vocabulary Knowledge Scale (VKS) (Wesche & Paribakht, 1996), a scale with five self-report categories (Figure 2.14 on page 62), was the most frequently used instrument of this type. Researchers used this scale to measure either the learner's ability to infer word meaning while reading or vocabulary acquisition and retention after reading.

### 2.4.2 *Predictors*

About three hundred independent variables were examined across studies ($N = 303$, $M_{\text{study}} = 4.61$, $M_{\text{analysis}} = 3.21$). In order to bring structure in this multitude of predictors, a taxonomy was constructed based on previous insights into the psychological factors in SLA (Dörnyei, 2009; Ellis, 1999). A summary of the most salient nodes in this taxonomic tree is given in Figure 2.15. At the highest level, four factors were distinguished, pertaining to (a) the *input* conveyed to the subject, (b) the *subject*, (c) the *interaction* of the subject with the input, and (d) the *method* of data collection and analysis. Next, a co-occurrence matrix was constructed for each dichotomous variable in the response measurement (e.g., form/meaning). The number of analyses $f$ where factor $i$ predicted an attribute $j$ was normalized

$$w_{i,j} = \frac{f_{i,j}}{\sum_{k=1}^{m} f_{i,k}} \tag{2.4}$$

such that $w_{i,j} \in [0, 1]$ and $\sum_{j=1}^{m} w_{i,j} = 1$.

A noticeable observation that can be made from Figure 2.15 is that not all predictors were studied evenly and extensively across studies and that many were tested almost exclusively ($w \approx 1$) on one type of response. The vast majority (89%) of studies investigated a wide range of input-related predictors, focusing either on intrinsic vocabulary traits or on different ways in which the input could be enhanced. By contrast, a much smaller range (33%) of studies

**Figure 2.11**

*Measurements with a Selected Response (1/2)*

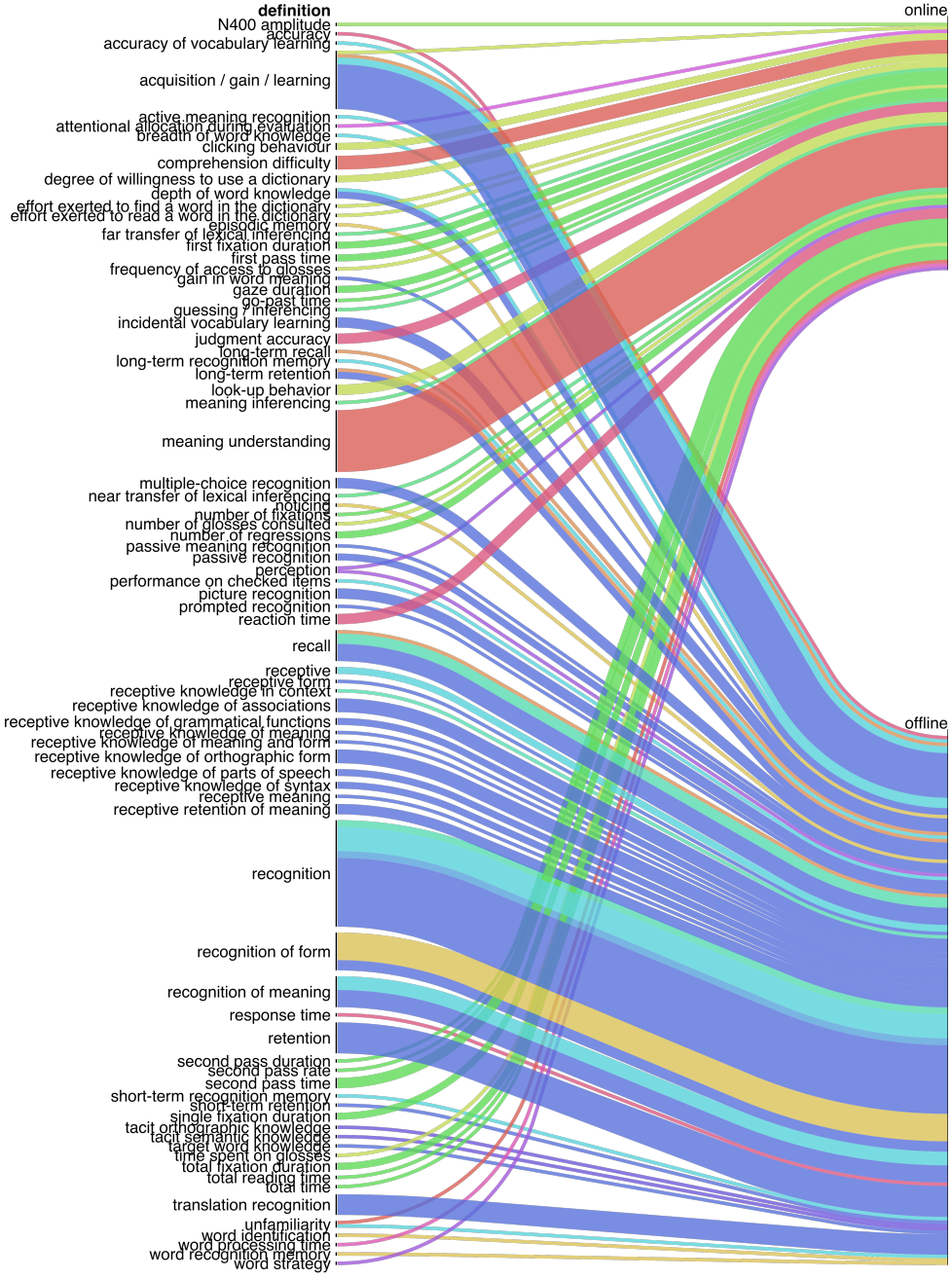


**Figure 2.11**

*Measurements with a Selected Response (2/2)*

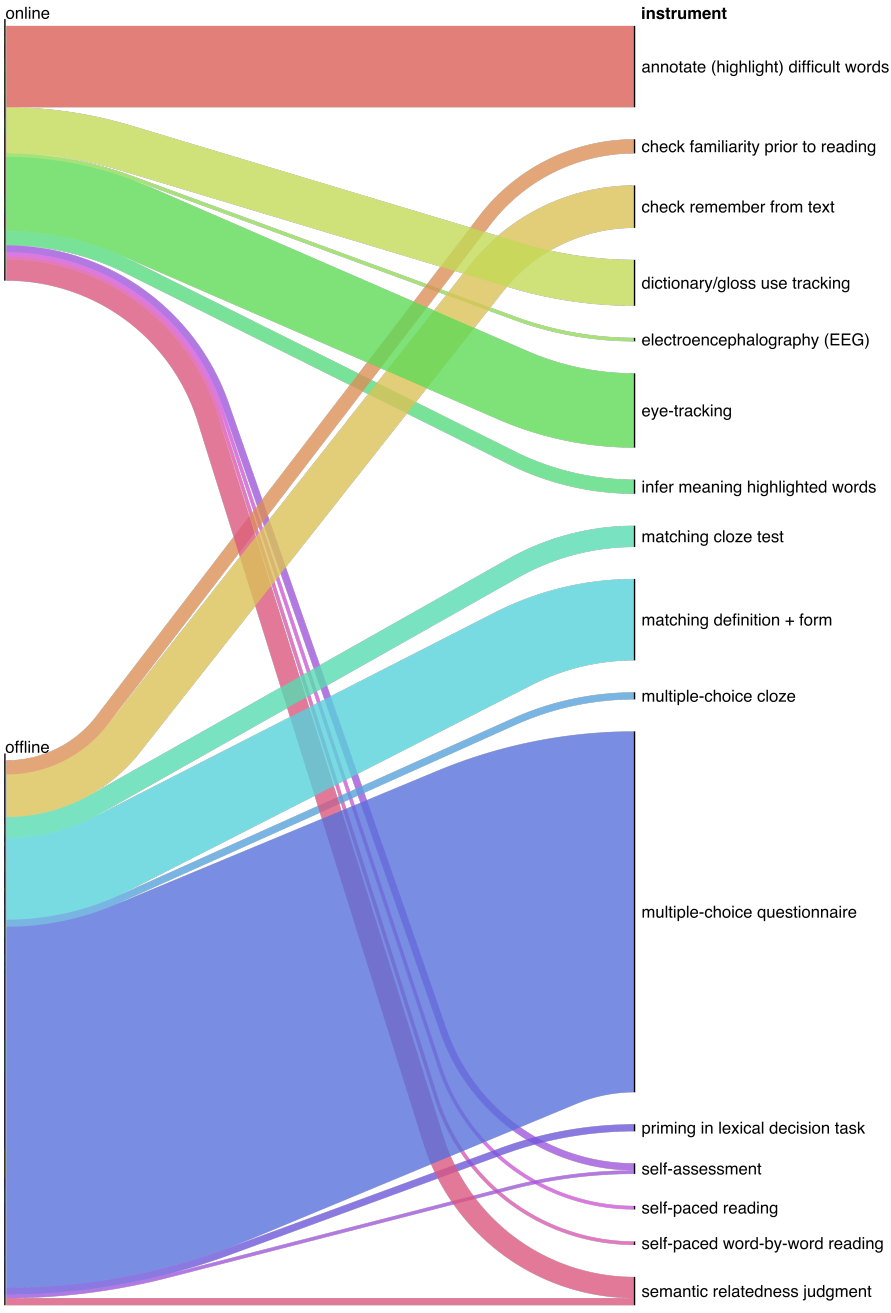


**Figure 2.12**

*Measurements with a Constructed Response (1/2)*

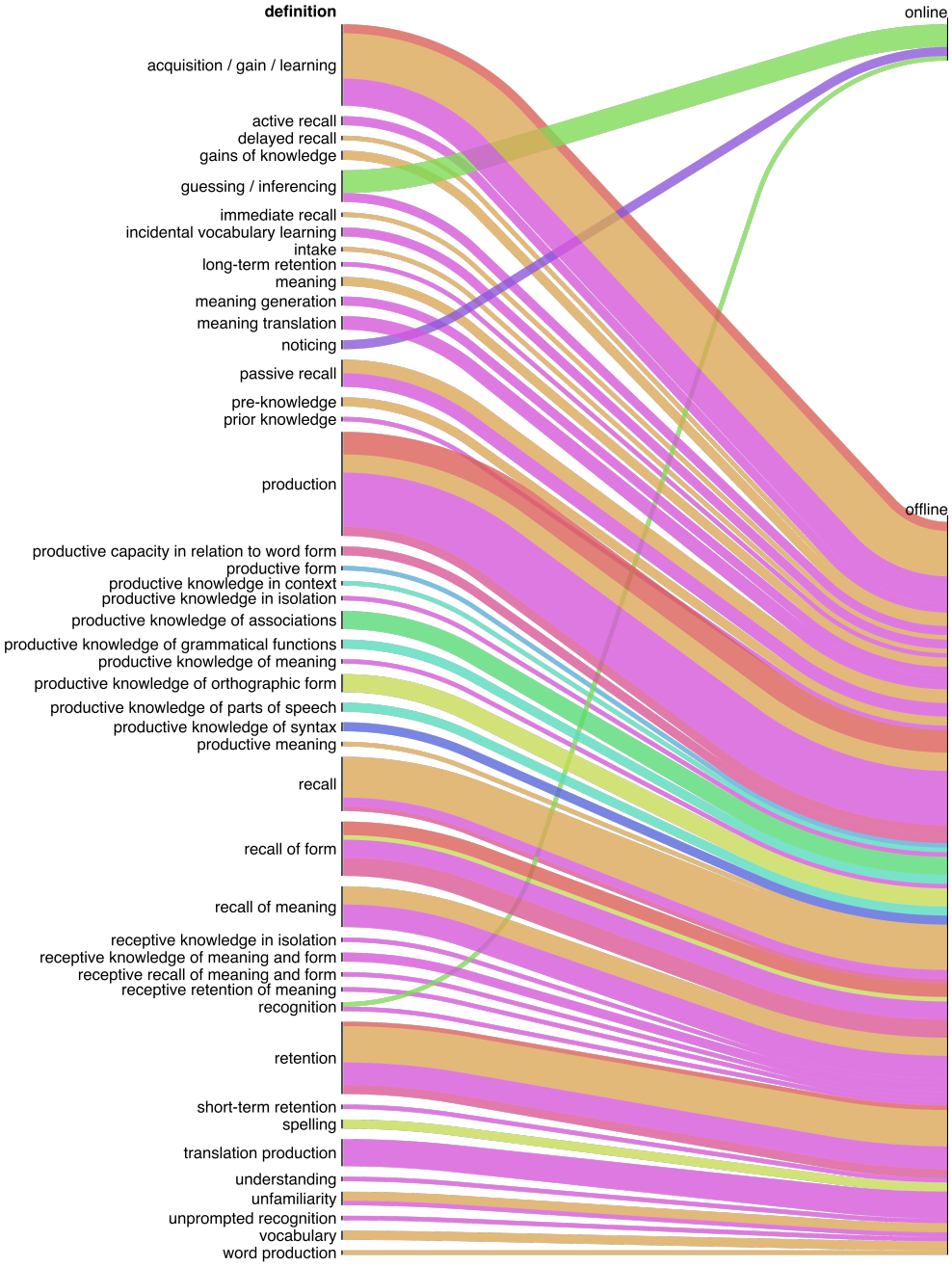


**Figure 2.12**

*Measurements with a Constructed Response (2/2)*

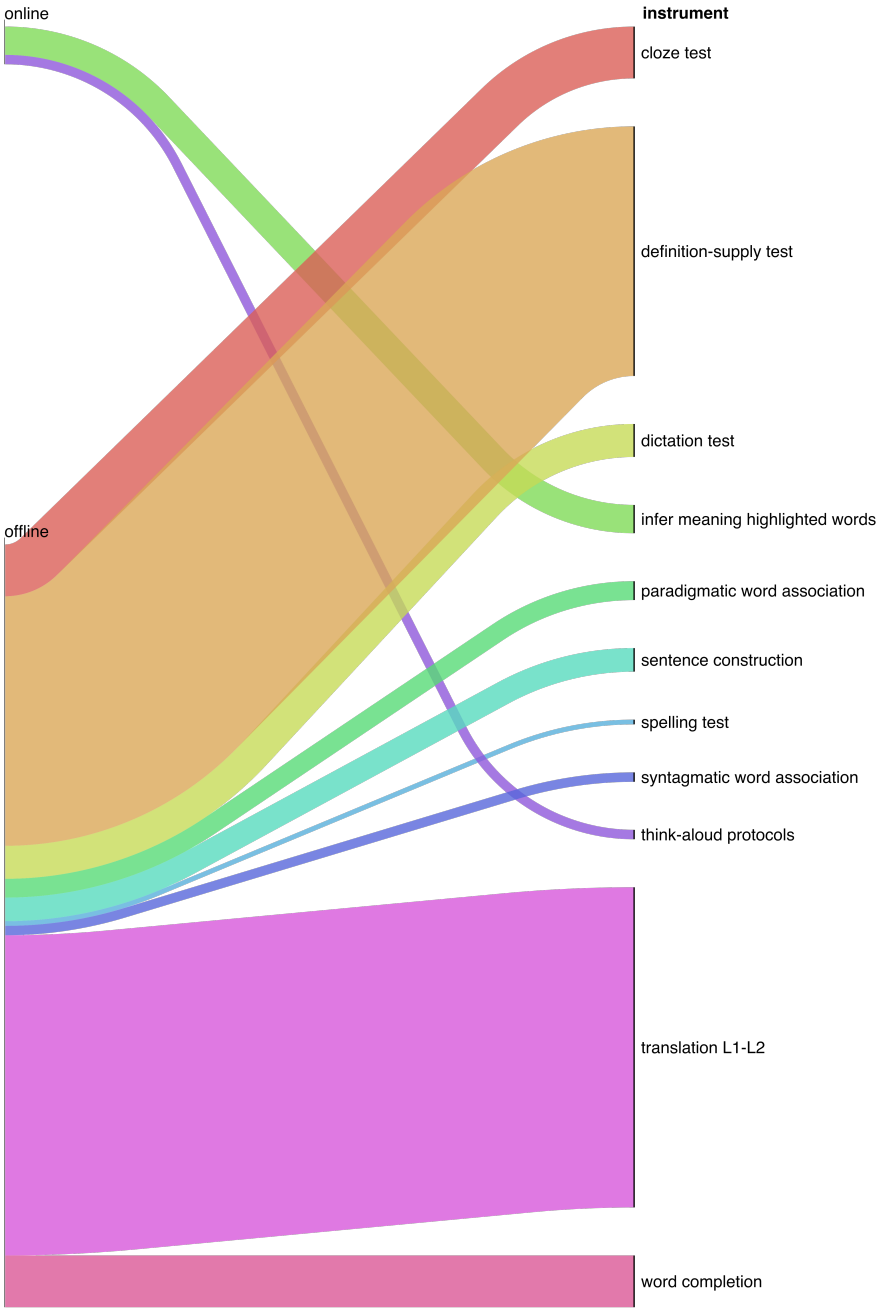
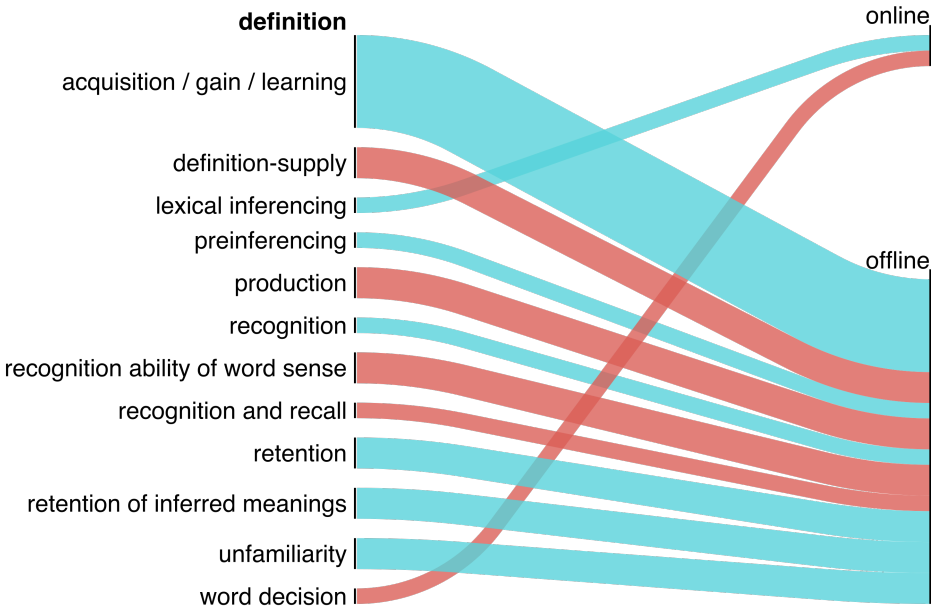


**Figure 2.13**

*Measurements with a Selected & Constructed Response (1/2)*

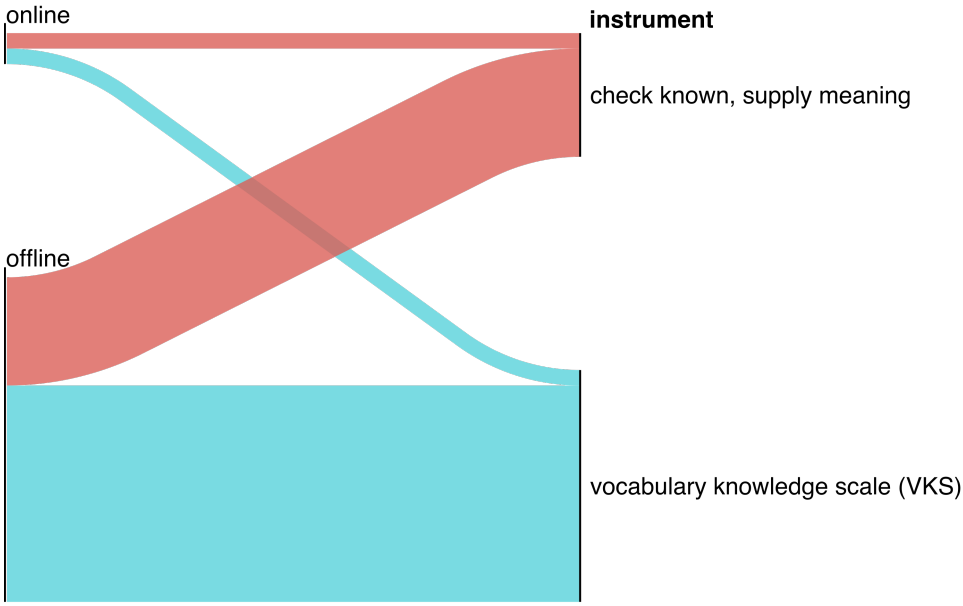



investigated the impact of different subject-specific characteristics and even fewer (16%) studies looked at how the subject interacted with the input.

Beyond the division already noted in RQ2.1, research in applied and computational linguistics also differed considerably in the way in which they predicted lexical competence while reading. On the one hand, studies in computational linguistics focused on building systems based on various intrinsic vocabulary traits to predict "complex words", which was measured from data where subjects were asked to highlight difficult words while reading. On the other hand, studies in applied linguistics made use of inferential statistics to explain the outcome of a more varied range of factors on a variety of measurements. The following sections will give an overview of the main results from the latter, bringing into focus the four levels of predictors outlined in Figure 2.15.

**Figure 2.13**

*Measurements with a Selected & Constructed Response (2/2)*




*Input*

Lexical competence during or after reading was first and foremost affected by the nature of the reading material. Many predictors were examined, ranging from the intrinsic traits of the word to overall textual properties. From these, the most evidenced were word occurrence, contextual elaboration, and input enhancement.

WORD OCCURRENCE & PRESENTATION All studies unequivocally reported that a higher frequency of exposure[14] to the word significantly increased the likelihood of retention (Chen & Truscott, 2010; Choi, 2016; Elgort et al., 2018; Elgort & Warren, 2014; Huang & Liou, 2007; Hulstijn et al., 1996;

14 Elgort et al. (2018) defined this as "the number of contextual encounters needed for the processing of an initially unfamiliar word to start approximating that of a known word" (p. 349). Although they used the term "order of occurrence", the term "exposure" was used here in order to avoid confusion with the sequential position or order of the word in the text.

**Figure 2.14**

*Self-Report Categories in the Vocabulary Knowledge Elicitation Scale*

| | |
|---|---|
| I. | I don't remember having seen this word before. |
| II. | I have seen this word before, but I don't know what it means. |
| III. | I have seen this word before, and I *think* it means _____. (synonym or translation) |
| IV. | I *know* this word. It means _____. (synonym or translation) |
| V. | I can use this word in a sentence: _____. (If you do this section, please also do Section IV.) |

*Note.* This scale is taken from Wesche and Paribakht (1996, p. 30).

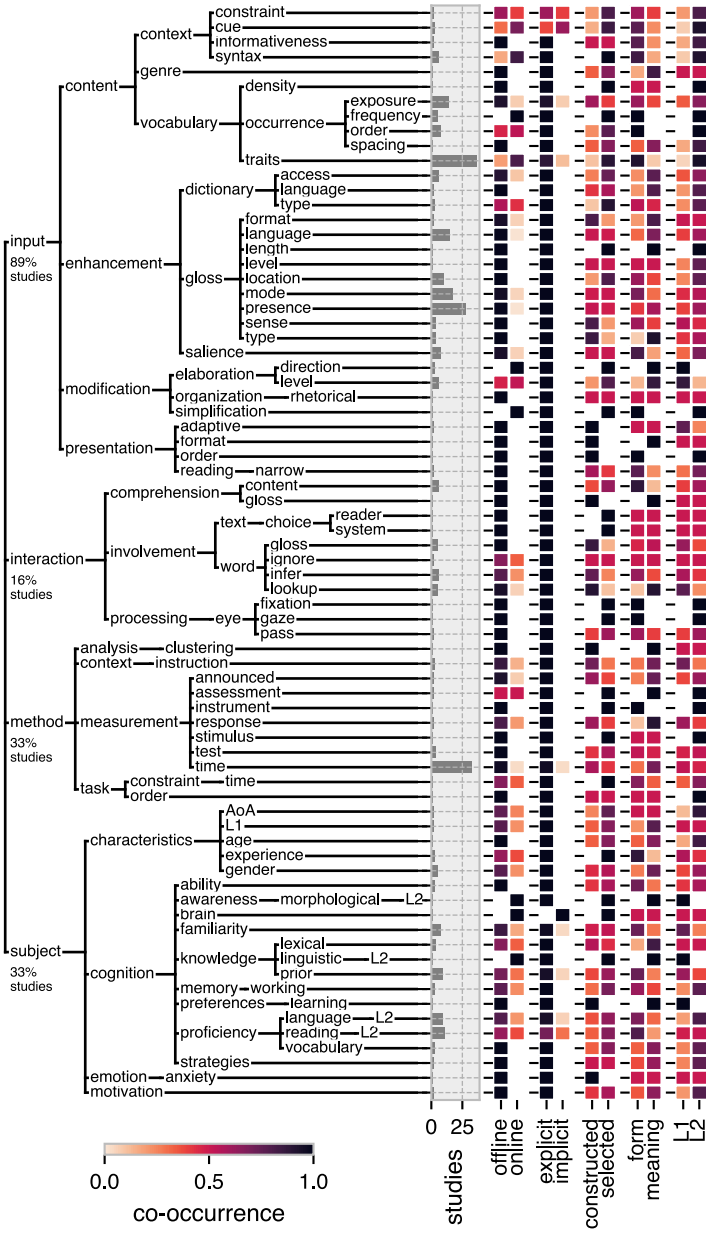
Pigada & Schmitt, 2006; Reynolds, 2016; Rott, 1999, 2007; Sun, 2014; Webb, 2007), although it was not clear how many exposures were needed seeing that thresholds varied between studies. A shorter spacing between these occurrences further lead to better retention, especially for low-proficiency learners (Elgort & Warren, 2014). Furthermore, the sequential order of the word in the text also affected comprehension and acquisition, but the directionality of this effect was not clear-cut. When words were encountered later on, there was a significant decrease in cognitive allocation, which was interpreted positively (i.e., indicative of faster comprehension; Kim et al., 2018) or negatively (i.e., reduced attention caused by a cognitively loaded task; Boers et al., 2017).

Frequency of exposure could also be further amplified by adjusting how the reading material was presented. For instance, a system that adapted the number of encounters with words that were unfamiliar to a specific learner gave rise to stronger recall and retention (Wang, 2016). Similarly, presenting learners with a series of thematically related texts (i.e., narrow reading) strengthened encounters with conceptually connected words and resulted in better receptive and productive knowledge gains that were less susceptible to memory attrition (Abdollahi & Farvardin, 2016; Kang, 2015).

CONTEXTUAL CUES & MODIFICATION Besides the exposure to a target word, the presence of informative and adequate contextual cues further determined success in inferring the word's meaning (Kaivanpanah & Rahimi, 2017;

**Figure 2.15**

*Taxonomic Tree of Predictors per Type of Response Measurement*



Liu & Leveridge, 2017; Sun, 2014; Webb, 2008). What is more, modifying the input by elaborating contextual and lexical cues resulted in better processing, inferencing, and gains (Birjandi et al., 2015; Hamada, 2015; O'Donnell, 2009; Watanabe, 1997) and was more effective when explicit rather than implicit cues were given (Kim, 2006). However, the direction did not seem to have an effect (Hamada, 2015) and the presence of semantic cues was of little importance for noticing and recognition of form (Godfroid et al., 2013). Yet, although lexical elaboration was effective, it was not significantly more effective than when the input was enhanced with glosses (Watanabe, 1997).

INPUT ENHANCEMENT A complementary factor to input modification was therefore the (technological) enhancement of the input, either by (a) adding typographic salience or (b) providing access to a gloss or dictionary. The effectiveness of highlighting forms to increase noticing and retention was inconclusive, ranging from no effect (Kim, 2006) to a superior recall of formulaic sequences (Peters, 2012b). Instead, attention-drawing techniques were more effective on attention and retention when they were used to highlight glosses (De Ridder, 2000, 2002). Indeed, it may be a good idea to encourage the learner to make use of a dictionary or gloss given that their effectiveness for establishing form-meaning connections was well-studied (AbuSeileek, 2011; AbuSeileek, 2008; Boers et al., 2017; Chen, 2016; Chen & Yen, 2013; Cheng & Good, 2009; Choi, 2016; Farvardin & Biria, 2012; Holley & King, 1971; Hu et al., 2014; Hulstijn et al., 1996; Jacobs et al., 1994; Knight, 1994; Ko, 2012; Plass et al., 2003; Rouhi & Mohebbi, 2012; Watanabe, 1997; Yoshii, 2006). However, the advantage of a specific location (i.e., in the text, in the margin, at the bottom of the page, in a pop-up window, or as a glossary at the end of the text), a particular language (i.e., L1 or L2), or a single or dual modality (i.e., combining pictures and/or definitions) was strongly regulated by other factors, such as word occurrence, ability and proficiency, and the type of competence measured. What is more, providing semantically ambiguous (MCQ) glosses (Duan, 2018; Nagata, 1999) resulted in ever higher gains, seeing that it stimulated the learner's involvement to search for meaning.

*Subject*

Although the nature of the reading material had an impact on lexical competence during or after reading, its effect was strongly regulated by various specificities of the readers. These specificities included, on the one hand, their personal traits and, on the other hand, their psychological state, which encompasses the cognitive, motivational, and emotional systems.

PERSONAL TRAITS It is striking that not many studies controlled for general traits such as native language (Laufer & Yano, 2001), gender (Elgort & Warren, 2014; Kim et al., 2018; Laufer & Yano, 2001; Lee, Warschauer, et al., 2017, 2018), age of acquisition (Elgort & Warren, 2014), and experience (Pulido & Hambrick, 2008; Watts, 2008). As most of these confounding variables were either largely understudied or yielded mixed results, we cannot draw any strong conclusions at the present moment.

COGNITION, MOTIVATION, EMOTION Whereas the effect of motivation and emotion was explored in only a handful of studies (Elgort & Warren, 2014; Zhao et al., 2016), more support was found for the effect of cognitive factors.

A frequent and salient predictor was L2 proficiency, which determined successful learning (Lee et al., 2018; Zhao et al., 2016) and regulated the effect of word occurrence (i.e., high-proficient learners needed fewer exposures; Elgort & Warren, 2014). Robust effects were also evidenced for reading proficiency (Pulido, 2003, 2004b, 2009), which impacted the time spent processing words as well as successful inferencing (Dolgunsöz, 2016; Dolgunsöz & Sariçoban, 2016; Kaivanpanah & Moghaddam, 2012). Similar effects were observed when considering prior familarity with the topic (Kaivanpanah & Rahimi, 2017; Mahdavy, 2011; Pulido, 2003, 2004a, 2004b, 2007, 2009) or with the text's vocabulary (Elgort et al., 2018).

Proficiency, ability, and working memory capacity also moderated the effectiveness of input enhancement. Firstly, not all proficiency levels benefited equally from the use of dictionaries, glosses, or multimedia annotations (Chen, 2016; Chen & Yen, 2013; Chen, 2012; Cheng & Good, 2009; Duan, 2018; Hu et al., 2014; Lee, Warschauer, et al., 2017; Lee et al., 2016). Secondly, the (in)effectiveness of input enhancement was impacted by the learner's abil-

ity: high verbal ability learners were more efficient when using dictionaries (Knight, 1994), whereas visualizers profited more from multimedia annotations (Plass et al., 1998, 2003). Finally, input enhancement may also be more indispensable for learners with a low working memory, who were prone to look up more words in order to achieve equivalent learning outcomes (Chun & Payne, 2004).

*Interaction*

Because tests of proficiency and ability were generally administered prior to reading, they did not give an indication of the reader's mental state while reading. Although not widely substantiated, lexical competence could also be predicted from the reader's cognitive behavior *while* reading, which encompasses both word processing and comprehension as well as the degree of involvement.

PROCESSING & COMPREHENSION The amount of attention given to unfamiliar words, as measured by the amount and duration of eye fixations (Godfroid et al., 2013), in particular second pass time[15] (Dolgunsöz, 2016), predicted subsequent lexical gains. Furthermore, successful text comprehension had a strong effect on vocabulary retention (Pulido & Hambrick, 2008) and enabled the reader to capitalize on the multiple encounters with a target word (Elgort & Warren, 2014), but it was not always a significant predictor of vocabulary uptake (Boers et al., 2017).

INVOLVEMENT A number of studies corroborated Laufer and Hulstijn (2001)'s hypothesis that, when words are processed with a higher load of involvement, the likelihood that they will be retained increases. When learners were more engaged in choosing the texts they read, their motivation to understand and learn new words increased (Reynolds & Bai, 2013). Furthermore, while acquisition was least effective when involvement was low (ignore, write down the translation after reading a gloss; Plass et al., 1998; Watanabe, 1997), cognitive strategies with a higher involvement load (evaluate the part of speech, infer meaning beyond compositionality, use contextual cues; Ender,

15 i.e., the time spent reading after regressing back to the word

2016; Rott, 2007; Shen, 2008) resulted in superior gains. For instance, meaning recall was highest when learners were required to first evaluate the meaning from context, then consult a gloss to check whether these guesses were correct, and finally retrieve the (revised) meaning upon subsequent encounters (Danesh & Farvardin, 2016).

*Method*

Not all factors attested in the literature were included in the overview above, either because they were understudied or yielded inconclusive results. This inconclusiveness may be related to the manner in which the data was collected. However, apart from studies taking into account the time of measurement to incorporate the degree of memory loss, not many studies controlled for other factors related to how the competence was measured. Yet, it is crucial to understand whether the effect of a predictor on a specific competence was modulated by either the instructional context (Bell & LeBlanc, 2000), the types of tests (Rott, 2007; Türk & Erçetin, 2014; Waring & Takaki, 2003), inconsistencies in objective or subjective measurements (Laufer & Yano, 2001), test announcement (Peters, 2007, 2012b), or by the reactivity of the instruments tracking the reading process (Godfroid & Spino, 2015).

## 2.5 DISCUSSION

The fundamental question addressed in this study was: If vocabulary is key to achieving successful reading in a foreign language, how can this competence be measured and predicted? Given the almost fifty years of publications reviewed, it is clear that this question is not an ephemeral one. Even more so, the scope of research appears to be ever-expanding, as shown by the bulk of studies published in the last five years.

The investigation touches upon different areas aiming to understand lexical competence while reading in a foreign language. The most central area is probably SLA where a number of studies developed psycholinguistic models of how learners process and acquire words contextually while reading. Besides SLA, the development of statistical models is also relevant to areas focusing in particular on the effectiveness of technologically-enhanced learning. In

the area of CALL, predictive analyses may be relevant for knowing how to enhance comprehension and learning (e.g., by means of glosses). In the area of educational NLP, the automated prediction of vocabulary in reading has recently sparked interest with two shared tasks (Paetzold & Specia, 2016a; Yimam et al., 2018) benchmarking the effectiveness of various statistical learning algorithms to identify "complex words" (i.e., the running words in a text which exceed the readers' competence). In particular, systems built on a considerable number of (psycho)linguistic and pedagogical features were able to reach top-level performance. Consequently, in order to further research on the subject, it seems that much is to be gained from reviewing insights from these three areas. However, the review also highlighted a rift between two research areas, viz., applied and computational linguistics, which was manifested in both their underlying theoretical frameworks and methodological approaches. By synthesizing insights stemming from either approaches, the review therefore hopes to have provided a bolstering contribution to bridging this gap.

In the lion's share of studies, we find evidence on how to predict lexical gain after reading based on factors related to the nature of the input. Among the most scrutinized and most pivotal predictors, we note the frequency of exposure to a word, the presence of adequate contextual elaboration, and the enhancing effect of (computer-aided) glosses and dictionaries. Additionally, the learner's mental state prior to reading acts as an important modulating variable. On the other hand, because of the sparsity in the combinations of L2s and L1s, the bias towards convenience samples of highly educated participants, and the small samples of target vocabulary, we might be in need of more representativeness and statistical power to draw generalizable assumptions from the studies as a whole. What is more, the types of measurements and predictors tested across studies display two shortcomings that could arguably pose a limit on achieving more accurate predictions.

A first shortcoming can be seen in the low degree of contextualization observed in the measurements. Given the preponderance of decontextualized vocabulary tests administered after reading and the few procedures conducted while reading, it can be concluded that lexical competence elicited in contextualized reading has mainly been operationalized, and consequently, predicted in a non-contextualized manner. Yet, it is not straightforward to argue whether

resorting to decontextualized post-tests should be seen as a limitation. While isolating the word from its context might give rise to less precise measurements of semantic processing, it could nonetheless be essential to avoid unwanted guessing.

A second shortcoming can be found in the low degree of personalization of the predictions. It is striking that, out of the handful of subject-related aspects examined, most support was given for the effect of cognitive factors such as reading proficiency and prior knowledge. By contrast, hardly any studies investigated the influence of personal characteristics such as the learner's native language, even though their sample of participants consisted of various L1 backgrounds. Furthermore, cognitive factors such as proficiency and motivation only give an indication of the learner's mental state prior to reading. As a result, not many predictions have accounted for how the learner interacts with the input while reading.

In light of some of the strict criteria that were used to delineate the scope of research, these conclusions should, however, be interpreted with caution. Firstly, because foreign language acquisition was defined in a strict sense, the results cannot simply be applied to bilingualism, second and additional language learners, language minorities, or the instruction of a *lingua franca*. Secondly, because only contextualized reading tasks were selected, the synthesis does not factor in evidence from (isolated) lexical decision trials or tasks where reading was supplemented with other learning tasks. Finally, and more consequentially, because non-English publications were excluded, there is a possibility that relevant work issued in other languages may have been glossed over. Although this ubiquitous English-language bias could, and arguably, should have been circumvented, it would, however, have introduced an even more subjective (and potentially problematic) bias towards including only those languages that were mastered by the author(s).

## 2.6 CONCLUSION

This literature study gathered fifty years of insights from SLA, CALL, and NLP into the 'how' and 'what' of predicting lexical competence in foreign language reading. The main aim of the synthesis was to systematically scope

a methodological basis for the development of predictive models and systems, which was carried out in two stages.

A first step was to delineate the scientific scope of studies. Citation analyses showed that the topic was fairly dispersed, as evidenced by the small proportion of productive authors and the discordance between references cited in applied and computational linguistics. Moreover, seeing that a relatively important number of studies did not report with equal methodological precision, the descriptive analyses could not provide an exhaustive picture of all study populations. Nevertheless, the results were able to highlight the need for more sizable samples of vocabulary and for more abundant support of target languages other than English.

A second step was to outline the methodological scope of the predictive analyses. Looking at how lexical competence was measured as a criterion variable, we saw that most of the focus was directed towards predicting incidental learning and retention. Conversely, not many online instruments were used to gauge how meaning was recognized or recalled while reading words in context. Furthermore, regarding the factors used to predict this competence, a considerable amount of evidence was found of the input conveyed to the learner, but only minor support was given for the learner's characteristics and involvement while reading.

In sum, while we can state that lexical competence in L2 reading can be predicted to a certain degree, we hope that future predictive work will provide more insights into how lexical competence unfolds *while* reading, tailoring the predictions even more closely to specific language learners.

## 2.A APPENDIX

This appendix provides some details on how the literature review was conducted. The following information is provided: the documentation of the final search strategy (Section 2.A.1), the keywords that were used for excluding studies (Section 2.A.2), the list of publications included in the synthesis (Section 2.A.3), the final data coding scheme (Section 2.A.4), and the tools and software that were used (Section 2.A.5).

### 2.A.1 *Search Documentation*

*Digital Libraries*

The following citation indices were searched in the Web of Science database: *Social Sciences Citation Index* (SSCI), *Arts & Humanities Citation Index* (A&HCI), *Conference Proceedings Citation Index - Social Science & Humanities* (CPCI-SSH), *Book Citation Index - Social Sciences & Humanities* (BKCI-SSH), *Emerging Sources Citation Index* (ESCI), and *Inspec*. The following citation indices were searched in the ProQuest database: *ProQuest Central*, *Educational Resources Information Center* (ERIC), *Library & Information Science Abstracts* (LISA), *Linguistics and Language Behavior Abstracts* (LLBA), *MLA International Bibliography* (MLA), *Periodicals Archive Online* (PAO), *PsycARTICLES*, and *PsycINFO*. It should be noted that the ProQuest database was searched twice. A first search was performed for retrieving relevant peer-reviewed records. A second search was carried out without a peer-reviewed filter on book chapters only. Because this type generally lacks a peer-reviewed label, these records would have been systematically left out during the first search. A complete documentation of the final search strategy is given below.

```
Search Documentation for Web of Science Core

                                        Date of last search: 07/12/2018
======================================================================
Search criteria (Population, Concept, Method)
----------------------------------------------------------------------
```

```
#1    (28,471 records)
(TS= ( ("foreign language$" OR "second language$" OR L2 OR L3 OR FL OR
     non-native* OR nonnative*) AND (acquisition OR education OR
    instruction OR learn*) )) AND LANGUAGE: (English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
---------------------------------------------------------------------
#2    (11,343 records)
(TS= ( (((passage OR reading OR text OR written) NEAR/0 comprehension)
     OR reading OR "written receptive" OR ((incidental NEAR/1
    learning) NEAR reading)) AND ( (lexicon OR lexis OR vocab* OR ((
    lexic* OR vocab* OR word$) NEAR/0 (breadth OR complex* OR
    competence OR depth OR difficulty OR familiarity OR growth OR
    inferenc* OR identification OR incidental OR know* OR level OR
    perception OR process* OR recall OR recognition OR representation
     OR retention OR size OR skills)) ) OR ("passive vocabular*" OR "
    sight vocabular*" OR "targeted vocabular*" OR "receptive
    vocabular*") OR ("complex word$" OR "target word$" OR "key word$"
     OR "unknown word$") ) )) AND LANGUAGE: (English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
---------------------------------------------------------------------
#3    (3,436,810 records)
(TS= ( annotation OR "complex word identification" OR lookup$ OR "
    dictionary use$" OR gloss* OR "eye?tracking" OR "read?aloud" OR "
    think?aloud" OR judg$ment OR assess* OR model* OR predict* OR
    procedure OR outcome$ OR experiment* OR empirical OR treatment$
    OR "controlled study" OR "vocabulary test$" OR criterion OR
    predictor$ OR feature$ OR "criterion variable$" OR "dependent
    variable$" OR "independent variable$" OR "outcome variable$" OR "
    predictor variable$" OR "target variable$" )) AND LANGUAGE: (
    English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
=====================================================================
Combined search syntax
---------------------------------------------------------------------
#4    (801 records)
(#3 AND #2 AND #1) AND LANGUAGE: (English) AND DOCUMENT TYPES: (
    Article OR Abstract of Published Item OR Book Chapter OR Data
    Paper OR Early Access OR Proceedings Paper OR Reprint OR
    Retracted Publication)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
=====================================================================
```

```
Final, refined search syntax
------------------------------------------------------------------------
#5    (749 records)
(#3 AND #2 AND #1) AND LANGUAGE: (English) AND DOCUMENT TYPES: (
     Article OR Abstract of Published Item OR Book Chapter OR Data
     Paper OR Early Access OR Proceedings Paper OR Reprint OR
     Retracted Publication)
Refined by: [excluding] RESEARCH AREAS: ( AGRICULTURE OR FILM RADIO
     TELEVISION OR OPERATIONS RESEARCH MANAGEMENT SCIENCE OR AREA
     STUDIES OR FOOD SCIENCE TECHNOLOGY OR OTORHINOLARYNGOLOGY OR
     GENERAL INTERNAL MEDICINE OR PHYSIOLOGY OR AUDIOLOGY SPEECH
     LANGUAGE PATHOLOGY OR GENETICS HEREDITY OR PSYCHIATRY OR
     GOVERNMENT LAW OR BUSINESS ECONOMICS OR HEALTH CARE SCIENCES
     SERVICES OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR CLASSICS
     OR INFORMATION SCIENCE LIBRARY SCIENCE OR RADIOLOGY NUCLEAR
     MEDICINE MEDICAL IMAGING OR COMMUNICATION OR REHABILITATION OR
     CULTURAL STUDIES OR DEMOGRAPHY OR MATHEMATICS OR DENTISTRY ORAL
     SURGERY MEDICINE OR MUSIC OR SURGERY OR NURSING OR
     TELECOMMUNICATIONS OR FAMILY STUDIES OR OBSTETRICS GYNECOLOGY )
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
========================================================================
Trial searches on known publications
------------------------------------------------------------------------
#6    (17 records)
(TI=("Dictionary Use While Reading: The Effects on Comprehension and
     Vocabulary Acquisition for Students of Different Verbal Abilities
     ") OR TI=("Effects of Multimedia Annotations on Vocabulary
     Acquisition") OR TI=("Supporting visual and verbal learning
     preferences in a second-language multimedia learning environment
     ") OR TI=("Textual and Pictorial Glosses: Effectiveness on
     Incidental Vocabulary Growth When Reading in a Foreign Language")
      OR TI=("Reading-based exercises in second language vocabulary
     learning: An introspective study") OR TI=("Second Language
     Incidental Vocabulary Retention: The Effect of Picture and
     Annotation Types") OR TI=("Vocabulary acquisition in a second
     language: Do learners really acquire most vocabulary by reading
     Some empirical evidence") OR TI=("L2 Vocabulary Learning from
     Context: Strategies, Knowledge Sources, and Their Relationship
     with Success in L2 Lexical Inferencing") OR TI=("Cognitive load
     in reading a foreign language text with multimedia aids and the
     influence of verbal and spatial abilities") OR TI=("The
```

```
    Relationship between Depth of Vocabulary Knowledge and L2
    Learners' Lexical Inferencing Strategy Use and Success") OR TI=("
    L1 and L2 Glosses: Their Effects on Incidental Vocabulary
    Learning") OR TI=("Second-Language Reading Comprehension and
    Vocabulary Learning with Multimedia") OR TI=("The Effects of
    Topic Familiarity and Passage Sight Vocabulary on L2 Lexical
    Inferencing and Retention through Reading") OR TI=("The Effects
    of Repetition on Vocabulary Knowledge") OR TI=("Vocabulary
    Assistance before and during Reading") OR TI=("The Virtuous
    Circle: Modeling Individual Differences in L2 Reading and
    Vocabulary Development") OR TI=("The Effects of Context on
    Incidental Vocabulary Learning")) AND LANGUAGE: (English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
---------------------------------------------------------------------
#7    (14 records)
(#6 AND #5) AND LANGUAGE: (English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
---------------------------------------------------------------------
#8    (3 records)
(#6 NOT #5) AND LANGUAGE: (English)
Indexes=SSCI, A&HCI, CPCI-SSH, BKCI-SSH, ESCI Timespan=All years
```

Search Documentation for Web of Science Inspec

```
                                        Date of last search: 07/12/2018
=====================================================================
Search criteria (Population, Concept, Method)
---------------------------------------------------------------------
#1    (6,817 records)
(TS= ( ("foreign language$" OR "second language$" OR L2 OR L3 OR FL OR
     non-native* OR nonnative*) AND (acquisition OR education OR
    instruction OR learn*) )) AND LANGUAGE: (English) AND DOCUMENT
    TYPES: (Journal Paper OR Book Chapter OR Conference Paper OR
    Conference Paper In Journal OR Conference Paper In Journal
    Original Abstracted OR Conference Proceedings OR Conference
    Proceedings In Journal OR Conference Proceedings In Journal
    Original Abstracted OR Journal Paper Original Abstracted OR
    Journal Paper Translation Abstracted OR Report)
Indexes=Inspec Timespan=All years
```

```
--------------------------------------------------------------------------
#2    (3,147 records)
(TS= ( (((passage OR reading OR text OR written) NEAR/0 comprehension)
     OR reading OR "written receptive" OR ((incidental NEAR/1
    learning) NEAR reading)) AND ( (lexicon OR lexis OR vocab* OR ((
    lexic* OR vocab* OR word$) NEAR/0 (breadth OR complex* OR
    competence OR depth OR difficulty OR familiarity OR growth OR
    inferenc* OR identification OR incidental OR know* OR level OR
    perception OR process* OR recall OR recognition OR representation
     OR retention OR size OR skills)) ) OR ("passive vocabular*" OR "
    sight vocabular*" OR "targeted vocabular*" OR "receptive
    vocabular*") OR ("complex word$" OR "target word$" OR "key word$"
     OR "unknown word$") ) )) AND LANGUAGE: (English) AND DOCUMENT
    TYPES: (Journal Paper OR Book Chapter OR Conference Paper OR
    Conference Paper In Journal OR Conference Paper In Journal
    Original Abstracted OR Conference Proceedings OR Conference
    Proceedings In Journal OR Conference Proceedings In Journal
    Original Abstracted OR Journal Paper Original Abstracted OR
    Journal Paper Translation Abstracted OR Report)
Indexes=Inspec Timespan=All years
--------------------------------------------------------------------------
#3    (9,586,928 records)
(TS= ( annotation OR "complex word identification" OR lookup$ OR "
    dictionary use$" OR gloss* OR "eye?tracking" OR "read?aloud" OR "
    think?aloud" OR judg$ment OR assess* OR model* OR predict* OR
    procedure OR outcome$ OR experiment* OR empirical OR treatment$
    OR "controlled study" OR "vocabulary test$" OR criterion OR
    predictor$ OR feature$ OR "criterion variable$" OR "dependent
    variable$" OR "independent variable$" OR "outcome variable$" OR "
    predictor variable$" OR "target variable$" )) AND LANGUAGE: (
    English) AND DOCUMENT TYPES: (Journal Paper OR Book Chapter OR
    Conference Paper OR Conference Paper In Journal OR Conference
    Paper In Journal Original Abstracted OR Conference Proceedings OR
     Conference Proceedings In Journal OR Conference Proceedings In
    Journal Original Abstracted OR Journal Paper Original Abstracted
    OR Journal Paper Translation Abstracted OR Report)
Indexes=Inspec Timespan=All years
==========================================================================
Combined search syntax
--------------------------------------------------------------------------
#4    (94 records)
```

```
#3 AND #2 AND #1
Indexes=Inspec Timespan=All years
=========================================================================
Final, refined search syntax
-------------------------------------------------------------------------
#5    (86 records)
#3 AND #2 AND #1
Refined by: [excluding] RESEARCH AREAS: ( ACOUSTICS OR ENERGY FUELS OR
     PHYSICS OR AUTOMATION CONTROL SYSTEMS OR RADIOLOGY NUCLEAR
    MEDICINE MEDICAL IMAGING OR BUSINESS ECONOMICS OR FILM RADIO
    TELEVISION OR GEOGRAPHY OR MATHEMATICAL COMPUTATIONAL BIOLOGY OR
    TELECOMMUNICATIONS OR CONSTRUCTION BUILDING TECHNOLOGY )
Indexes=Inspec Timespan=All years
```

Search Documentation for ProQuest (KU Leuven Libraries)

```
                                        Date of last search: 07/12/2018
=========================================================================
Search criteria (Population, Concept, Method)
-------------------------------------------------------------------------
noft((("foreign language[*1]" OR "second language[*1]" OR L2 OR L3 OR
    FL OR non-native* OR nonnative*) AND (acquisition OR education OR
     instruction OR learn*))) AND
-------------------------------------------------------------------------
noft(((((passage OR reading OR text OR written) NEAR/0 comprehension)
    OR reading OR "written receptive" OR ((incidental NEAR/1 learning
    ) NEAR/4 reading)) AND ((lexicon OR lexis OR vocab* OR ((lexic*
    OR vocab* OR word[*1]) NEAR/0 (breadth OR complex* OR competence
    OR depth OR difficulty OR familiarity OR growth OR inferenc* OR
    identification OR incidental OR know* OR level OR perception OR
    process* OR recall OR recognition OR representation OR retention
    OR size OR skills))) OR ("passive vocabular*" OR "sight vocabular
    *" OR "targeted vocabular*" OR "receptive vocabular*") OR ("
    complex word[*1]" OR "target word[*1]" OR "key word[*1]" OR "
    unknown word[*1]")))) AND
-------------------------------------------------------------------------
noft((annotation OR "complex word identification" OR lookup[*1] OR "
    dictionary use[*1]" OR gloss* OR "eye?tracking" OR "read?aloud"
    OR "think?aloud" OR judg[*1]ment OR assess* OR model* OR predict*
```

```
     OR procedure OR outcome[*1] OR experiment* OR empirical OR
    treatment[*1] OR "controlled study" OR "vocabulary test[*1]" OR
    criterion OR predictor[*1] OR feature[*1] OR "criterion variable
    [*1]" OR "dependent variable[*1]" OR "independent variable[*1]"
    OR "outcome variable[*1]" OR "predictor variable[*1]" OR "target
    variable[*1]"))
================================================================================
Selected collections: Central, LISA, LLBA, MLA, PAO, PsycARTICLES
================================================================================
Years covered by search: all years
================================================================================
Search A    (1,217 records, 6 databases)
--------------------------------------------------------------------------------
- Peer reviewed
- Source type: Conference Papers & Proceedings, Reports, Scholarly
    Journals, Working Papers
- Document type: Article, Conference Paper, Correction/Retraction,
    Working Paper/Pre-Print
- Language: English
================================================================================
Search B    (18 records, 6 databases)
--------------------------------------------------------------------------------
- Source type: Books
- Document type: Book Chapter
- Language: English
```

Search Documentation for ProQuest (UCLouvain Libraries)

```
                                        Date of last search: 07/12/2018
================================================================================
Search criteria (Population, Concept, Method)
--------------------------------------------------------------------------------
noft((("foreign language[*1]" OR "second language[*1]" OR L2 OR L3 OR
    FL OR non-native* OR nonnative*) AND (acquisition OR education OR
     instruction OR learn*))) AND
--------------------------------------------------------------------------------
noft(((((passage OR reading OR text OR written) NEAR/0 comprehension)
    OR reading OR "written receptive" OR ((incidental NEAR/1 learning
    ) NEAR/4 reading)) AND ((lexicon OR lexis OR vocab* OR ((lexic*
```

```
    OR vocab* OR word[*1]) NEAR/0 (breadth OR complex* OR competence
    OR depth OR difficulty OR familiarity OR growth OR inferenc* OR
    identification OR incidental OR know* OR level OR perception OR
    process* OR recall OR recognition OR representation OR retention
    OR size OR skills))) OR ("passive vocabular*" OR "sight vocabular
    *" OR "targeted vocabular*" OR "receptive vocabular*") OR ("
    complex word[*1]" OR "target word[*1]" OR "key word[*1]" OR "
    unknown word[*1]")))) AND
------------------------------------------------------------------------
noft((annotation OR "complex word identification" OR lookup[*1] OR "
    dictionary use[*1]" OR gloss* OR "eye?tracking" OR "read?aloud"
    OR "think?aloud" OR judg[*1]ment OR assess* OR model* OR predict*
     OR procedure OR outcome[*1] OR experiment* OR empirical OR
    treatment[*1] OR "controlled study" OR "vocabulary test[*1]" OR
    criterion OR predictor[*1] OR feature[*1] OR "criterion variable
    [*1]" OR "dependent variable[*1]" OR "independent variable[*1]"
    OR "outcome variable[*1]" OR "predictor variable[*1]" OR "target
    variable[*1]"))
========================================================================
Selected collections: ERIC, LLBA, PAO, PsycARTICLES, PsycINFO
========================================================================
Years covered by search: all years
========================================================================
Search A    (2,807 records, 5 databases)
------------------------------------------------------------------------
- Peer reviewed
- Source type: Conference Papers & Proceedings, Reports, Scholarly
    Journals, Working Papers
- Document type: Article, Conference Paper, Correction/Retraction,
    Working Paper/Pre-Print
- Language: English
========================================================================
Search B    (59 records, 6 databases)
------------------------------------------------------------------------
- Source type: Books
- Document type: Book Chapter
- Language: English
```

*ACL Anthology*

The following peer-reviewed journals and proceedings were targeted: *Computational Linguistics*, *Transactions of the Association for Computational Linguistics* (TACL), *Annual Meeting of the Association for Computational Linguistics* (ACL), *International Conference on Computational Linguistics* (COLING), *Conference on Computational Natural Language Learning* (CoNLL), *Conference of the European Chapter of the Association for Computational Linguistics* (EACL), *Conference on Empirical Methods in Natural Language Processing* (EMNLP), *International Joint Conference on Natural Language Processing* (IJCNLP), *International Conference on Language Resources and Evaluation* (LREC), *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), *International Workshop on Semantic Evaluation* (SemEval), *International Conference Recent Advances in Natural Language Processing* (RANLP), and *Workshop on the Innovative Use of NLP for Building Educational Applications* (BEA). Two URLs were used to query the ACL Anthology: one corresponding to the final search strategy and an another, shorter URL matching papers on complex word identification. It should be noted that only the "Paper Metadata" tab was consulted.

**Google Custom Search URL for Querying the ACL Anthology**

```
https://aclweb.org/anthology/search/?q=%28%22foreign+language%22+OR
    +%22second+language%22+OR+L2+OR+L3+OR+FL+OR+non-native+OR+
    nonnative%29++AND+%28acquisition+OR+education+OR+instruction+OR+
    learner+OR+learning%29+AND+%28comprehension+OR+reading+OR+
    receptive%29+AND+%28lexicon+OR+lexis+OR+vocabulary+OR+%28%28
    lexical+OR+word%29+AND+%28complex+OR+complexity+OR+competence+OR+
    difficulty+OR+familiarity+OR+growth+OR+inferencing+OR+knowledge+
    OR+perception+OR+recall+OR+recognition+OR+representation+OR+
    retention+OR+skills%29+OR+%22passive+vocabulary%22+OR+%22sight+
    vocabulary%22+OR+%22targeted+vocabulary%22+OR+%22receptive+
    vocabulary%22+OR+%22complex+word%22+OR+%22unknown+word%22%29%29+
    AND+%28annotation+OR+%22complex+word+identification%22+OR+lookup+
    OR+%22dictionary+use%22+OR+gloss+OR+%22eye-tracking%22+OR+%22read
    -aloud%22+OR+%22think-aloud%22+OR+judgment+OR+assess+OR+
    assessment+OR+model+OR+predict*+OR+procedure+OR+outcome+OR+
```

```
    experiment+OR+empirical+OR+treatment+OR+%22controlled+study%22+OR
    +%22vocabulary+test%22+OR+criterion+OR+predictor+OR+feature+OR
    ++%22criterion+variable%22+OR+%22dependent+variable%22+OR+%22
    independent+variable%22+OR+%22outcome+variable%22+OR+%22predictor
    +variable%22+OR+%22target+variable%22%29

https://aclweb.org/anthology/search/?q=%22complex+word+identification
    %22
```

### 2.A.2 *Exclusion Keywords*

*Population*

The following keywords were used to exclude studies beyond foreign language acquisition: additional language, bilinguals, bilingualism, bilingual education, biliteracy, CLIL, dual language, EAL, ELL, emergent literacy, emergent readers, ESOL, heritage speakers, immersion (programs), language minority, (im)migrants, pluri-ethnic settings, plurilingual schools, secondary anglophones, or situated learning.

The following keywords were used to exclude studies focusing on disabilities: aphasia, aphasic, at-risk students, disorder, dyslexia, dysgraphia, deaf, epilepsy, impairment, language delay, low achievers, mental deficits, mental retardation, intellectual, literacy and reading disabilities, sign language, signers, special education, struggling or poor readers and/or comprehenders, slow development, or down syndrome.

*Concept*

The following keywords were used to exclude non-contextualized reading tasks: character reading, flashcard reading, lexical decision task, masked priming, *n*-back tasks, rapid naming, rapid serial visual presentation, semantic judgment trials, single word processing, Stroop paradigm, word attack skills, word naming, and word reading. Importantly, we only excluded studies where the main task referred to the listed keywords. Studies that combined a relevant

reading task with offline vocabulary measurements involving a lexical decision task, for instance, were therefore not excluded.

The following keywords were used to exclude multimodal reading tasks: audio-visual materials, dictation, dual-modality input, graphic novels, comic books, manga-based e-books, guided phonological practice, listening while reading, listen and read aloud, reading while listening, reading with audio support, (bimodal) video, video subtitles.

The following keywords were used to exclude non-receptive reading tasks: cloze reading passages for word retrieval, chat, SMS, (computer-mediated) communication, SCMC, cooperative reading, dialogic reading, input-output cycles for reading, reading-based (collaborative) output activities, shared reading, social reading, synetics instruction, reading and summarizing, reading and retelling, reading and story rewriting, and spelling.

### 2.A.3 *Publications Included in the Synthesis*

Abdollahi, M., & Farvardin, A. T. (2016). Demystifying the effect of narrow reading on EFL learners' vocabulary recall and retention. *Education Research International*, *2016*(5454031), 1–10. https://doi.org/10.1155/2016/5454031

Abraham, L. B. (2007). Second-language reading comprehension and vocabulary learning with multimedia. *Hispania*, *90*(1), 98–108. https://doi.org/10.2307/20063468

AbuSeileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. *Computers & Education*, *57*(1), 1281–1291. https://doi.org/10.1016/j.compedu.2011.01.011

AbuSeileek, A. F. M. (2008). Hypermedia annotation presentation: Learners' preferences and effect on EFL reading comprehension and vocabulary acquisition. *CALICO Journal*, *25*(2), 260–275.

Adams, S. J. (1982). Scripts and the recognition of unfamiliar vocabulary: Enhancing second language reading skills. *Modern Language Journal*, *66*(2), 155–159.

Arpaci, D. (2016). The effects of accessing L1 versus L2 definitional glosses on L2 learners' reading comprehension and vocabulary learning. *Eurasian Journal of Applied Linguistics*, *2*(1), 15–29.

Bell, F. L., & LeBlanc, L. B. (2000). The language of glosses in L2 reading on computer: Learners' preferences. *Hispania*, *83*(2), 274–285.

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*(1), 15–31. https://doi.org/10.1111/j.1467-9817.1984.tb00252.x

Bingel, J., Schluter, N., & Martínez Alonso, H. (2016). CoastalCPH at SemEval-2016 Task 11: The importance of designing your neural networks right. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1028–1033. http://www.aclweb.org/anthology/S16-1160

Birjandi, P., Alavi, S. M., & Najafi Karimi, S. (2015). Effects of unenhanced, enhanced, and elaborated input on learning English phrasal verbs. *International Journal of Research Studies in Language Learning*, *4*(1), 43.

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, *66*, 113.

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, *87*(3), 541–552.

Brooke, J., Uitdenbogerd, A., & Baldwin, T. (2016). Melbourne at SemEval 2016 Task 11: Classifying type-level word complexity using random forests with corpus and word list features. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 975–981. http://www.aclweb.org/anthology/S16-1150

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693–713.

Chen, I.-J. (2016). Hypertext glosses for foreign language reading comprehension and vocabulary acquisition: Effects of assessment methods. *Computer Assisted Language Learning*, *29*(2), 413–426. https://doi.org/10.1080/09588221.2014.983935

Chen, I.-J., & Yen, J.-C. (2013). Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vo-

cabulary learning in foreign languages. *Computers & Education*, *63*, 416–423. https://doi.org/10.1016/j.compedu.2013.01.005

Chen, Y. (2012). Dictionary use and vocabulary learning in the context of reading. *International Journal of Lexicography*, *25*(2), 216–247.

Cheng, Y.-H., & Good, R. L. (2009). L1 glosses: Effects on EFL learners' reading comprehension and vocabulary retention. *Reading in a Foreign Language*, *21*(2), 119–142.

Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences*, *45*, 137–143. https://doi.org/10.1016/j.lindif.2015.11.018

Choubey, P., & Pateria, S. (2016). Garuda & Bhasha at SemEval-2016 Task 11: Complex word identification using aggregated learning models. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1006–1010. http://www.aclweb.org/anthology/S16-1156

Chun, D. M., & Payne, J. S. (2004). What makes students click: Working memory and look-up behavior. *System*, *32*(4), 481–503.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal*, *80*(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Danesh, T., & Farvardin, M. T. (2016). A comparative study of the effects of different glossing conditions on EFL learners' vocabulary recall. *Sage Open*, *6*(3), UNSP 2158244016669548. https://doi.org/10.1177/2158244016669548

Davoodi, E., & Kosseim, L. (2016). CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 982–985. http://www.aclweb.org/anthology/S16-1151

De Ridder, I. (2000). Are we conditioned to follow links? Highlights in CALL materials and their impact on the reading process. *Computer Assisted Language Learning*, *13*(2), 183–94. https://doi.org/10.1076/0958-8221(200004)13:2;1-D;FT183

De Ridder, I. (2002). Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? *Language Learning & Technology*, *6*(1), 123–46.

Dilenschneider, R. F. (2018). Examining the conditions of using an on-line dictionary to learn words and comprehend texts. *ReCALL*, *30*(1), 4–23.

Dolgunsöz, E. (2016). Using eye-tracking to measure lexical inferences and its effects on reading rate during EFL reading. *Journal of Language and Linguistic Studies*, *12*(1), 63–78.

Dolgunsöz, E., & Sariçoban, A. (2016). CEFR and eye movement characteristics during EFL reading: The case of intermediate readers. *Journal of Language and Linguistic Studies*, *12*(2), 238–252.

Duan, S. (2018). Effects of enhancement techniques on L2 incidental vocabulary learning. *English Language Teaching*, *11*(3), 88–101.

Ebadi, S., Weisi, H., Monkaresi, H., & Bahramlou, K. (2018). Exploring lexical inferencing as a vocabulary acquisition strategy through computerized dynamic assessment and static assessment. *Computer Assisted Language Learning*, *31*(7), 790–817. https://doi.org/10.1080/09588221.2018.1451344

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365–414. https://doi.org/10.1111/lang.12052

Ender, A. (2016). Implicit and explicit cognitive processes in incidental vocabulary acquisition. *Applied Linguistics*, *37*(4), 536–560. https://doi.org/10.1093/applin/amu051

Farvardin, M. T., & Biria, R. (2012). The impact of gloss types on Iranian EFL students' reading comprehension and lexical retention. *International Journal of Instruction*, *5*(1), 99–114.

Fisher, T., Sharples, M., Pemberton, R., Ogata, H., Uosaki, N., Edmonds, P., Hull, A., & Tschorn, P. (2012). Incidental second language vocabulary

learning from reading novels: A comparison of three mobile modes. *International Journal of Mobile and Blended Learning*, *4*(4), 47–61.

Furtner, M. R., Rauthmann, J. F., & Sachse, P. (2011). Investigating word class effects in first and second languages. *Perceptual and Motor Skills*, *113*(1), 87–97. https://doi.org/10.2466/04.11.28.PMS.113.4.87-97

Gasigijtamrong, J. (2013). Effects of multimedia annotations on Thai EFL readers' words and text recall. *English Language Teaching*, *6*(12), 48–57.

Ghahari, S., & Heidarolad, M. (2015). Multiple-choice glosses and incidental vocabulary learning: A case of an EFL context. *The Reading Matrix*, *15*(1), 262.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483–517. https://doi.org/10.1017/S0272263113000119

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*(4), 896–928. https://doi.org/10.1111/lang.12136

Goudarzi, Z., & Moini, M. R. (2012). The effect of input enhancement of collocations in reading on collocation learning and retention of EFL learners. *International Education Studies*, *5*(3), 247–258.

Hamada, A. (2015). Effects of forward and backward contextual elaboration on lexical inferences: Evidence from a semantic relatedness judgment task. *Reading in a Foreign Language*, *27*(1), 1–21.

Hanifi, S., Nasiri, M., & Aliasin, H. (2016). Dynamic assessment of incidental vocabularies: A case of Iranian ESP learners. *Advances in Language and Literary Studies*, *7*(2), 163–170.

Hayati, M., & Fattahzadeh, A. (2006). The effect of monolingual and bilingual dictionaries on vocabulary recall and retention of EFL learners. *The Reading Matrix*, *6*(2), 125–134.

Holley, F. M., & King, J. K. (1971). Vocabulary glosses in foreign language reading materials. *Language Learning*, *21*(2), 213–219. https://doi.org/10.1111/j.1467-1770.1971.tb00060.x

Hu, S.-M., Vongpumivitch, V., Chang, J. S., & Liou, H.-C. (2014). The effects of L1 and L2 e-glosses on incidental vocabulary learning of junior

high-school English students. *ReCALL*, *26*(1), 80–99. https://doi.org/10.1017/S0958344013000244

Huang, H.-T., & Liou, H.-C. (2007). Vocabulary learning in an automated graded reading program. *Language Learning & Technology*, *11*(3), 64–82.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal*, *80*(3), 327–339. https://doi.org/10.2307/329439

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, *17*(1), 19–28.

Jung, J. (2016). Effects of glosses on learning of L2 grammar and vocabulary. *Language Teaching Research*, *20*(1), 92–112. https://doi.org/10.1177/1362168815571151

Kaivanpanah, S., & Moghaddam, M. (2012). Knowledge sources in EFL learners' lexical inferencing across reading proficiency levels. *RELC Journal*, *43*(3), 373–391. https://doi.org/10.1177/0033688212469219

Kaivanpanah, S., & Rahimi, N. (2017). The effect of contextual clues and topic familiarity on L2 lexical inferencing and retention. *Porta Linguarum*, (27), 47–61. https://dialnet.unirioja.es/servlet/articulo?codigo=6151249

Kang, E. Y. (2015). Promoting L2 vocabulary learning through narrow reading. *RELC Journal*, *46*(2), 165–179. https://doi.org/10.1177/0033688215586236

Karbalaei, A., Sattari, A., & Nezami, Z. (2016///Jan/Jun). A comparison of the effect of text-picture and audio-picture annotations in second language vocabulary recall among Iranian EFL learners. *GiST*, *12*, 51.

Kauchak, D. (2016). Pomona at SemEval-2016 Task 11: Predicting word complexity based on corpus frequency. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1047–1051. http://www.aclweb.org/anthology/S16-1164

Ke, S. E., & Koda, K. (2017). Contributions of morphological awareness to adult L2 chinese word meaning inferencing. *Modern Language Journal*, *101*(4), 742–755.

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*(5), 1155–1180.

Kim, Y. (2006). Effects of input elaboration on vocabulary acquisition through reading by Korean learners of English as a foreign language. *TESOL Quarterly*, *40*(2), 341–373.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, *78*(3), 285–299. https://doi.org/10.2307/330108

Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, *46*(1), 56–79. https://doi.org/10.1002/tesq.3

Konkol, M. (2016). UWB at SemEval-2016 Task 11: Exploring features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1038–1041. http://www.aclweb.org/anthology/S16-1162

Kost, C. R., Foss, P., & Lenzini, J. J. (SPR 1999). Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals*, *32*(1), 89–113. https://doi.org/10.1111/j.1944-9720.1999.tb02378.x

Kuru, O. (2016). AI-KU at SemEval-2016 Task 11: Word embeddings and substring features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1042–1046. http://www.aclweb.org/anthology/S16-1163

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, *13*(2), 549–66.

Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, *21*(2), 32.

Lee, H., Warschauer, M., & Lee, J. H. (2018). Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL*. https://doi.org/10.1017/S0958344018000162

Lee, H., Lee, H., & Lee, J. H. (2016). Evaluation of electronic and paper textual glosses on second language vocabulary learning and reading comprehension. *Asia-Pacific Education Researcher*, *25*(4), 499–507. https://doi.org/10.1007/s40299-015-0270-1

Lee, J. F. (1998). The relationship of verb morphology to second language reading comprehension and input processing. *Modern Language Journal*, *82*(1), 33–48.

Liu, T.-C., & Lin, P.-H. (2011). What comes with technological convenience? Exploring the behaviors and performances of learning with computer-mediated dictionaries. *Computers in Human Behavior*, *27*(1), 373–383. https://doi.org/10.1016/j.chb.2010.08.016

Liu, Y.-T., & Leveridge, A. N. (2017). Enhancing L2 vocabulary acquisition through implicit reading support cues in e-books. *British Journal of Educational Technology*, *48*(1), 43–56. https://doi.org/10.1111/bjet.12329

Mahdavy, B. (2011). The role of topic familiarity and rhetorical organization of texts in L2 incidental vocabulary acquisition. In Z. Bekirogullari (Ed.), *Proceedings of the 2nd International Conference on Education and Educational Psychology*. Elsevier Science Bv.

Malmasi, S., Dras, M., & Zampieri, M. (2016). LTG at SemEval-2016 Task 11: Complex word identification with classifier ensembles. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 996–1000. http://www.aclweb.org/anthology/S16-1154

Malmasi, S., & Zampieri, M. (2016). MAZA at SemEval-2016 Task 11: Detecting lexical complexity using a decision stump meta-classifier. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 991–995. http://www.aclweb.org/anthology/S16-1153

Martínez Martínez, J. M., & Tan, L. (2016). USAAR at SemEval-2016 Task 11: Complex word identification with sense entropy and sentence perplexity. *Proceedings of the 10th International Workshop on Semantic Evalua-*

*tion (SemEval-2016)*, 958–962. http://www.aclweb.org/anthology/S16-1147

Marzban, A., & Hadipour, R. (2012). Depth versus breadth of vocabulary knowledge: assessing their roles in Iranian intermediate EFL students' lexical inferencing success through reading. In G. A. Baskan, F. Ozdamli, S. Kanbul, & D. Ozcan (Eds.), *Proceedings of the 4th World Conference on Educational Sciences (WCES-2012)* (pp. 5296–5300). Elsevier Science Bv.

Mukherjee, N., Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 Task 11: Identifying complex words in a sentence. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 986–990. http://www.aclweb.org/anthology/S16-1152

O'Donnell, M. E. (2009). Finding middle ground in second language reading: Pedagogic modifications that increase comprehensibility and vocabulary acquisition while preserving authentic text features. *Modern Language Journal*, *93*(4), 512–533. https://doi.org/10.1111/j.1540-4781.2009.00928.x

Paetzold, G., & Specia, L. (2016a). SVooogg at SemEval-2016 Task 11: Heavy gauge complex word identification with system voting. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 969–974. http://www.aclweb.org/anthology/S16-1149

Paetzold, G. H., & Specia, L. (2016b–December 17). Understanding the lexical simplification needs of non-native speakers of English. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 717–727.

Palakurthi, A., & Mamidi, R. (2016). IIIT at SemEval-2016 Task 11: Complex word identification using nearest centroid classification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1017–1021. http://www.aclweb.org/anthology/S16-1158

Paribakht, T. S. (2005). The influence of first language lexicalization on second language lexical inferencing: A study of Farsi-speaking learners of English as a foreign language. *Language Learning*, *55*(4), 701–748. https://doi.org/10.1111/j.0023-8333.2005.00321.x

Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, *11*(2), 36–58.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *Language Learning Journal*, *40*(1), 65–79. https://doi.org/10.1080/09571736.2012.658224

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior*, *19*(2), 221–243. https://doi.org/10.1016/S0747-5632(02)00015-8

Prichard, C. (2008). Evaluating L2 readers' vocabulary strategies and dictionary use. *Reading in a Foreign Language*, *20*(2), 216–231.

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, *53*(2), 233–284.

Pulido, D. (2004a). The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix*, *4*(2), 20–53.

Pulido, D. (2004b). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, *54*(3), 469–523. https://doi.org/10.1111/j.0023-8333.2004.00263.x

Pulido, D. (2007). The effects of topic familiarity and passage sight vocabulary on L2 lexical inferencing and retention through reading. *Applied Linguistics*, *28*(1), 66–86. https://doi.org/10.1093/applin/aml049

Pulido, D. (2009). How involved are American L2 learners of Spanish in lexical input processing tasks during reading? *Studies in Second Language Acquisition*, *31*(1), 31–58. https://doi.org/10.1017/S0272263109090020

Pulido, D., & Hambrick, D. Z. (2008). The virtuous circle: Modeling individual differences in L2 reading and vocabulary development. *Reading in a Foreign Language*, *20*(2), 164–190.

Quijada, M., & Medero, J. (2016). HMC at SemEval-2016 Task 11: Identifying complex words using depth-limited decision trees. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1034–1037. http://www.aclweb.org/anthology/S16-1161

Reynolds, B. L. (2016). The effects of target word properties on the incidental acquisition of vocabulary through reading. *TESL-EJ*, *20*(3), 1–31.

Reynolds, B. L., & Bai, Y. L. (2013). Does the freedom of reader choice affect second language incidental vocabulary acquisition? *British Journal of Educational Technology*, *44*(2), E42–E44. https://doi.org/10.1111/j.1467-8535.2012.01322.x

Ronzano, F., Abura'ed, A., Espinosa Anke, L., & Saggion, H. (2016). TALN at SemEval-2016 Task 11: Modelling complex words by contextual, lexical and semantic features. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1011–1016. http://www.aclweb.org/anthology/S16-1157

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*(2), 165–199. https://doi.org/10.1111/j.1467-9922.2007.00406.x

Rouhi, A., & Mohebbi, H. (2012). The effect of computer assisted L1 and L2 glosses on L2 vocabulary learning. *Journal of Asia TEFL*, *9*(2), 1–19.

Rouhi, A., & Mohebbi, H. (2013). Glosses, spatial intelligence, and L2 vocabulary learning in multimedia context. *3L: The Southeast Asian Journal of English Language Studies*, *19*(2), 75–87.

Shahrokni, S. A. (2009). Second language incidental vocabulary learning: The effect of online textual, pictorial, and textual pictorial glosses. *TESL-EJ*, *13*(3), 1–27.

Shen, H. H. (2008). An analysis of word decision strategies among learners of Chinese. *Foreign Language Annals*, *41*(3), 501–524. https://doi.org/10.1111/j.1944-9720.2008.tb03309.x

Shokouhi, H., & Maniati, M. (2009). Learners' incidental vocabulary acquisition: A case on narrative and expository texts. *English Language Teaching*, *2*(1), 13–23.

sp, s., Kumar, A., & K P, S. (2016). AmritaCEN at SemEval-2016 Task 11: Complex word identification using word embedding. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1022–1027. http://www.aclweb.org/anthology/S16-1159

Suk, N. (2017). The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Reading Research Quarterly*, *52*(1), 73–89. https://doi.org/10.1002/rrq.152

Sun, H. (2014). The effects of exposure frequency and contextual richness in reading on Chinese EFL learners' vocabulary acquisition. *Chinese Journal of Applied Linguistics*, *37*(1), 86–106. https://doi.org/10.1515/cjal-2014-0006

Tabatabaei, O., & Shams, N. (2011). The effect of multimedia glosses on online computerized L2 text comprehension and vocabulary learning of Iranian EFL learners. *Journal of Language Teaching and Research*, *2*(3), 714–725. https://doi.org/10.4304/jltr.2.3.714-725

Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning*, *27*(1), 1–25. https://doi.org/10.1080/09588221.2012.692384

Wang, Y.-H. (2013). Incidental vocabulary learning through extensive reading: A case of lower-level EFL Taiwanese learners. *Journal of Asia TEFL*, *10*(3), 59–80.

Wang, Y.-H. (2016). Promoting contextual vocabulary learning through an adaptive computer-assisted EFL reading system. *Journal of Computer Assisted Learning*, *32*(4), 291–303. https://doi.org/10.1111/jcal.12132

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2).

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *19*(3), 287–307.

Watts, M. L. (2008). Clause type and word saliency in second language incidental vocabulary acquisition. *The Reading Matrix*, *8*(1), 1–22.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/amlo48

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245. https://eric.ed.gov/?id=EJ815123

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48–67.

Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017a). CWIG3G2 - Complex word identification task across three text genres and two user groups. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 401–407. http://www.aclweb.org/anthology/I17-2068

Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017b). Multilingual and cross-lingual complex word identification. *Proceedings of Recent Advances in Natural Language Processing*, 813–822. https://doi.org/10.26615/978-954-452-049-6_104

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*(3), 85–101.

Zampieri, M., Tan, L., & van Genabith, J. (2016). MacSaar at SemEval-2016 Task 11: Zipfian and character features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1001–1005. http://www.aclweb.org/anthology/S16-1155

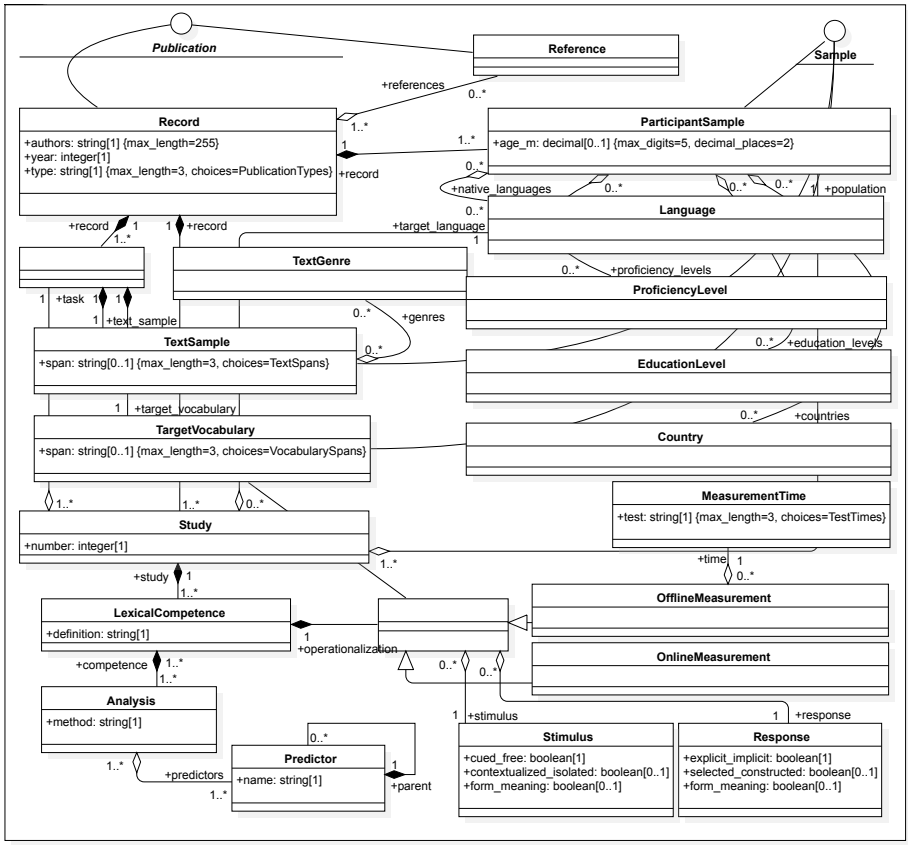
Zhao, A., Guo, Y., Biales, C., & Olszewski, A. (2016). Exploring learner factors in second language (L2) incidental vocabulary acquisition through reading. *Reading in a Foreign Language*, *28*(2), 224–245.

Zou, D. (2016). Comparing dictionary-induced vocabulary learning and inferencing in the context of reading. *Lexikos*, *26*, 372–390.

### 2.A.4 *Data Coding Variables*

The data that were extracted from the selected publications were saved in an SQL database. The structure of this relational database is given below. A more detailed list of all coding variables is given below and in Table 2.3.

**Figure 2.16**
*UML Class Diagram of the Database Structure for Data Extraction*




*Bibliographic Variables*

For each record, the following fields were extracted automatically: the citation key; the authors; the year, type, and source of publication; and the list of cited references. Because not all bibliographic references were included in the database records, the lists of cited references were retrieved with the *scopus*

package (Rose & Kitchin, 2019). For records not covered in Scopus, *AnyStyle*[16] was used to automatically extract references either from the record's PDF file or from an OCR-ized version (scanned articles). Finally, an in-house Python script was used to normalize the different citation keys that referred to the same reference.

*Study Variables*

All variables were hand-coded. Importantly, a number of publications did not provide detailed information on the study participants and on how the reading task was administered. Due to this methodological imprecision, several database fields were incomplete and some variables even had a majority of missing values (e.g., $M_{age}$, $SD_{age}$).

**Table 2.3**
*Coding Scheme for Data Extraction*

| Variable | Value |
|---|---|
| BIBLIOGRAPHIC VARIABLES | |
| Citation key | *categorical: nominal, unique, open* |
| | `Web of Science citation key format` |
| Authors | *categorical: nominal, multiple, open* |
| Year | *numerical: discrete, single, open* |
| Publication type | *categorical: nominal, single, closed* |
| | `ART` = journal article |
| | `CHA` = book chapter |
| | `PAP` = conference paper |
| Publication source | *categorical: nominal, single, open* |
| Cited references | *categorical: nominal, multiple, open* |
| | `Web of Science citation key format` |
| STUDY VARIABLES | |
| Study number | *numerical: discrete, single, open* |
| Target language, L2 | *categorical: nominal, single, closed* |

16 https://anystyle.io/

**Table 2.3**

*Coding Scheme for Data Extraction (Continued)*

| Variable | Value |
|---|---|
| | `ISO 639-3 alpha-3 code` |
| **Study population variables** | |
| Participants (*N*) | *numerical: discrete, single, open* |
| Age (*M* + *SD*) | *numerical: continuous, single, open* |
| Native languages, L1 | *categorical: nominal, multiple, closed* |
| | `ISO 639-3 alpha-3 code` |
| Countries | *categorical: nominal, multiple, closed* |
| | `ISO 3166-1 alpha-3 code` |
| Education levels | *categorical: ordinal, multiple, closed* |
| | `ISCED 2011 level 0-9` |
| Proficiency levels | *categorical: ordinal, multiple, closed* |
| | `BEG` = beginner (to pre-intermediate) |
| | `INT` = intermediate (low- to high-intermediate) |
| | `ADV` = advanced |
| **Reading task variables** | |
| Text sample (*N*) | *numerical: discrete, single, open* |
| Text span | *categorical: ordinal, single, closed* |
| | `SEN` = sentence |
| | `CON` = context |
| | `PAR` = paragraph |
| | `PAS` = passage |
| | `TEX` = text |
| | `REA` = reader |
| Text length (*M*) | *numerical: continous, single, open* |
| | in number of words |
| Text genres | *categorical: nominal, multiple, closed* |
| | `DES` = descriptive |
| | `EXP` = expository |
| | `HOR` = hortatory |

**Table 2.3**

*Coding Scheme for Data Extraction (Continued)*

| Variable | Value |
|---|---|
| | NAR = narrative |
| | PRO = procedural |
| **Target vocabulary variables** | |
| Sample (*N*) | *numerical: discrete, single, open* |
| Vocabulary span | *categorical: ordinal, single, closed* |
| | TOK = token |
| | COL = collocation |
| | PHR = phrase |
| | FOR = formulaic sequence |
| Selection criterion | *categorical: nominal, single, open* |
| LEXICAL COMPETENCE VARIABLES | |
| Definition | *categorical: nominal, single, open* |
| **Measurement variables** | |
| Stimulus | *dummy variables* |
| | 1/0 = cued/free |
| | 1/0/NULL = contextualized/isolated/unspecified or both values |
| | 1/0/NULL = form/meaning/unspecified or both values |
| | 1/0/NULL = L2/L1/unspecified or both values |
| Response | *dummy variables* |
| | 1/0 = explicit/implicit |
| | 1/0/NULL = selected/constructed/unspecified or both values |
| | 1/0/NULL = form/meaning/unspecified or both values |
| | 1/0/NULL = L2/L1/unspecified or both values |
| ***Online measurement*** | y/n |
| Online stimulus | → Stimulus |
| Online response | → Response |
| Online instrument | *categorical: nominal, single, open* |
| | ANN = annotate (highlight) difficult words |

**Table 2.3**

*Coding Scheme for Data Extraction (Continued)*

| Variable | Value |
|---|---|
| | `EEG` = electroencephalography (EEG) |
| | `EYE` = eye-tracking |
| | `GLO` = gloss-tracking |
| | `INF` = infer meaning highlighted words |
| | `JDG` = semantic relatedness judgment |
| | `THI` = think-aloud protocols |
| | `SAS` = self-assessment |
| | `SCA` = vocabulary knowledge scale (VKS) |
| | `SPR` = self-paced reading |
| | `WBW` = self-paced word-by-word reading |
| | `WDT` = check known, supply meaning |
| ***Offline measurement*** | `y/n` |
| Offline stimulus | → Stimulus |
| Offline response | → Response |
| Offline instrument | *categorical: nominal, single, open* |
| | `ASS` = paradigmatic word association |
| | `CHE` = check remember from text |
| | `CHS` = check known, supply meaning |
| | `CLO` = cloze test |
| | `COM` = word completion |
| | `DEF` = definition-supply test |
| | `DIC` = dictation test |
| | `FAM` = check familiarity prior to reading |
| | `JDG` = semantic relatedness judgment |
| | `MAT` = matching definition + form |
| | `MAC` = matching cloze test |
| | `MCC` = multiple-choice cloze |
| | `MCQ` = multiple-choice questionnaire |
| | `PRI` = priming in lexical decision task |
| | `RYN` = check remember from text |

**Table 2.3**

*Coding Scheme for Data Extraction (Continued)*

| Variable | Value |
|---|---|
| | `SAS` = self-assessment |
| | `SCA` = vocabulary knowledge scale (VKS) |
| | `SEN` = sentence construction |
| | `SPE` = spelling test |
| | `SYN` = syntagmatic word association |
| | `TRA` = translation L1-L2 |
| Test time | *categorical: nominal, single, closed* |
| | `PRE` = pretest |
| | `POS` = posttest |
| | `IMM` = immediate posttest |
| | `DEL` = delayed posttest |
| Time span | *categorical: nominal, single, closed* |
| | `[0-9]+d` = x days |
| | `[0-9]+w` = x weeks |
| | `[0-9]+m` = x months |
| | `[0-9]+y` = x years |
| ANALYSIS VARIABLES | |
| Method | *categorical: nominal, single, open* |
| Model | *categorical: nominal, single, open* |
| Criteria | → LEXICAL COMPETENCE (*multiple*) |
| Predictors | *categorical: nominal, multiple, open* |

*Note.* All variables were coded either as discrete/continuous numerical data or as nominal/ordinal categorical data. Each variable was coded with unique, single or multiple values, accepting either an open or closed range of possible values.

### 2.A.5 *Tools and Software*

The following tools and libraries were used for data analysis and visualization: *scipy* (Jones et al., 2001), *pandas* (McKinney, 2010), *statsmodels* (Seabold &

Perktold, 2010), and *scikit-learn* (Pedregosa et al., 2011) for statistical analyses, *matplotlib* (Hunter, 2007) and *seaborn*[17] for plotting graphs, *cartopy* (Met Office, 2010–2015) for plotting choropleth graphs, *floWeaver* (Lupton & Allwood, 2017) for plotting Sankey diagrams, *NetworkX* (Hagberg et al., 2008) for network analysis and visualization, *metaknowledge* (McLevey & McIlroy-Young, 2017) for the generation of bibliographic coupling and co-citation graphs, and *python-louvain*[18] for Louvain community detection.

17 https://seaborn.pydata.org
18 https://python-louvain.readthedocs.io

CHAPTER 3

# METHODS FOR IDENTIFYING COMPLEX AND DIFFICULT WORDS IN READING

## *Theoretical, Empirical, Computational*

**Abstract** This chapter provides a short overview of three methods used in the literature to identify complex and difficult words in reading. First, the chapter addresses how theoretical conjectures found in readability assessment and lexical simplification determine complexity. A case study on a parallel corpus of simplified French texts recently published by Gala, Tack, Javourey-Drevet, François, and Ziegler (2020) will be discussed in particular. Next, the chapter focuses on how difficulty can be identified from empirical evidence, which covers indirect/implicit and direct/explicit measures. Finally, the chapter concludes with two computational methods: probabilistic language models and statistical machine learning systems.

So far, the dissertation has addressed the general scientific and methodological research scope. The preceding chapter (Chapter 2) gave a systematic review of relevant studies examined from a measurement and prediction perspective. The chapter mainly dealt with incidental vocabulary learning as these studies were the most represented in the collection of publications.

The current chapter addresses a topic that has not been described in much detail: lexical difficulty. The chapter zooms in on a subset of studies reviewed in Chapter 2 and focuses on describing various methods for identifying lexical difficulty in L2 reading. At the same time, the chapter's focus will also be broadened. Instead of focusing solely on empirical and statistical methods, other methods will also be considered.

However, several comments should be made. Firstly, the literature overview does not aim to be exhaustive. Instead, the chapter delineates several methods of interest to this doctoral study. Secondly, it was not easy to determine whether a study investigated the construct of difficulty. Some studies made explicit references to difficulty, whereas other studies mentioned the term either peripherally or not at all. Therefore, besides difficult and complex words, relevant terminology also included unknown words (existing or pseudowords) and other known effects (e.g., the N400 effect). Finally, for completeness purposes, the chapter will also include some studies conducted on L1 readers only.

The chapter is structured into three sections, each focusing on a different method. Section 3.1 describes the identification of difficulty with theoretical assumptions. Section 3.2 describes the identification of difficulty from empirical evidence. These empirical measures cover discrete (e.g., errors, binary judgment tasks) and continuous (e.g., word processing time) outcomes and indirect and direct measurements. Finally, Section 3.3 will describe the identification of difficulty with computational indices derived from probabilistic language models or statistical machine learning systems.

## 3.1 THEORETICAL METHODS

The first series of methods relate to theoretical conjectures about lexical complexity in reading materials, which are relevant for two related tasks: readability assessment and text simplification. Both readability formulae and simplification systems have used various thresholds to identify a set of complex words in a text. Furthermore, when experts are manually simplifying a text, they also conjecturing about the difficulty of words in this text. These thresholds and expert judgments will be described further in respectively Sections 3.1.1 and 3.1.2.

### 3.1.1 *Complexity Thresholds*

Readability formulae determine the overall difficulty of a text by surface-level characteristics, such as the average sentence length for all sentences $S$,

which quantifies the text's syntactic complexity; and the set of complex words $W_{\text{complex}}$ among all words $W$, which quantifies the lexical complexity of the text. To determine the latter, the most straightforward approach is to set a threshold at which a word should be considered complex. Two heuristics have been commonly used to establish this threshold: word length and frequency.[19]

*Word Length*

In some readability formulae[20], such as Gunning's (1952) index (3.1) and Mc Laughlin's (1969) SMOG formula (3.2), the set of complex words $W_{\text{complex}}$ are established with a polysyllable threshold, which considers a word as complex if it counts three or more syllables.

$$\text{Gunning Fog Index} = 0.4 \left( \frac{W}{S} + \frac{W_{\text{complex}}}{W} \cdot 100 \right) \tag{3.1}$$

$$\text{SMOG} = 3 + \sqrt{W_{\text{complex}}} \tag{3.2}$$

A similar approach has also been adopted in the domain of automatic simplification. For instance, Paetzold and Specia (2016a) used a length-based threshold as one of several baselines in the complex word identification (CWI) shared task. However, this baseline achieved a low accuracy (33%) on lexical complexity judged by non-native English speakers. Similarly, Paetzold and Specia (2016c) observed that word length and number of syllables was not a significant predictor of complexity on this data. In contrast, a significant predictor was the word's likelihood of occurrence in the target language.

*Word Frequency*

There are two ways in which frequency is used to determine complexity. On the one hand, a word can be considered complex if its frequency is beyond a pre-defined threshold. Paetzold and Specia (2016a) and Shardlow (2013a)

19 The reason why both are considered fundamental indicators of complexity stems from insights into the reading process. For instance, in eye-tracking research, it is generally known that word recognition time is determined by word length and frequency effects (e.g., see Dirix et al., 2019; Kuperman et al., 2013).

20 Flesch's (1948) and Coleman and Liau's (1975) formulae also incorporate a measure of word length. However, instead of identifying a set of complex words based on length, these formulae compute the total number of syllables and the average number of letters per word.

adopted such a threshold-based approach. They ranked words based on frequency and established a threshold that maximized the classification accuracy between complex and simple words. These frequencies were extracted from different corpora, including *Subtlex* (Shardlow, 2013a), *Wikipedia*, and *Simple Wikipedia* (Paetzold & Specia, 2016a). However, Shardlow (2013a) found that this frequency-based threshold was as accurate as considering all words as complex (i.e., the naïve 'simplify everything' baseline).

On the other hand, the word's absence from a most frequent and familiar vocabulary is also used as a heuristic of complexity. This lexicon-based approach is adopted in Dale and Chall (1948) (3.3): the percentage of complex words – or "vocabulary load" as Dale and Chall defined it – is established with a list of approximately 3,000 words judged as familiar by elementary school children.

$$\text{New Dale-Chall} = 0.1579 \left( \frac{W_{\text{complex}}}{W} \cdot 100 \right) + 0.0496 \cdot \frac{W}{S} \tag{3.3}$$

In addition to the New Dale-Chall list, several papers resorted to other well-known word lists, such as Ogden's (1930) Basic English, Dolch's (1936) list, and the Academic Word List (Coxhead, 2000), to identify complex words for non-native English speakers (see Brooke et al., 2016; Mukherjee et al., 2016; Paetzold & Specia, 2016a; Ronzano et al., 2016).

A similar threshold can be established with a graded vocabulary list. Such a list indicates how frequently a word occurs in reading materials at specific grade levels.[21] Tack et al. (2016a) defined a simple threshold (3.4) to determine the complexity of a word for specific L2 proficiency levels.

$$\text{complex} = \begin{cases} 1, & \text{if } w \notin V \vee L_w > L_{\text{learner}} \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

The threshold determined the complexity of a word $w$ with a two-step verification. First, like Dale and Chall's (1948) method, a word was considered complex if it was absent from the lexicon $V$, that is, if the word never occurred in reading materials intended specifically for language learners. Next, if the

21 For a more detailed analysis on the use of a graded lexicon for obtaining a theoretical measure of lexical difficulty, see also Chapter 4.

word occurred in $V$ at grade levels $L$ higher than the learner's proficiency level, the word was also considered complex.

### 3.1.2 *Expert Simplifications*

Theoretical insights into lexical complexity can also be gained from a corpus of simplified language. In lexical simplification studies, two types of corpora have been used: (a) comparable corpora of texts written in common or simple language and (b) parallel corpora of authentic texts aligned with manually simplified translations.

Examples of comparable corpora include the standard Wikipedia compared with either Simple Wikipedia (see for instance Woodsend and Lapata, 2011) or Vikidia (i.e., an encyclopedia written for children; see for instance Brouwers et al., 2014). A measure of complexity can be derived from these corpora by comparing an entry in the English Wikipedia with the same entry written in Simple English. However, since these wiki articles are not aligned, an extra text processing step needs to be performed in order to realign the writings at the syntactic or lexical levels. As noted by Xu et al. (2015), there are important practical drawbacks to aligning the original and simplified wiki articles.

Another solution is to derive a measure of complexity from manual simplifications attested in Simple Wikipedia (Shardlow, 2013b; Woodsend & Lapata, 2011). For example, Shardlow established a first gold standard for complex word identification based on Simple Wikipedia page revisions (see Figure 3.1). The complete inventory of page revisions was searched to gather single-sentence edit histories, that is, sentences where only one word was edited with the purpose of simplification.

Nevertheless, Amancio and Specia (2014) and Xu et al. (2015) called into question the use of Simple English Wikipedia to harvest evidence on simplified language. In their opinion article, Xu et al. highlighted the lack of editorial rigor in Simple English Wikipedia, which was in stark contrast with a professional, propietary, and parallel (aligned) corpus of simplified news articles such as Newsela. They showed that the supposedly simpler version contained potentially more intricate passages than the original version. For this reason,

**Figure 3.1**

*Illustration of a Simple English Wikipedia Edit History from the CW Corpus*

Readability tests give a prediction as to how _____ readers will find a particular text.
**original (complex word):** difficult
**edit (simple word):** hard

*Note.* This illustration was adapted from Shardlow (2013b, p. 74).

the use of simplified edit histories as a ground truth for the identification of complex words can be disputed.

*A Case Study: The Alector Corpus*

To further reflect on the use of more rigorously aligned corpora with expert simplifications, consider the following case study conducted on the Alector corpus (Gala et al., 2020). The Alector corpus is a parallel corpus of 79 French texts that were simplified specifically for poor-reading and dyslexic children. The manual simplifications were done by a panel of researchers specialized in cognitive psychology, psycholinguistics, speech therapy, and the educational sciences and experienced in reading difficulties caused by dyslexia. The simplifications were performed on multiple levels: lexical, morphological, syntactic, and discursive. The aims of the case study were to look specifically into the simplifications conducted at the lexical level and to validate the conjectures of lexical complexity made by the panel of experts.

The main assumptions were that the manual simplifications were necessary and that the lexical substitutions were adequate. To verify these assumptions, the study focused on aligning the original (complex) and simplified (non-complex) words with empirical evidence of reading difficulties. This evidence was collected from transcripts of read-aloud interventions. For each word in the authentic and simplified texts, the probability of the word being misread by a child suffering from dyslexia, as identified in the transcripts. With this error probability as a measure of lexical difficulty, it was verified whether the manual simplifications had a significant alleviating effect. The first hypothesis was that

This section was previously published in Gala et al. (2020).

the manual simplifications were necessary if there were more reading errors on the words in the authentic texts targeted for simplification (i.e., conjectured to be difficult) than on those not targeted for simplification. Moreover, the second hypothesis was that the simplifications were accurate if there were less errors on the substitute words than on the original words.

ALIGNMENT OF READING ERRORS From the parallel corpus, a sample of 20 texts (i.e., 10 authentic texts with their simplified version) were used during speech therapy interventions with reading-impaired children. A sample of 21 French-speaking children aged between 9 and 12 participated in the experiment. The participants were attending mainstream schools and had a reading delay of two and a half years on average. During the experiment, the participants were asked to read aloud 10 different texts (i.e., five original and five simplified texts), which were drawn from both the literary and scientific genres and which were presented on a digital tablet. The reading task was self-paced: the texts were read sentence per sentence and the children had to click to move on to the next sentence. The read-aloud data were recorded by students majoring in speech therapy (Nandiegou & Reboul, 2018), who manually transcribed the childrens' vocalizations and reading errors using ad-hoc guidelines.

Because the data had not been encoded with specialized transcription software, all 210 transcripts (i.e., 10 transcripts per participant) had to be realigned with their original text. The alignment was done at the level of the word, associating each word as it was read aloud with the word as it appeared in the text. To this end, a modified version of the Needleman and Wunsch (1970) sequence alignment algorithm was used. The simplicity of the algorithm did not seem problematic given that the transcripts did not contain any major syntactic modifications. The modified version of the algorithm aligned two words by constructing a similarity matrix for each pair of sentences in the original text ($o$) and the read-aloud transcript ($r$), represented as a sequence of tokens. The similarity score $s$ between two tokens $w_o$ and $w_r$ was computed by the integration of the edit distance and the length of a word (3.5). The modified algorithm aligned two tokens when they had similar forms (e.g., *jardin* and *jadi*). Once the alignment was completed for all participants separately, all these per-participant alignments were aggregated so as to obtain a list of all

**Table 3.1**

*Post-Hoc Comparisons of the Probability of Making Reading Errors on Words Targeted or Not Targeted for Simplification*

| | | | DSCF Comparisons ($W$) | |
|---|---|---|---|---|
| Group | $N$ | $Mdn$ | Substituted | Deleted |
| Substituted | 192 | 0.18 | | |
| Deleted | 80 | 0.09 | $6.16^{***}$ | |
| Not targeted | 1,972 | 0.00 | $19.6^{***}$ | $5.30^{***}$ |

*Note.* This table was taken from Gala et al. (2020, p. 1359).
$^{***}$ $p < .001$

different variants read aloud for each word attested in the original text. An illustration of two simplified sentences in which each word is aligned with reading errors is given in Figures 3.2 and 3.3.

$$s = \begin{cases} +1, & \text{if } w_o = w_r \\ -\left[1 + \delta_{\text{Levenshtein}}(w_o, w_r)\right], & \text{if } w_o \neq w_r \\ -\left(1 + |w|\right), & \text{as gap penalty} \end{cases} \tag{3.5}$$

READING ERRORS ON LEXICAL SIMPLIFICATIONS To validate the top-down conjectures of lexical difficulty, the study first focused on the words in the authentic texts and divided them into three different categories: words that were targeted for simplification, either by substitution or deletion, and words that were not targeted for simplification (Table 3.1). A Kruskal-Wallis test showed that the probability of misreading words belonging to either category was significantly different, $\chi^2(2) = 199$, $p < .001$, $\epsilon^2 = 0.089$. Post hoc Dwass-Steel-Critchlow-Fligner (DSCF) comparisons were performed with a Wilcoxon ($W$) rank sum test and $r = \frac{Z}{\sqrt{N}}$ as the effect size statistic. The post hoc analysis showed that the largest difference was attested between words that were substituted ($Mdn = 0.18$) and words that were not targeted for simplification ($Mdn = 0.00$), $W = 19.6$, $p < .001$, $r = .30$. A significant difference was also observed between words that were deleted ($Mdn = 0.090$) and words that were not targeted but the effect was much smaller, $W = 5.30$,

**Figure 3.2**

*Alignment of Reading Errors on a Lexically Simplified Text with a Substitution*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original Text<br>[10 Readings] | Voilà | que | tu | t' | **agenouilles** | | | devant | ce | tas | de | neige. |
| Misreadings | | *qui* | | | ***agenou***<br>***agenoui***<br>***agenouiller***<br>***agenouillies***<br>***angenouillies***<br>***aquenoulés*** | | | | *cette*<br>*ça* | | | |
| Error Count | 0 | 1 | 0 | 0 | **7** | | | 0 | 2 | 0 | 0 | 0 |
| Error Probability | 0.0 | 0.1 | 0.0 | 0.0 | **0.7** | | | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 |
| Simplified Text<br>[11 Readings] | Voilà | que | tu | te | **mets** | **à** | **genoux** | devant | ce | tas | de | neige. |
| Misreadings | | | | *me* | | | | | *ces*<br>*te* | | | |
| Error Count | 0 | 0 | 0 | 1 | **0** | **0** | **0** | 0 | 4 | 0 | 0 | 0 |
| Error Probability | 0.0 | 0.0 | 0.0 | 0.09 | **0.0** | **0.0** | **0.0** | 0.0 | 0.36 | 0.0 | 0.0 | 0.0 |

*Note.* This illustration was adapted from Gala, Tack, Javourey-Drevet, François, and Ziegler (2020, p. 1358).

**Figure 3.3**

*Alignment of Reading Errors on a Lexically Simplified Text with a Deletion*

| Original Text [11 Readings] | Il | y | avait | **jadis** | en | Irlande | un | homme | du | nom | de | Jack. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Misreadings | | | *avant*<br>*était* | ***jadi***<br>***jardin***<br>***jardis*** | | *Arlande* | | | | | | |
| Error Count | 0 | 0 | 2 | **7** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Error Probability | 0.0 | 0.0 | 0.18 | **0.64** | 0.0 | 0.09 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Simplified Text [10 Readings] | Il | y | avait | | en | Irlande | un | homme | du | nom | de | Jack. |
| Misreadings | | ∅ | | | *un* | *rilande* | | | *de* | | | *Jean* |
| Error Count | 0 | 1 | 0 | | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 1 |
| Error Probability | 0.0 | 0.1 | 0.0 | | 0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.1 |

*Note.* This illustration was adapted from Gala, Tack, Javourey-Drevet, François, and Ziegler (2020, p. 1359).

$p < .001$, $r = .083$. The study therefore concluded that all words targeted for simplification were well-chosen. However, the need for simplification was not equally strong, as was shown by the significant difference between the number of misreadings on words that were substituted and deleted, $W = 6.16$, $p < .001$, $r = .26$. Although removing superfluous words was also necessary to reduce reading difficulties, substituting core but difficult content words seemed even more crucial.

Next, the study focused on the substituted words and examined the effect of the lexical simplifications. A pairwise Friedman rank sum test showed that there were significantly fewer misreadings after simplification ($Mdn = 0.090$) than on the word in the original text ($Mdn = 0.18$), $\chi^2_F(1) = 40.6$, $p < .001$. Moreover, the expected decrease in number of errors made before and after simplification was indeed large (Kendall's $W = .527$). Consequently, it was concluded that the words targeted for simplification were not only well-chosen, but also substituted with substantially easier alternatives.

In sum, the case study showed two things. On the one hand, the results suggested that the conjectures made by experts were a reliable measure of lexical difficulty. As such, the results supported the preference for relying on a theoretical ground truth of lexical difficulty based on more rigorously conducted manual simplifications. On the other hand, the case study also showed the need to compare these conjectures of lexical difficulty with empirical evidence of processing difficulties in reading. A limitation was, however, that the study only tested one type of empirical measure of lexical processing difficulties in reading, namely the errors made by poor-reading dyslexic readers in a vocal reading task.

## 3.2 EMPIRICAL METHODS

The previous section concluded that, although it is possible to determine lexical complexity based on simple heuristics, it is essential to contrast these theoretical conjectures with empirical data from a target reader population. Therefore, the current section presents several additional empirical measures of lexical difficulty. Because the systematic review in Chapter 2 only briefly alluded to these empirical methods, this section provides a more focused

overview. Without attempting to be exhaustive, the section presents the following measures: reading times, eye movements, brain responses, vocal reading errors, think-aloud verbalizations, and self-assessments.

The common ground between all these empirical measures is – at least, in the current domain of investigation – that they are the outcome (or result) of a measurement assumed to have been caused by a psychological construct (or attribute) observed in a person. They differ, however, in the criteria that characterize how these different aspects relate to each other. To compare and contrast these measures, we need to know which classification criteria currently exist and how to use them.

CLASSIFICATION CRITERIA An important but often confusing distinction is the contrast between implicit/explicit and indirect/direct measures. De Houwer and Moors (2010) define a measure as *implicit* if the psychological process that causes the outcome of a measurement is one of automaticity (i.e., lack of awareness and intention). Conversely, one could define a measure as *explicit* if it is the outcome of a measurement caused by a deliberate and conscious cognitive process. By this definition, the reading aloud of a text is an explicit measure because the outcome (i.e., the vocalization) is consciously produced. The brain signal, in contrast, is an implicit measure because the outcome (i.e., the amplitude and time frame) is automatically produced.

On the other hand, the distinction between direct and indirect measures is not concerned with the automaticity of the cognitive process brought about in a person. Instead, it is concerned with the directness of the interpretative link between the construct and the outcome.[22] For example, the vocalization of a reading text is a direct measure of word recognition difficulty because there is a direct link between the outcome (i.e., a misreading) and the construct (i.e., word recognition). The brain signal, in contrast, is an indirect measure because it requires making inferences about the link between the outcome (i.e., the amplitude and time frame) and the construct (i.e., word recognition).

Table 3.2 presents the measures discussed in this section and classified based on the previously defined criteria. Besides the implicit/explicit and indirec-

22 It should be noted that De Houwer and Moors (2010) define direct measures more restrictively. A direct measure must to satisfy the self-assessment criterion (i.e., a person must give a self-assessment of the construct being measured).

**Table 3.2**

*Classification of Empirical Measures of Lexical Difficulty in Reading*

| | | | Response | |
|---|---|---|---|---|
| Measurement | Causation | Link | System | Content |
| Reading time | implicit | indirect | behavioral | asymbolic |
| Eye fixation | implicit | indirect | behavioral | asymbolic |
| Brain signal | implicit | indirect | neurological | asymbolic |
| Vocalization[a] | explicit | direct | behavioral | symbolic |
| Verbalization[b] | explicit | direct | behavioral | symbolic |
| Self-assessment | explicit | direct | behavioral | symbolic |

[a] Vocalization in read-aloud protocols.
[b] Verbalization in think-aloud protocols.

t/direct paradigms, measures in cognitive research can be further classified based on other characteristics such as the nature of the response (i.e., behavioral, neurological, and physiological) and whether the response is symbolic or a-symbolic. This classification also enables the overview to be structured into two sets of measures that share the same characteristics. On the one hand, Section 3.2.1 describes some implicit and indirect measures of lexical difficulty observed from reading times, eye movements, and brain signals. On the other hand, Section 3.2.2 looks at some explicit and direct measures of lexical difficulty observed from vocal reading tasks, think-aloud procedures, and self-assessments of difficulty.

### 3.2.1 *Implicit and Indirect Measures*

The first set of empirical measures include reading times, eye movements, and brain signals. As highlighted above, these measures are implicit because they lack conscious introspection, making it difficult for the individual to control or modify the response. These measures are indirect because there is no direct link between the outcome and the psychological construct they aim to gauge.

The studies discussed below address various psychological constructs. On a more general level, they aim to gauge the overall cognitive effort (or cognitive

load) exerted to process a word while reading. This cognitive load comprises the word decoding and language comprehension components of the simple view of reading presented in Chapter 1. On the one hand, cognitive load originates from difficulties in lexical access, which is the effort exerted to recognize the word form (orthography and morphology). On the other hand, cognitive load is also the cause of difficulties related to word-to-(con)text integration, which is the effort exerted to interpret the meaning from the surrounding context. In what follows, we will examine how these two attributes have been measured from reading time, eye movements, and brain responses.

*Reading Time*

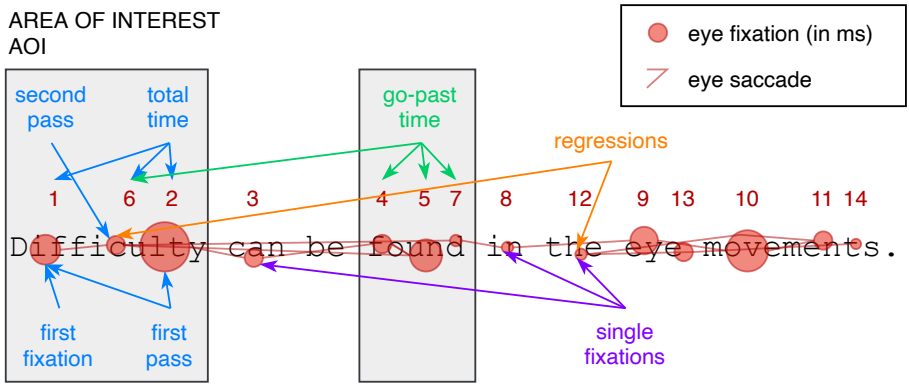
In the literature reviewed in Chapter 2, very few studies looked into lexical competence operationalized as reading time. The analyses briefly referred to a recent study by Kim et al. (2018) that examined different predictors of word processing time measured with a self-paced word-by-word reading task. A self-paced word-by-word reading task presents the text in its sequential order but only displays one word at a time. Reading time is then measured from the moment the word appears on the screen to the moment where the reader clicks to move on to the next word.[23] In particular, Kim et al. found a significant decrease in reading time when words appeared later in the text. They interpreted this effect positively: words encountered later on were processed faster because learners had already understood most of the text's content by then.

Nevertheless, because reading time is an indirect measure of lexical difficulty, the exact interpretation of shorter and longer times can be double-edged. Shorter reading times may indicate that fewer cognitive resources are allocated, but the use of fewer cognitive resources may be the result of better textual understanding (i.e., positive interpretation; cf. *supra*), or symptomatic of the reader's attention span lessening in a cognitively loaded task (i.e., negative interpretation). Similarly, longer reading times may evidence that more cognitive resources are being allocated, but the use of more cognitive resources may be the result of word processing and comprehension difficulties (i.e., negative

23 Because of this, it may be debatable whether reading time can always be considered an implicit measure. It is not entirely sure to what extent the clicking performed by the subject is either a conscious or an (increasingly) automatized process.

**Figure 3.4**

*Artificial Illustration of Eye-Movement Measures and Metrics*




interpretation)[24] or it may indicate that the reader allocates more resources to learn a new word successfully (i.e., positive outcome). Consequently, it is not always clear which inferences regarding lexical difficulty can be drawn from reading time measures.

*Eye Movements*

Like word reading times, eye movements also indicate the degree of cognitive load while reading. Figure 3.4 illustrates the most fundamental eye-movement measures targeted in L2 research[25], including the duration, amount, and types of eye fixations and regressions on a particular word or area of interest (AOI).

In addition to the *total time* spent reading an AOI, other eye-movement measures indicate cognitive load and its multiple facets. Regarding lexical access, *first-fixation duration* and *first-pass time* (also called *first-dwell time* or *gaze duration*) capture difficulties in word form recognition. On the one hand, the duration of the first fixation on the word indicates the degree of automaticity in accessing the word form (e.g., see Dolgunsöz & Sariçoban, 2016; Elgort et al., 2018). The shortest first fixations are on short and frequent words, which tend to be fixated only once (i.e., *single fixation*). In other words, it is much more likely that commonplace function words are recognized with

24 Consider, for instance, cognitive load measured as word processing time, which is predicted to be higher on words displaying a greater entropy (Frank, 2013).

25 For reference material on eye tracking studies in L2 research, see Godfroid (2019).

automaticity and ease. On the other hand, Elgort et al. (2018) and Godfroid et al. (2013) found that L2 learners displayed a higher gaze duration on unknown words (low-frequency words or pseudowords). After about eighth encounters, however, Elgort et al. observed that, after about eight encounters, the duration of eye fixations on target words (first fixation, gaze duration, and total time) was no longer significantly different from high-frequency words (control words), indicating that contextual learning had occurred.

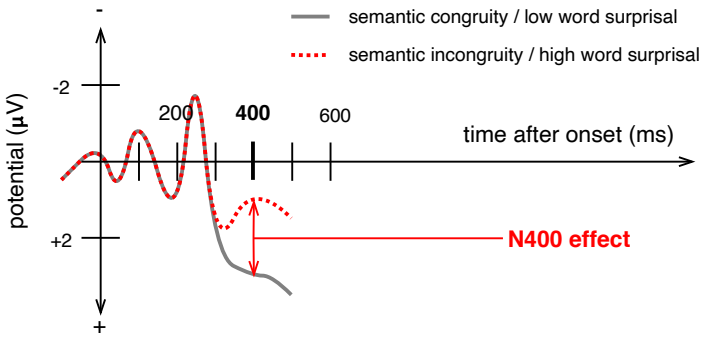
At the same time, learners differ significantly in how they fixate on new word forms while reading. Godfroid et al. (2013) noted a considerable between- and within-learners variance in the duration of eye fixations (beyond first-fixation duration) on pseudowords. In particular, the learner's proficiency level appears to be a distinguishing factor. Dolgunsöz and Sariçoban (2016) found that there was a significant interaction between proficiency level (intermediate learners, B1/B2 CEFR levels) and eye measures indicative of initial word recognition. First-pass time and total time significantly decreased in the case of more proficient learners.

Besides lexical access, eye-movement measures also indicate difficulties regarding word-to-context integration. In particular, when a reader regresses to a word, the duration of this *second pass* (as well as *go-past time*) is indicative of the effort exerted to interpret the meaning of the word from the surrounding contextual cues (Dolgunsöz, 2016; Elgort et al., 2018). A significantly longer second-pass time indicates that learners need to infer the meaning of a new (pseudo or unfamiliar) word from contextual cues (Dolgunsöz, 2016; Godfroid et al., 2013). Moreover, Furtner et al. (2011) found that the *number of regressions* made by L2 learners was significantly higher on common nouns than on other word categories. This observation is not surprising; the need to interpret meaning is inevitably higher for words carrying a semantic function.

Similar to the variability observed for lexical access, several studies also showed that learners differed considerably in second-pass time and other measures indicative of contextual inferencing (Dolgunsöz, 2016; Dolgunsöz & Sariçoban, 2016), even within the same context conditions (Godfroid et al., 2013). Dolgunsöz (2016) observed a significant decrease in second-pass time from low-intermediate (B1 level) and high-intermediate (B2 level) to advanced (C1 level) learners. What is more, the eye-movement measures also varied

**Figure 3.5**

*Artificial Illustration of the N400 Effect in EEG Brain Potentials*




between types of reading tasks. Second-pass time was more significantly different in a sentence reading task than in a natural reading task. In other words, the effort exerted to infer word meaning from context was more marked when learners read sentences in isolation.

*Brain Signals from Electroencephalography (EEG)*

A final implicit and indirect measure of word processing difficulties can be found in the brain's response while reading, measured from EEG procedures. A review of EEG studies by Vandenberghe et al. (2019) discussed the use of event-related potentials (ERPs) as sensitive measures in L2 vocabulary acquisition research.

Researchers assume that, among these ERPs, the N400 potential measures lexical processing difficulties, especially regarding word-to-context integration (Chen et al., 2017; Choi et al., 2014; Frishkoff et al., 2010; Rodríguez-Gómez et al., 2018). Figure 3.5 illustrates the N400 effect, characterized by a more substantial negative-going signal peaking around 400 ms after stimulus onset. There is long-standing evidence in cognitive neuroscience (see Kutas and Hillyard, 1980, 1984) that this effect indicates an incongruity in deriving word meaning from context because the word occurs unexpectedly (semantic violation). In particular, Nigam et al. (1992) found that the N400 applied to contexts where target word forms were substituted with either anomalous words or pictures, suggesting that this ERP was indicative of activity in a conceptual memory.

Similarly, the N400 effect also indicates an incongruity in deriving word meaning from context because the word is unknown (semantic novelty). For example, Chen et al. (2017) found that the mean N400 amplitude reflected the impact of different contextual constraints on learning new word meanings. The results showed that the N400 amplitudes on real or pseudo words embedded in high-constraint sentence contexts (i.e., which enabled the meaning to be congruently construed) were initially larger but then significantly decreased after 2–4 exposures, suggesting that contextual learning had occurred. By contrast, the N400 amplitude on pseudowords embedded in low-constraint sentence contexts (i.e., which *did not* enable the meaning to be congruently construed) did not significantly decrease after multiple exposures, suggesting that no contextual word learning had occurred. Moreover, Chen et al. also observed a significant interaction between the experimental condition, the learner's proficiency level, and the brain region.

In conclusion, both eye movements and brain signals can implicitly measure difficulties when learners process new words while reading, both in terms of lexical access and word-to-context integration. At the same time, the various eye-tracking and EEG studies also highlighted considerable variability in how these new words were processed, which originated not only from individual differences between learners (e.g., proficiency levels) but also from the nature of the reading task (e.g., sentence reading vs. natural reading).

### 3.2.2 *Explicit and Direct Measures*

Besides reading times, eye movements, and brain signals, lexical difficulty can also be measured from think-aloud protocols, vocal reading tasks, and self-assessments. As explained before, these measures are explicit because they require conscious introspection. Also, these measures are direct because the construct they aim to gauge can be directly inferred from the outcomes.

#### *Think-Aloud Protocols*

Different aspects of lexical difficulty can be measured explicitly by asking learners to verbalize their inner thoughts aloud while reading. In L2 vocabulary studies, researchers used think-aloud procedures to document how learners

notice unknown vocabulary and how they address these lexical difficulties. On the one hand, Bowles (2004) and Yanguas (2009) used think-aloud procedures to measure the reader's unfamiliarity with the target words (conjectured to be difficult based on pilot ratings) and evidence that hyperlinked glosses significantly increased the learners' noticing of these target words. On the other hand, Ender (2016) and Levine and Reves (1998) used think-aloud procedures to observe the learners' strategic behavior while reading. In particular, the studies looked at the degree of learner involvement, which ranged from ignoring the word, inferring meaning from context, and using a dictionary to check the meaning of a word.

However, there is some debate about the potential reactivity of these protocols. A meta-analysis by Bowles (2010) showed that thinking aloud while performing a specific task resulted in a positive or negative effect on learning:

> In other words, compared to participants completing the same tasks silently, participants who think aloud tend to perform only slightly better or slightly worse on post-tests. The results for time on task are more decisive, indicating across the board that thinking aloud increases time on task. Nevertheless, effect sizes for latency ranged from small ($d$ = .16) to very large ($d$ = 1.16), with the largest effects demonstrated when participants were required to think aloud while performing reading tasks. (p. 110)

Therefore, one should be mindful that think-aloud measures will inevitably impact the time spent on the reading task. Regarding potential reactivity, Godfroid and Spino (2015) did not observe a significant negative or positive effect for both eye-tracking and think-aloud measures on reading comprehension. For vocabulary learning, in contrast, the results indicated a small but positive impact of thinking aloud on vocabulary retention after reading. In consequence, there is something quite paradoxical about the use of think-aloud protocols. While the advantage is that they enable more direct measurement of lexical difficulty, it is probably precisely this very explicit and direct nature of the measurement (i.e., conscious introspection) that triggers the subjects to overcome these difficulties.

*Read-Aloud/Vocal Reading Tasks*

A method similar to think-aloud protocols is to measure possible errors on words in a vocal, reading-aloud task. Previously, a case study was described showing the use of misreadings to verify manual lexical simplifications (cf. *supra*; Section 3.1.2).

An important point not highlighted before is the considerable variability in the extent and types of reading errors. The two excerpts listed in Figures 3.2 and 3.3 on page 109 and on page 110 show that many misreadings were highly idiosyncratic, with only one or two participants mispronouncing the word. This variability can be seen more clearly in Figure 3.6. The figure illustrates the distribution of the probability $p$ to misread a word in the original and simplified texts. More than half of the words appearing in both versions were correctly read ($p = 0.0$). In contrast, only a tiny percentage of words were mispronounced by all poor-reading children ($p = 1.0$). These observations seem to be in line with the between-subjects variability previously observed for eye-movement measures (cf. *supra* and Bingel, Barrett, et al., 2018).

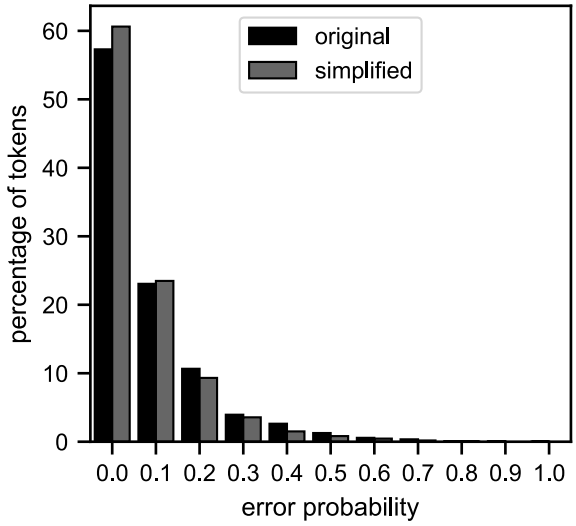
Furthermore, the examples in Figures 3.2 and 3.3 on page 109 and on page 110 suggest that not all errors are equally indicative of difficulty. When poor-reading children pronounce either non-existent words (e.g., *aquenoulés*) or existing words with radically different grammatical categories (e.g., *jardin* [noun] instead of *jadis* [adverb]), it is relatively straightforward to define these errors as word decoding difficulties. In contrast, when poor-reading children slightly change the word's form without radically changing the grammatical category (e.g., *qui* instead of *que*), it is disputable whether these errors indicate word decoding difficulties.

*Self-Assessments of Difficult Words*

A final empirical measure pertains to learners judging the difficulty of words while they are reading. Following De Houwer and Moors (2010) (cf. *supra*), this self-assessment measure is probably the most direct way to measure lexical difficulty among all measures reviewed in this chapter. Moreover, this self-assessment of lexical difficulty has been studied in both applied and computational linguistics.

**Figure 3.6**

*Distribution of the Probability to Misread a Word Token in the Sample of Poor-Reading and Dyslexic Children*




*Note.* This illustration was taken from Gala et al. (2020, p. 1358).

Chapter 2 saw that researchers interested in incidental vocabulary acquisition usually defined a sample of target unfamiliar words based on pilot ratings or pretests. These target words were used for statistical testing and were the same for all subjects. Conversely, another more recent but less prevalent approach focuses on words that a learner clicks on to access the gloss (Wang, 2016) or which learners themselves judged to be personally tricky. Ender (2016), for instance, asked learners to highlight all words they did not know while they were reading, which was "intended to yield the number of unknown words in the texts *on an individual basis* [emphasis added]" (p. 544). Ender used these personal target words to test the individual learner's strategies for dealing with lexical difficulty.

Chapter 2 also introduced several studies using self-assessment measures for complex word identification (CWI). The approach adopted in CWI papers shifted from using manual lexical simplifications as a gold standard for ma-

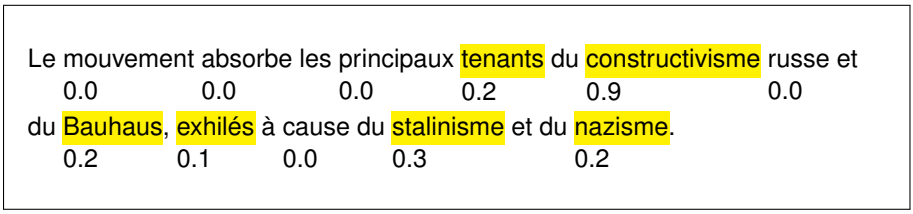
chine learning (see Section 3.1.2 and Shardlow, 2013a) to collecting judgments of lexical difficulty in a sample of target L1 or L2 readers. These self-assessment measures were collected with either a binary judgment task (Paetzold & Specia, 2016a; Yimam et al., 2018) or, more recently, with Likert-scale rankings (Shardlow et al., 2020). In the binary judgment task, participants were asked to mark or highlight the words of which they did not recall the meaning. In the Likert-scale ranking task, participants assessed personal word familiarity and form/meaning recognition: 1 = *very easy* (*very familiar*), 2 = *easy* (*aware of the meaning*), 3 = *neutral* (*neither difficult nor easy*), 4 = *difficult* (*unclear meaning but could be inferred from the sentence*), 5 = *very difficult* (*never seen before or very unclear*). Based on this data, several shared tasks provided a benchmark for developing statistical machine learning systems, which will be discussed in Section 3.3.

An important note to make here is that, although these recent studies in applied and computational linguistics seem to have used very similar self-assessment measures, they approached the data collection procedure in a very different way. Whereas studies in applied linguistics focused on self-assessment of personal lexical difficulty for learners reading the same texts (Ender, 2016), studies in computational linguistics instead used aggregated measures to optimize the data collection procedure (i.e., where different participants annotated separate parts of the data).

During the first shared task, Paetzold and Specia (2016a) divided a sample of 9,200 sentence contexts into two data sets: (a) a set for training a statistical machine learning system and (b) another for evaluating system performance on unseen data. The researchers then distributed the sentences among 400 non-native speakers of English. On the one hand, the training set consisted of 9000 sentences, each seen by one participant only. Therefore, the training data included an aggregate measure of lexical difficulty for EFL learners at large. On the other hand, the test set consisted of 200 sentences, each judged by 20 participants simultaneously. The test data was used to examine whether a general model trained on aggregated assessments could accurately generalize its predictions on personal difficulty judgments. This evaluation was not a trivial matter given that there was a low agreement and high variability between individual raters (Krippendorff's $\alpha$ = 0.24, *SD* = 0.1).

**Figure 3.7**

*Example of Highlighted Difficult Words in the CWI Shared Task 2018*




*Note.* This example is taken from data published in Yimam et al. (2018).

During the second shared task, Yimam et al. (2018) expanded on this methodology in multiple ways. First, they added a probabilistic measure of difficulty (Figure 3.7) computed from data gathered via Amazon Mechanical Turk. A similar self-assessment task was administered to a group of native and non-native workers who read a series of short paragraphs (ranging from 5-10 sentences). For each paragraph, a sample of 10 to 20 participants highlighted the words or word sequences that they assessed to be difficult. The probabilistic measure then corresponded to the probability that the word was found difficult by all participants reading the paragraph. The researchers used the same procedure to collect a new gold standard for other languages, including Spanish, German, and French. However, a limitation of their study was that it did not further investigate the assessment of personal lexical difficulty because of the aggregated nature of the data.

## 3.3 COMPUTATIONAL METHODS

The previously presented methods identified lexical difficulty from theoretical, a priori knowledge or empirical, a posteriori knowledge. As it happens, both these theoretical and empirical measures (in particular, manual simplifications and self-assessments) have served as ground truth for computing a predictive model of lexical difficulty. This section presents the use of such computational methods. More specifically, the section addresses probabilistic language models (Section 3.3.1) and statistical machine learning systems (Section 3.3.2).

### 3.3.1 *Probabilistic Language Models*

A first way to computationally model lexical difficulty is to rely on language models, which estimate how likely lexical units and syntactic structures occur in the target language. The hypothesis is that linguistic forms and structures which exhibit a lower probability will be subject to a higher cognitive load. This hypothesis is, of course, strongly related to the frequency effects discussed in Sections 3.1.1 and 3.2.1 and which were indicative of readability and speed of word processing.

The simplest language model is the *n*-gram model, which estimates the probability of observing a word from the $n - 1$ preceding words. Because the model relies on a restricted context window, often ranging from zero (i.e., unigram) to four (i.e., pentagram) preceding words, these probabilities give a relatively shallow representation of language use. A less shallow and more hierarchical model is a syntactic parser based on a probabilistic context-free grammar probabilistic context-free grammar (PCFG). For a sentence never seen before, a probabilistic parser can estimate the probability of observing this structure as a whole. A practical drawback is, however, that these grammars need to be pre-trained on a large corpus with manual annotations of syntactic structure, which is a very costly process. A less shallow, less costly, and more recent approach to language modeling is to learn probability estimates from recurrent neural networks (RNNs), which aim to capture long-term syntactic and semantic dependencies. A more detailed introduction into these various approaches is, unfortunately, beyond the scope of this chapter. For a more comprehensive introduction into probabilistic language modeling, the unfamiliar reader may wish to read reference books in computational linguistics (e.g., see Jurafsky & Martin, 2009; Manning & Schütze, 1999).

With probability estimates obtained from a given language model, one can subsequently compute a number of information-theoretic complexity metrics. A central information-theoretic complexity metric is concerned with the notion of *word surprisal*. In information theory, the surprisal $S$ of an outcome $y$ of a random variable $Y$ (3.6) is defined as "the logarithm[26] of the reciprocal of

26 Most often, the base of the logarithm is set to two, which means that the information value of a random event is quantified in the number of bits. However, it should be noted that most

a probability [estimate of $y$]" (Hale, 2016, p. 400). Put differently, the less probable a random event, the higher its surprisal.

$$S(Y = y) = \log_{\text{base}} \left[ \frac{1}{P(y)} \right] \tag{3.6}$$

In a seminal study, Hale (2001) put forth a computational model of the cognitive effort in processing unexpected syntactic structures. In particular, the study focused on the computational processing cost of reading garden-path sentences as estimated with a probabilistic Early parser (Figure 3.8). These garden-path sentences "happen at points where the parser can disconfirm alternatives that together comprise a great amount of probability" (p. 5). In other words, the cognitive cost comes from having to first discard the most likely syntactic analysis construed before a particular word (e.g., "fell" in Figure 3.8), and then revise this very likely prefix probability (e.g., $P$ ["The horse raced past the barn"]) into a much less probable syntactic analysis (e.g., $P$ ["The horse raced past the barn fell"]). Notably, several eye-tracking, EEG, and fMRI studies have given empirical support for this computational cost to predict cognitive processing difficulties (see the literature overview in Hale, 2016).

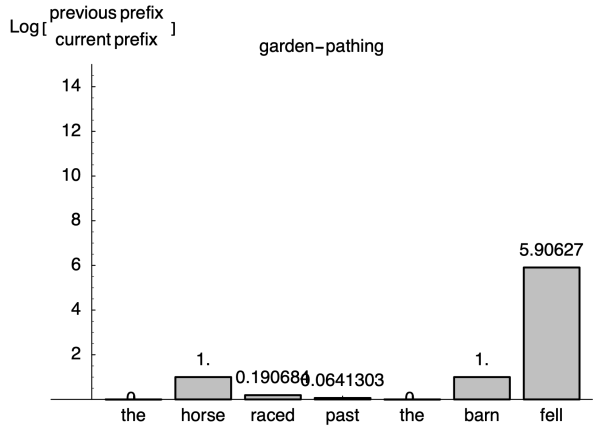
An exciting set of studies investigated whether word surprisal could predict lexical processing difficulties (in the L1) measured from EEG[27] potentials. Frank et al. (2013, 2015) studied the inclusion of surprisal estimates into several linear mixed-effects regression models, one for each targeted brain potential (P200; N400, EPNP, PNP; ELAN, LAN, P600). The surprisal estimates were computed from three different language models: an $n$-gram model ($n \in \{2, 3, 4\}$), a PCFG model, and an RNN. The researchers drew two important conclusions. Firstly, word surprisal significantly predicted the N400 potential, whereas no significant effect was found for the other ERPs. This result was in line with the typical observation mentioned earlier, namely that the N400 effect occurs in contexts with a low cloze probability (Chen et al., 2017; Parviz et al., 2011; Smith & Levy, 2011). Secondly, word surprisal estimated from either

language modeling tools (e.g., see Heafield et al., 2013) already provide log-likelihood estimates, which use the common logarithm (base 10).

27 For a discussion of these EEG measures and the N400 effect in particular, see Section 3.2.1 above.

**Figure 3.8**

*Difficulties in Processing Garden-Path Sentences Predicted from Word Surprisal Values Computed with an Early Parser on a Simple PCFG*




*Note.* This illustration was taken from Hale (2001, p. 5).

an *n*-gram or RNN model significantly outperformed parser-based surprisal. The RNN outperformed the *n*-gram model on content words only, although the difference between both models was not clear-cut when considering all morphosyntactic categories. Consequently, the studies provided evidence that "a non-hierarchical, RNN-like architecture forms a more plausible cognitive model of language processing than systems that are based on hierarchical syntactic structure" (Frank et al., 2015, p. 9).

### 3.3.2 *Statistical Machine Learning*

The preceding section introduced word surprisal as a computational measure of the cognitive cost of lexical processing during reading. Specifically, surprisal statistically explained several implicit measures of processing difficulties previously discussed in Section 3.2.1. In addition to these explanatory models, researchers have also resorted to other statistical modeling techniques.[28] This

28 For a technical discussion on the (sometimes) thin line between "causal explanation" and "empirical prediction", see Shmueli (2010).

section describes machine learning systems developed in several papers related to the CWI shared tasks (Paetzold & Specia, 2016a; Yimam et al., 2018).[29] In particular, the overview compares systems built on feature engineering, on the one hand, and deep learning, on the other hand.[30]

*Feature Engineering*

Most systems that competed in the CWI shared tasks used ensemble-based classification/regression algorithms (Table 3.3). This reliance on ensemble machine learning is likely due to the well-known observation that ensemble methods (random forests, voting classifiers, etc.) tend to outperform classical machine learning algorithms (decision trees, regression-based models, etc.) (Dietterich, 2000). Unsurprisingly, systems based on a voting classifier and integrating an extensive feature set achieved the best performance on both shared tasks (Gooding & Kochmar, 2018; Paetzold & Specia, 2016b).

The features of lexical complexity most commonly computed were word length and frequency (Table 3.4). Again, this is not a surprising observation seeing that readability and cognitive processing interact with length and frequency effects (cf. *supra*, Sections 3.1 and 3.2). Nevertheless, it seems that word frequency, in particular, was a salient predictor of lexical difficulty for non-natives. Indeed, a parsimoniously engineered system using only fine-tuned frequency-based measures achieved top-tier performance on the first shared task (Wróbel, 2016).

At first sight, it may seem that the frequent use of semantic features, as shown in Table 3.4, relates to measuring difficulties in word-to-context integration. However, it is essential to note that these features did not correspond to contextualized, semantically disambiguated indices. They measured either (a) the total number of meanings associated with the word or (b) the average or aggregate number of synonyms, hypernyms, and hyponyms for all possible

29 It should be noted that a new lexical complexity prediction (LCP) shared task based on data collected by Shardlow et al. (2020) is to be held at SemEval 2021. However, the papers had not been published yet at the time of writing. For more information, see the shared task website https://sites.google.com/view/lcpsharedtask2021.

30 The overview focuses on linguistic features and learning algorithms because they are the most relevant to the current dissertation. Methodologically speaking, there is more to be said about the shared tasks (e.g., performance evaluation metrics), but this discussion did not seem relevant to the scope of this chapter.

**Table 3.3**
*Frequently Used Machine Learning Algorithms in the CWI Tasks*

| Systems | # | Papers |
|---|---|---|
| ENSEMBLE LEARNING<br>Decision Trees (Random Trees, Random Forests, Extra Trees, Decision Stumps); Boosting (Gradient Boosting, AdaBoost, XGBoost); Bagging; Voting | 19 | AbuRa'ed and Saggion (2018), Alfter and Pilán (2018), Aroyehun et al. (2018), Bingel and Bjerva (2018), Brooke et al. (2016), Choubey and Pateria (2016), Davoodi and Kosseim (2016), Gooding and Kochmar (2018), Hartmann and dos Santos (2018), Kajiwara and Komachi (2018), Kauchak (2016), Malmasi et al. (2016), Malmasi and Zampieri (2016), Martínez Martínez and Tan (2016), Mukherjee et al. (2016), Paetzold and Specia (2016b), Ronzano et al. (2016), Wani et al. (2018), and Zampieri et al. (2016) |
| LINEAR MODELS<br>Regression (Logit, MaxEnt, Linear); Support Vector Machines (LibSVM, v-SVR) | 10 | AbuRa'ed and Saggion (2018), Bingel et al. (2016), Butnaru and Ionescu (2018), Choubey and Pateria (2016), Gooding and Kochmar (2018), Konkol (2016), Kuru (2016), Shardlow (2013a), sp et al. (2016), and Wani et al. (2018) |
| DEEP LEARNING<br>FFN, LSTM, CNN, embeddings | 6 | Aroyehun et al. (2018), Bingel and Bjerva (2018), Bingel et al. (2016), De Hertog and Tack (2018), Hartmann and dos Santos (2018), and Wani et al. (2018) |
| DECISION TREES | 3 | Davoodi and Kosseim (2016), Quijada and Medero (2016), and Wani et al. (2018) |
| NEAREST NEIGHBORS | 3 | Palakurthi and Mamidi (2016) and Yimam et al. (2017a, 2017b) |
| NAÏVE BAYES | 2 | Mukherjee et al. (2016) and Popović (2018) |

**Table 3.4**

*Frequently Used Features of Lexical Complexity in the CWI Tasks*

| Features | # | Papers |
|---|---|---|
| LENGTH characters, syllables, vowels, stem | 26 | AbuRa'ed and Saggion (2018), Alfter and Pilán (2018), Bingel and Bjerva (2018), Bingel et al. (2016), Butnaru and Ionescu (2018), Choubey and Pateria (2016), Davoodi and Kosseim (2016), De Hertog and Tack (2018), Gooding and Kochmar (2018), Hartmann and dos Santos (2018), Kajiwara and Komachi (2018), Konkol (2016), Malmasi et al. (2016), Malmasi and Zampieri (2016), Mukherjee et al. (2016), Paetzold and Specia (2016b), Palakurthi and Mamidi (2016), Quijada and Medero (2016), Ronzano et al. (2016), Shardlow (2013a), sp et al. (2016), Wani et al. (2018), Wróbel (2016), Yimam et al. (2017a, 2017b), and Zampieri et al. (2016) |
| FREQUENCY, LIKELIHOOD word, collocation, document | 25 | AbuRa'ed and Saggion (2018), Alfter and Pilán (2018), Aroyehun et al. (2018), Bingel and Bjerva (2018), Bingel et al. (2016), Brooke et al. (2016), Choubey and Pateria (2016), Davoodi and Kosseim (2016), De Hertog and Tack (2018), Kajiwara and Komachi (2018), Kauchak (2016), Konkol (2016), Malmasi et al. (2016), Malmasi and Zampieri (2016), Paetzold and Specia (2016b), Palakurthi and Mamidi (2016), Quijada and Medero (2016), Ronzano et al. (2016), Shardlow (2013a), sp et al. (2016), Wani et al. (2018), Wróbel (2016), Yimam et al. (2017a, 2017b), and Zampieri et al. (2016) |
| SEMANTICS senses, synonyms, hypernyms, hyponyms | 18 | AbuRa'ed and Saggion (2018), Alfter and Pilán (2018), Bingel and Bjerva (2018), Bingel et al. (2016), Butnaru and Ionescu (2018), Choubey and Pateria (2016), Davoodi and Kosseim (2016), Gooding and Kochmar (2018), Hartmann and dos Santos (2018), Konkol (2016), Mukherjee et al. (2016), Paetzold and Specia (2016b), Palakurthi and Mamidi (2016), Quijada and Medero (2016), Ronzano et al. (2016), Shardlow (2013a), sp et al. (2016), and Wani et al. (2018) |

**Table 3.4**

*Frequently Used Features of Lexical Complexity in the CWI Tasks (Continued)*

| Feature | # | Papers |
|---|---|---|
| MORPHOSYNTAX<br>part of speech | 13 | Alfter and Pilán (2018), Bingel and Bjerva (2018), Bingel et al. (2016), Choubey and Pateria (2016), Davoodi and Kosseim (2016), Gooding and Kochmar (2018), Konkol (2016), Mukherjee et al. (2016), Paetzold and Specia (2016b), Quijada and Medero (2016), sp et al. (2016), and Yimam et al. (2017a, 2017b) |
| N-GRAMS<br>words, characters | 11 | Alfter and Pilán (2018), Bingel et al. (2016), Choubey and Pateria (2016), Gooding and Kochmar (2018), Hartmann and dos Santos (2018), Konkol (2016), Malmasi et al. (2016), Malmasi and Zampieri (2016), Paetzold and Specia (2016b), Popović (2018), and Quijada and Medero (2016) |
| EMBEDDING | 11 | Alfter and Pilán (2018), Bingel and Bjerva (2018), Bingel et al. (2016), Butnaru and Ionescu (2018), De Hertog and Tack (2018), Hartmann and dos Santos (2018), Kuru (2016), sp et al. (2016), Wróbel (2016), and Yimam et al. (2017a, 2017b) |
| SYNTAX<br>length, dependency | 7 | AbuRa'ed and Saggion (2018), Bingel et al. (2016), Gooding and Kochmar (2018), Malmasi and Zampieri (2016), Ronzano et al. (2016), Wani et al. (2018), and Wróbel (2016) |
| PSYCHOLOGY<br>norms | 7 | Alfter and Pilán (2018), Aroyehun et al. (2018), Davoodi and Kosseim (2016), De Hertog and Tack (2018), Gooding and Kochmar (2018), Hartmann and dos Santos (2018), and Quijada and Medero (2016) |
| POSITION | 5 | Bingel et al. (2016), Choubey and Pateria (2016), Mukherjee et al. (2016), Ronzano et al. (2016), and Wróbel (2016) |
| WORD LISTS | 4 | Brooke et al. (2016), Mukherjee et al. (2016), Paetzold and Specia (2016b), and Ronzano et al. (2016) |

meanings. These indices do provide a general measure of semantic complexity (e.g., see Crossley et al., 2009; Millis & Button, 1989). However, they do not allow to consider the effect of the surrounding context, as would a measure of word surprisal, for instance.

*Neural Networks & Deep Learning*

The CWI shared tasks showed that representations of lexical complexity and difficulty could also be learned using neural networks and deep learning. However, compared to the frequent use of classical and ensemble machine learning methods, only a small number of neural architectures were investigated in both tasks. The list of systems that competed in the first task included only one neural system, a feedforward network that integrated a large set of features of lexical complexity (Bingel et al., 2016). However, the network did not achieve a performance comparable with the top-performing ensemble methods.[31] In the second task, more studies examined the development of neural network architectures and, contrary to the previous task, some even achieved top-tier performance (Bingel & Bjerva, 2018; De Hertog & Tack, 2018). This performance was likely due to pre-trained word embeddings (word2vec), which seemed to capture the same information attested in the engineered features of lexical complexity (word frequency). Indeed, De Hertog and Tack (2018) showed that adding features of lexical complexity on top of the word and character embeddings did not significantly increase system performance. Consequently, it can be concluded that both engineered features and word and character embeddings indicate the complexity of a particular word, which in turn predicts lexical difficulty.

## 3.4 CONCLUSION

This chapter has shown that lexical difficulty in reading can be determined in several ways:

31 It should be noted that Bingel et al. (2016) reported on a post hoc experiment that showed a significant performance increase of a revised neural network. In particular, this revision was partly based on optimizing the decision threshold, which maps the logit probabilities to a binary value.

1. based on a priori knowledge, including thresholds of word complexity, expert lexical simplifications, and information-theoretic estimates of word surprisal;

2. based on a posteriori knowledge, including empirical measurements such as eye movements, brain signals, vocalizations, and self-assessments; and

3. with statistical models of lexical difficulty.

Explanatory models evidenced the influence of word frequency, word surprisal, and contextual constraints on different difficulty measurements. Predictive models showed that ensemble methods integrating a large set of complexity features achieved the best performance, although deep neural networks came close to this performance.

As such, this chapter concludes the literature part. The following two parts will further investigate the construct of lexical difficulty in L2 reading, both from a measuring point of view and from a modeling point of view. Part ii will look at how lexical difficulty can be measured from a priori knowledge of difficulty (Chapter 4) and a posteriori knowledge (Chapter 5). Part iii will look at how lexical difficulty can be predicted from an explanatory model of lexical complexity (Chapter 6) and statistical machine learning (Chapter 7). Notably, the investigation will touch upon several limitations observed in the methods reviewed in this chapter regarding the two main research objectives: contextualization and personalization.

CONTEXTUALIZATION The empirical and computational approaches reviewed in this chapter showed evidence for adopting a contextualized approach. Fundamentally, this context effect was captured by the information-theoretic notion of surprisal. Studies on L1 reading showed, for instance, that word surprisal estimates obtained from *n*-gram models and recurrent neural networks could predict significant differences in the amplitude of the N400 brain potential (Frank et al., 2013, 2015). Similarly, studies on L2 reading showed that the word's cloze probability (obtained from human raters) could predict difficulties in contextual word learning (Chen et al., 2017). Non-natives displayed high N400 amplitudes when first encountering unknown words embedded

in high-constraint sentence contexts (i.e., with a high cloze probability), but this difficulty rapidly subsided after some encounters. Difficulties in word-to-context integration persisted even after multiple encounters with unknown words embedded in low-constraint sentence contexts (i.e., with a low cloze probability). However, it is crucial to bear in mind that task effects may influence these measures of word-to-context integration. For instance, more significantly marked differences in eye-movement measures were observed for words embedded in shorter contexts (i.e., sentence reading) than when they appeared in much longer contexts (i.e., natural reading) (Dolgunsöz, 2016).

Nevertheless, this contextualized approach was lacking in other reviewed methods in the chapter. In this respect, the most notable methods were theoretical measures based on a frequency threshold derived from a vocabulary list. As the surprisal effects cited above, these thresholds aimed to capture the effect of word occurrence on difficulty. However, while word surprisal was estimated from contextual language models, the vocabulary lists were based on a frequency tally that disregarded the surrounding context. Chapter 4 will see whether such a theoretical frequency threshold could be enhanced by tallying word occurrence based on the surrounding context.

Secondly, not many statistical machine learning methods performed a contextualized prediction of lexical difficulty. The systems reviewed in this chapter mainly integrated features of the word itself, the most frequent of which were word length and word frequency. Admittedly, some of these features were contextualized features complexity, such as n-gram probabilities and topic distributions. Still, the vast majority of the machine learning algorithms (ensemble methods, linear models, feedforward neural networks) did not predict the difficulty of a word from preceding or following observations. Chapter 7 will see whether empirical predictions of lexical difficulty could be enhanced with a contextualized decision function, derived in particular from a bidirectional long-short term memory network.

PERSONALIZATION The empirical studies reviewed in this chapter also evidenced the need to consider lexical difficulty for each learner individually. Indeed, various studies reported a considerable between-subjects variance in the measurement outcomes, both in L1 and L2 readers, indirect/im-

plicit (eye movements, brain potentials), and direct/explicit (misreadings, self-assessments) measures. For instance, L2 readers differed significantly in their eye-movements patterns, even when the same context conditions were considered (Godfroid et al., 2013).

A significant source of variability is the reader's proficiency level. On the one hand, some theoretical measures set a threshold of lexical complexity, not based on general corpus frequencies, but on the frequency of occurrence in graded reading materials intended for the reader's proficiency level (Tack et al., 2016a). On the other hand, some empirical studies showed a critical interaction between the readers' proficiency level and their eye movements and brain potentials. For instance, high-intermediate and advanced proficiency levels displayed quicker eye-movement measures (Dolgunsöz, 2016; Dolgunsöz & Sariçoban, 2016). Similarly, the EEG potentials indicative of lexical processing difficulties also differed significantly between proficiency levels and brain regions (Chen et al., 2017).

Nevertheless, not all studies that measured lexical difficulty from self-assessments accounted for this variability between readers. The machine learning systems submitted to the CWI shared tasks, in particular, were developed on aggregated measures of difficulty, where various participants annotated different parts of the data. Consequently, the aggregated nature of the train and test sets gave rise to a valuable loss in information on between-reader variance. Admittedly, the first shared task did experiment with predicting personal difficulty, but the systems were also developed on a training set of aggregated measures. Chapter 5 will describe two trials where learner-specific self-assessment measures were collected and where all participants read the same reading materials. Chapters 6 and 7 will look at two modeling procedures that integrate this between-reader variance into lexical complexity features or neural networks.

PART II

# MEASURING LEXICAL DIFFICULTY

Put not yourself into amazement how these
things should be: all difficulties are but easy
when they are known.

— William Shakespeare, *Measure for Measure*, IV.ii.2108-10

CHAPTER 4

# *A Priori Knowledge Of Difficulty* WORD OCCURRENCE IN CEFR-GRADED READING MATERIALS

**Abstract** This chapter discusses the extraction of word occurrence from reading materialsfor specific CEFR levels as a first approach to measuring lexical difficulty for L2 readers. First, the chapter introduces a novel resource for Dutch L2, namely NT2Lex. The main novelty is that lexical entries are linked to a semantic network for Dutch (viz., Open Dutch WordNet) with automatic word-sense disambiguation (WSD). Next, the chapter discusses several analyses of how the resource accounts for some well-established effects. The first analysis examines whether the distribution of novel words is coherent with what one would expect in terms of frequency, dispersion, sophistication, semasiology, and psycholinguistic norms. The second analysis examines the added value of WSD for computing indices of semantic complexity. The final analysis compares the distribution of Dutch-French cognates in the Dutch resource and its French equivalent.

A fundamental starting point for measuring difficulty is to consider the grading of reading materials during a language curriculum development. These language curricula follow a standard educational scale used to order learning materials according to successive grade or proficiency levels. A well-known educational scale for English L1 speakers is the US grade scale, used in most of the foundational measures of readability (Dale & Chall, 1948; Gunning, 1952; Mc Laughlin, 1969). For the L2 language curriculum, one of the

Parts of this chapter have been presented or published in Tack et al. (2017, 2018a, 2018b, 2019)

most widespread scales used to date is the Common European Framework of Reference (CEFR) (Council of Europe, 2001). The CEFR is a general framework that aims to provide a comprehensive description of the types of written or spoken discourse a foreign language learner can understand or produce at a particular proficiency level. The framework distinguishes six proficiency levels, ranging from the elementary (A1/A2) to the intermediate (B1/B2) and advanced (C1/C2) levels. Because of the popularity of the CEFR framework, many educational curriculum designers have therefore used the scale to grade various materials and textbooks for foreign language learners.

Because the difficulty level is determined by the a priori knowledge of expert curriculum designers, these graded learning materials give a hypothetical indication of acquisitional complexity. A fundamental assumption that underlies this expert knowledge is that specific linguistic structures display a degree of complexity that can only be adequately understood or produced at specific developmental stages. In other words, certain linguistic features are *criterial* of specific developmental stages in the target language, which some researchers have tried to connect with each of the six CEFR levels (Bartning & Schlyter, 2004; Hawkins & Buttery, 2010; Hulstijn et al., 2010). Some reference level descriptors (RLDs), which inventory the criterial features of each CEFR level, are currently available for various target languages, such as the English Profile (English Profile, 2011) and the *Référentiels* for French L2 (Beacco, 2004, 2008; Beacco et al., 2011; Beacco et al., 2005; Beacco et al., 2008). For vocabulary, these RLDs list the lexical units typical of different CEFR levels (Marello, 2012; Milton, 2010). These lists draw from empirical evidence of the vocabulary that learners can produce at various proficiency levels. An example is the English Vocabulary Profile, which draws information from the Cambridge Learner Corpus (Capel, 2010, 2012). As such, these descriptors follow the tradition of deriving reference word lists from corpora (e.g., Nation & Waring, 1997).

Another way to link the CEFR scale to (receptive) L2 vocabulary is to look directly at the type of words that occur in CEFR-graded reading materials. From a corpus of graded textbooks and readers, one can extract, for each level on the CEFR scale, a list of all the lexical units that occur in reading activities at that level. Moreover, one can also simultaneously estimate their frequency of occurrence and dispersion across these graded reading activities. This type

of lexical database is defined as a CEFR-graded lexical resource. Before the current study, a number of graded lexicons had already been developed for French L2 (Francois et al., 2014) and Swedish L2 (François et al., 2016; Volodina et al., 2017), and since then, another resource has also been published for English L2 (Dürlich & François, 2018).

The study reported in this chapter aims to contribute to the development of CEFR-graded lexicons. In particular, the study offers several contributions. Firstly, the study introduces the NT2Lex resource, a novel graded receptive lexicon for Dutch as a foreign language (NT2)[32]. Moreover, the study extends the previous methodology by introducing the first lexicon with semantically disambiguated lexical entries linked to a semantic network (viz., Open Dutch WordNet). These contributions will be described in Section 4.2. In addition, the study also examined the possible use of a CEFR-graded lexicon for deriving a measure of lexical difficulty in reading. In particular, the analyses reported in Sections 4.3 to 4.5 address the following questions:

RQ4.1 If a CEFR-graded lexicon can be used to derive a measure of lexical complexity, does this measure correlate with other well-known indicators of lexical complexity?

RQ4.2 Can word-sense disambiguation enhance the computation of lexico-semantic complexity such as the degree of hypernymy?

RQ4.3 Does a CEFR-graded resource account for learner characteristics such as cognate status?

The chapter is structured as follows. Section 4.1 presents the background of the study. Section 4.2 describes the methodology used to develop the NT2Lex resource. Sections 4.3 to 4.5 present the analyses. Sections 4.6 and 4.7 cover the main implications and conclusions of the study.

## 4.1 BACKGROUND

This section introduces the background and rationale behind several aspects addressed by the research questions. The core aspect underlying all these

32 The acronym NT2 refers to the term *Nederlands tweede taal*, or Dutch as a second language as it is referred to in Dutch.

questions is concerned with the development of frequency-based measures of lexical difficulty. Previous work on the computation of lexical frequencies linked to a scale of difficulty is introduced in Section 4.1.1. Another aspect addressed in the second research question is how semantics relates to lexical complexity. In most frequency-based measures, this aspect has been overlooked as few measures distinguish a word based on its specific meaning in a given context. Previous work adopting a lexico-semantic perspective on L2 vocabulary and complexity is introduced in Section 4.1.2. A final aspect pertains to how a difficulty measurement can account for learner characteristics. Section 4.1.3 presents some studies in applied linguistics on cognate status as one possible learner characteristic.

### 4.1.1 *Lexical Frequencies Linked to a Difficulty Scale*

Much of the seminal work for this study was done by Carroll et al. (1971) in their creation of *The American Heritage Word Frequency Book*, a reference vocabulary list including frequencies estimated from texts read by English natives at various US grade levels. Carroll et al.'s study put forward several novel statistics that aimed to adjust the frequency tally *F*: the dispersion index *D*, the normalized frequency per one million words *U*, and the standard frequency index *SFI*.[33] These adjusted measures were later used by Zeno et al. (1995) in their compilation of *The Educator's Word Frequency Guide*, a more extensive vocabulary list for speakers of English at various grade levels, from kindergarten to college (K–12 coursebooks). An important part of their work was that, for each entry in the list, the frequency statistics were computed per each distinct grade level.

Later, Lété (2004) and Lété et al. (2004) used the methodology seminally invented by Rinsland (1945) and proposed by Zeno et al. (1995) to develop a similar grade-level lexical database for French.[34] The database was named MANULEX – a syllabic abbreviation of *lexique des manuels* 'textbooks lexicon' – and included lexical frequencies for three difficulty levels: the first grade, the

33 These are not the only adjusted measures that have been proposed in the literature. For a comprehensive overview of adjusted frequencies and dispersion indices, see Gries (2008).

34 Other grade-level lexical databases have been developed for other languages as well, such as the ESCOLEX database for Portuguese (Soares et al., 2013).

second grade, and the third to fifth grades of elementary education in France.[35] More recently, studies by Billami et al. (2018), Gala et al. (2013), and Gala and Javourey-Drevet (2020) expanded on Lété et al.'s work by proposing ReSyF, a lexical database with synonyms attested in the lexico-semantic network JeuxDeMots (Lafourcade, 2007) ranked with MANULEX levels.

It appears that linking lexical frequencies to various grade levels has been a long-standing topic in L1 research. Regarding L2 research, recent studies have applied the same methodology to the development of graded lexical resources based on the CEFR scale.[36] These studies focused on graded textbooks or readers used in language curricula designed for foreign language learners. Up to the time of writing, a number of graded lexicons have been developed for French L2 (FLELex; Francois et al., 2014), Swedish L2 (SVALex; François et al., 2016; Volodina et al., 2017), Dutch L2 (the NT2Lex resource presented in the current chapter; Tack et al., 2017, 2018b), and English L2 (EFLLex; Dürlich and François, 2018). As these receptive graded lexicons include frequencies estimated from reading materials[37], they could give information about the kind of vocabulary that should be understood when reading in a foreign language at a particular proficiency level.

The lexical resources cited above have also found their purpose as components of NLP-driven educational applications. Some of the resources served as features of a predictive model for French L2 vocabulary (Tack et al., 2016a, 2016b), as components in a readability-driven learning platform for Swedish (Pilán, Volodina, & Borin, 2016), or as part of an automated essay grading system for Swedish (Pilán, Volodina, & Zesch, 2016). Therefore, it is clear to say that the scope of relevance of the graded lexical resources goes well beyond their apparent usefulness to gain didactic insights into the complexity of the L2 curriculum.

Notably, these lexicon-based approaches have overlooked an important aspect: the distinction of word senses. Most reference vocabulary lists include frequencies tallied either for an inflected word form, a lemma, and a part

35 These grade levels are: *Cours préparatoire* (CP; G1 in MANULEX), *Cours élémentaire 1ère année* (CE1; G2 in MANULEX), and *Cours élémentaire 2ème année*, *Cours moyen 1ère année*, and *Cours moyen 2ème année* (CE2, CM1, CM2; G3-5 in MANULEX).

36 See the CEFRLex project: https://cental.uclouvain.be/cefrlex/.

37 The same methodology has also been applied to compute a productive CEFR-grade lexicon from a corpus of texts written by learners of Swedish L2 (SweLLex; Volodina et al., 2016).

of speech. However, there is a drawback to measuring difficulty without considering the various senses of a word. As Tharp (1939) noted:

> Four major errors or fallacies have been present in the preceding techniques of measuring difficulty: (...) (3) (...) [T]here has been ignored likewise the added burden of multiple meanings of a single item. *Faire* has been tallied as a single word, whether it means "do" or "make," or "be" in a weather meaning, or "pay" with *attention*, etc. (p. 172)

Despite some lexical simplification studies that ranked the difficulty of synonyms based on contextual factors (Jauhar & Specia, 2012) or based on a lexical database of synonyms ranked according to elementary grade levels (Billami et al., 2018; Gala et al., 2013), many of the previously cited studies still have not accounted for the multiple senses of the word.

### 4.1.2 *Semantic Complexity and Lexico-Semantic Networks*

In research on L2 reading comprehension and vocabulary acquisition, several studies underscored the importance of taking into account form-meaning mappings. For instance, Qian (1999) highlighted that in the interplay between vocabulary size and reading comprehension, the notion of vocabulary knowledge depth played a significant role. Although it is arguable whether depth is a proper dimension of vocabulary knowledge (Gyllstad, 2013)[38], the term has been used throughout the literature to refer to a comprehensive mastery of the various features of the word. Besides this notion of comprehensive vocabulary knowledge, a "deeper" vocabulary knowledge also entails knowing the precise, specific meaning of a word and knowing how the word relates to other words in the mental lexical network. Read (2004) most notably summarized these notions of *comprehensive word knowledge*, *precision of meaning*, and *network knowledge*, which are defined in Table 4.1.

An essential consequence of looking at vocabulary knowledge in this way is that it supports the view of L2 lexical growth in terms of a lexical network.

38 One key argument against defining breadth and depth as two separate dimensions is that they have shown to be highly correlated. Qian (1999) also noted this correlated nature: "breadth and depth are two interconnected dimensions of vocabulary knowledge, the development of which are interdependent to a substantial extent" (p. 287).

**Table 4.1**

*Conceptions of Depth of Vocabulary Knowledge*

| | |
|---|---|
| Precision of meaning | the difference between having a limited, vague idea of what a word means and having a much more elaborated and specific knowledge of its meaning |
| Comprehensive word knowledge | knowledge of a word, not only its semantic features but also orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics |
| Network knowledge | the incorporation of the word into a lexical network in the mental lexicon, together with the ability to link it to – and distinguish it from – related words |

*Note.* This table was adapted from Read (2004, pp. 211–212).

A few noteworthy studies have shown how the development of multilingual lexical representations can be modeled with various types of networks, such as self-organizing neural networks (DevLex; Farkas and Li, 2002), random autonomous boolean networks (Meara, 2006), or models following the Dynamic Systems Theory of language learning (de Bot et al., 2007; Filatova, 2010; Lowie et al., 2010). Although these studies have given remarkable insights into how lexical representations emerge in multilingual networks, they open up perspectives that would require a much more extensive investigation beyond the scope of this study. Another approximate way to examine L2 lexical growth from a network perspective is to study the role of various semantic relations attested in a reference lexico-semantic network such as WordNet (Fellbaum, 1998).

*Hypernymy in L2 Lexical Development and Reading*

One quintessential semantic relation in a lexical network is concerned with hypernymy. The hypernymic relation between two words – or, more correctly, between the concepts to which they refer – is a hierarchical semantic relation: it characterizes the association between a superordinate concept or *hypernym* and its subordinate concept or *hyponym*. Because of this hierarchical organization, the degree of hypernymy can be formalized in a tree structure. Then, various concepts can be compared in terms of their depth in the tree. The tree's root node has the lowest depth and corresponds to the most all-encompassing, superordinate concept. Conversely, the leaf nodes of the tree correspond to the most specific, subordinate concepts.

Importantly, the position of the word in the hypernymy tree is characteristic of the acquisitional complexity of a word. Crossley et al. (2009), for instance, showed that the order by which words were learned for productive use in a foreign language was conditioned by hypernymic relations, which they computed with the Coh-Metrix tool (Graesser et al., 2004). Based on the reviewed literature, the authors hypothesized that words located at the highest ranks in the hypernymy tree would be acquired first. Because they are located higher up in the tree, these early-acquisition words have a very general meaning (e.g., *cat*, *dog*, etc.). Therefore, they need to internalized first before more specific meanings (e.g., *toy dog*, *Brussels griffon*, etc.) can be attached to the previously acquired concepts. Another study by Crossley and Salsbury (2010) gave further evidence of concept specificity as a significant feature in the early acquisition stages: "less specific verbs are produced first, while more specific verbs are not produced" (p. 135). This notion of concept specificity therefore seems to be related to the notion of precision of meaning (see Table 4.1). As their vocabulary knowledge increases in size and depth, learners can produce more precise word meanings located deeper in the hypernymy tree. Consequently, one could consider the degree of conceptual specificity as a criterial feature of specific stages in L2 vocabulary development.

Of course, these effects concerned productive word learning, but the same applies to receptive word learning and reading. Crossley et al. (2007) tested the hypothesis whether authentic and simplified reading texts differed in terms of hypernymy, but even though "the simplified texts showed a tendency to

contain more abstract words, (...) these findings were not significant" (p. 24). In another study on receptive word recognition from semantic priming, Crossley (2013) found that hypernymic relations mainly determined L2 lexical networks: "unidirectional priming effects suggesting that L2 lexicons are characterized by network connections that activate between subordinate and superordinate terms, but not between superordinate and subordinate terms" (p. 109). Recently, Gagné et al. (2020) found that hyponymy, as a relation of subclassification, significantly predicted lexical decision latencies on semantically transparent compound words such as *teacup*, as opposed to non-transparent compounds such as *honeymoon*. Finally, other studies have also sought to predict the readability of a word by using polysemic, hypernymic, and other semantic features extracted from WordNet. For example, various studies have used the depth in the WordNet hypernymy tree as a feature to automatically predict the difficulty of words for non-native readers (AbuRa'ed & Saggion, 2018; Bingel & Bjerva, 2018; Mukherjee et al., 2016; Ronzano et al., 2016; Wani et al., 2018).

Related to the notion of lexico-semantic networks is the task of word-sense disambiguation. When computing the depth in the WordNet hypernymy tree, the specific sense of a word needs to be known beforehand. However, in most complexity analysis tools such as Coh-Metrix, it is generally impossible to either manually specify or automatically obtain the precise sense of a word. Because no manual or automatic word-sense disambiguation is performed after the standard lemmatization and part-of-speech tagging, it is practically impossible to determine the precise hypernymy rank for a particular concept. Instead, the solution is to provide an average of the rank over all possible senses of a word. However, there may be a drawback to using an index of semantic complexity that averages the hypernymy rank over all possible word meanings, which does not seem to have been addressed.

### 4.1.3 *Cognateness*

Complementary to the factor of concept specificity is the notion of semiotic transparency, or the fact that the meaning of a word in the target language is transparent because of the existence of an equivalent linguistic form, or

cognate, in the learner's native language. However, the exact interpretation of what defines a cognate appears to vary between study areas.

*Defining Cognates*

In historical linguistics, cognate words have been defined in a narrow sense. Campbell and Mixco (2007), for instance, defined cognateness as follows:

> **cognate** A word (or morpheme) that is related to a word (morpheme) in sister languages by reason of these words (morphemes) having been inherited by the related languages from a common word (morpheme) of the **proto-language** from which they descend. (p. 33)

By this definition, only words in cognate languages having a common etymon in the proto-language are considered cognates, hence excluding borrowing.

Cognates have also been defined in a broader sense as either linguistic systems (i.e., languages) or meaningful linguistic forms (i.e., morphemes or words) originating from a common ancestor. Crystal (2008), for instance, defined the term as follows:

> **cognate** (*adj./n.*) (1) A language or a LINGUISTIC FORM which is historically derived from the same source as another language/form, e.g. Spanish/Italian/ French/Portuguese are 'cognate languages' (or 'cognates'); *père/padre*, etc. ('father') are 'cognate words' or cognates (p. 83)

For linguistic forms, in particular, this definition is broad because it covers all words that are "historically derived from the same source". As such, cognates can be related either by descent, derivation, or borrowing. To use a common example, the French *père*, the Spanish *padre*, and the English *father* are cognates by descent from the Proto-Indo-European etymon *$*ph₂tḗr$*, while the Dutch *abrikoos* and the French *abricot(s)* are cognates by borrowing. Likewise, related words in two non-cognate languages, such as the English *hot dog* and the Japanese ホットドッグ (*hottodoggu*), are cognates by borrowing.

In sum, a cognate is mainly defined in function of the etymological relatedness between two linguistic forms, which can be interpreted either narrowly (by descent only) or broadly (by descent, derivation, and borrowing).

*Cognates in Applied Linguistics*

In applied linguistics, it is striking that cognate status has chiefly been referred to by focusing on additional formal and semantic similarity constraints. Also, the defining factor of etymological relatedness is not always explicitly stated. To illustrate these constraints, let us consider this non-exhaustive list of definitions used in studies in applied linguistics:

> The cognate status variable involves differences between words in terms of the form relation with their translation in the target [foreign language]. Cognate words *share (parts of) their orthographic and/or phonological form* [emphasis added] with their translations, whereas noncognate words are dissimilar in form to their translations. (de Groot & Keijzer, 2000, p. 3)

> In other contexts (...) the term is often used more loosely, denoting words in different languages that are *similar in form and meaning* [emphasis added], without making a distinction between borrowed and genetically related words (...). (Kondrak, 2001, p. 1)

> *Cognates* are words in two languages that share both *a similar meaning and a similar phonetic (and, sometimes, also orthographic) form* [emphasis added], due to a common ancestor in some protolanguage. The definition is sometimes also extended to words that have *similar forms and meanings* [emphasis added] due to *borrowing*. (Rabinovich et al., 2018, p. 329)

Narrowing down the definition of cognates based on formal and semantic similarity is not unchallenging, however. For instance, to what extent would the very different forms *père* and *father* still be defined as cognates in the examples cited earlier? Moreover, these definitions do not seem to clarify at what specific point in the degree of formal dissimilarity one should stop considering two linguistic forms as cognates.

Although these additional constraints define a cognate in a way that is open to equivocal interpretations, they nevertheless make sense given the context of applied linguistics. The formal and semantic similarity between two cognates explains the ease or difficulty they will be correctly translated from

the source or native language to the target or foreign language. In other words, the criterion of "translational equivalence" (Mitkov et al., 2007) is essential for the detection of possible issues regarding the adequate contextual translation of cognate words, both productively and receptively. Indeed, cognate effects have shown to explain the ease in processing and learning words for both bilinguals and foreign language learners (de Groot & Keijzer, 2000; de Groot & Nas, 1991):

> Whereas in noncognate learning new entries have to be created in memory, cognate learning may only involve adding new information to, or adapting, memory representations that already existed in memory prior to learning. The former process may be more demanding than the latter, thus causing the disadvantage for noncognates. (de Groot & Keijzer, 2000, p. 31)

Moreover, various eye-movement studies showed not only facilitation effects for cognates with higher formal similarity (Duyck et al., 2007), but also inhibition effects for interlingual homographs (e.g., the English *coin* and the French *coin* 'corner') in low-constraint contexts (i.e., which did not enable the meaning to be congruently construed; Libben and Titone, 2009). Similarly, cognateness has also shown to be a significant factor in determining readability for foreign language learners (Beinborn et al., 2014).

In consequence, if one were to measure lexical difficulty from graded reading materials, it is crucial to examine to what extent this measure accounts for a well-established effect such as cognate status. However, this has not yet been addressed so far.

### 4.1.4 *Research Hypotheses*

In light of previous studies, the current study examined the use of a CEFR-graded lexicon to measure lexical difficulty for foreign language learners. The study presents the case of NT2Lex, used to verify three hypotheses. Each one of these pertains to a research question described above.

H4.1 The hypothesis for RQ4.1 is that there will be a clear increasing trend in complexity when novel entries occur at increasing difficulty lev-

els in NT2Lex. In particular, this trend should be visible in lexical features such as frequency, dispersion, sophistication, polysemy, and psycholinguistic norms (concreteness and age of acquisition).

H4.2 The hypothesis for RQ4.2 is that word-sense disambiguation will lead to better discriminate difficulty levels in NT2Lex and achieve a more accurate index of complexity than word-sense averaging.

H4.3 The hypothesis for RQ4.3 is that there will be a more significant proportion of transparent cognates, particularly those that are translation equivalents, occurring at lower difficulty levels. Moreover, this effect should remain comparable across languages. More specifically, this effect should remain comparable for Dutch-French cognates attested in NT2Lex and FLELex.

## 4.2 NT2LEX: A CEFR-GRADED LEXICON FOR DUTCH AS A FOREIGN LANGUAGE LINKED TO OPEN DUTCH WORDNET

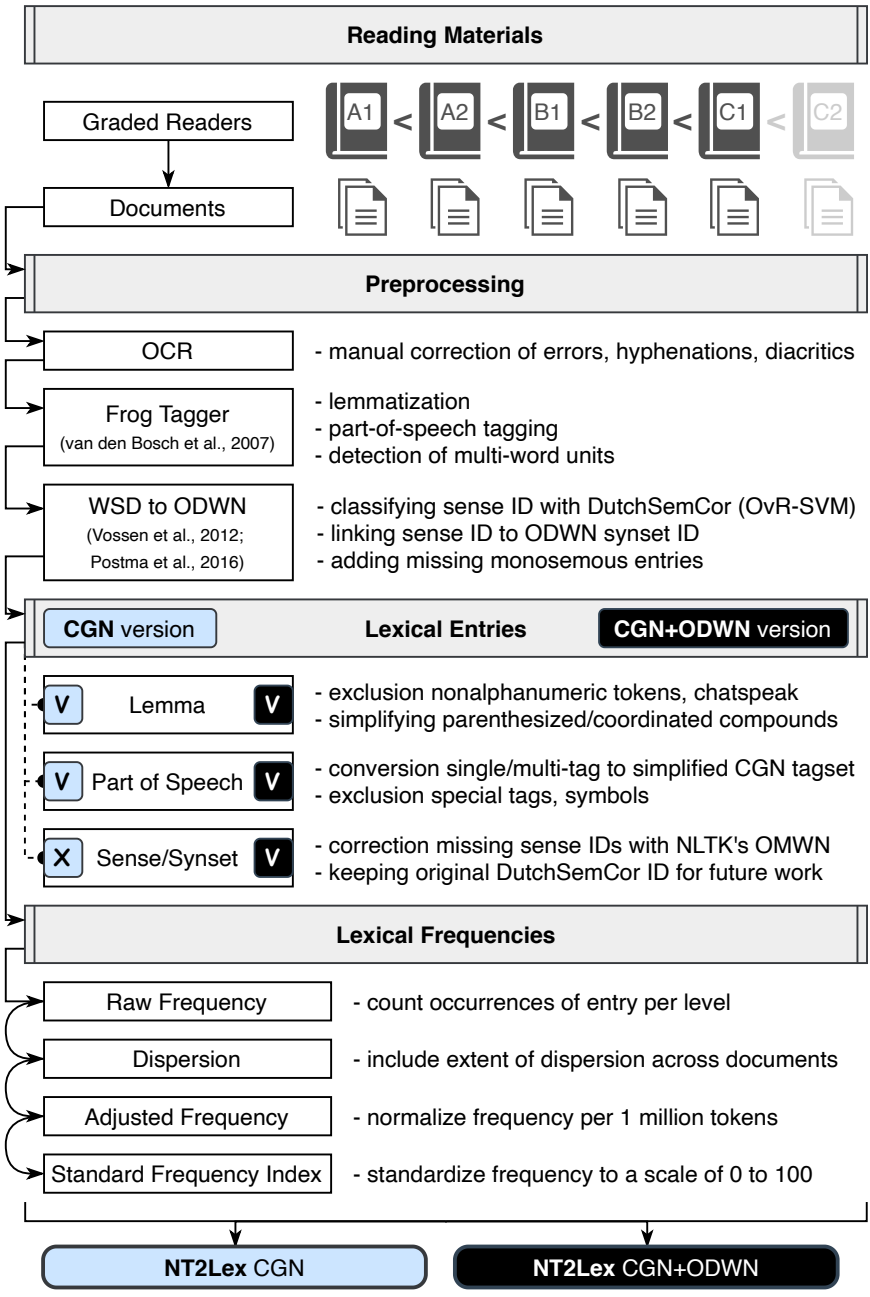
Figure 4.1 outlines the various steps taken in developing a new graded lexical resource for Dutch L2. These steps are further explained in Section 4.2.1. Moreover, the NT2Lex resource comprises two versions: a standard version including a list of entries identified by their lemma and part of speech (viz., CGN) and an enhanced version including word-sense disambiguated entries linked to Open Dutch WordNet (viz., CGN+ODWN). These two versions are described in Section 4.2.2. Finally, some web-based tools were created during the development process and will be illustrated in Section 4.2.3.

### 4.2.1 *Resource Development*

The general methodology for compiling a graded lexical resource essentially involves three steps: identifying a specialized corpus of graded reading materials, extracting relevant lexical units from this corpus, and computing adjusted and standardized frequencies per each grade level. For a detailed account of the original methodology, see Lété et al. (2004). Further information on

**Figure 4.1**

*Flowchart of the NT2Lex Resource Development Process*



creating a graded lexicon intended specifically for L2 learning can also be found in Francois et al. (2014).

*Reading Materials Graded With CEFR Levels*

The resource development process started with the search for relevant CEFR-graded materials with reading activities. To be considered eligible for inclusion, these materials included books intended for teaching or learning Dutch as a foreign language, either graded coursebooks or graded readers, with an explicit CEFR level. Furthermore, materials could also be in the standard Netherlandic or Belgian Dutch language. In all, the collection of relevant materials counted 45 CEFR-graded books.

Next, a selection of relevant text documents was extracted from these books. In the graded readers, each chapter was a separate document. In the coursebooks, each reading activity was a separate document. Notably, a coursebook activity was only relevant if it included a text intended for reading comprehension or which served as a cloze reading exercise.[39] Other activities (e.g., grammar exercises and vocabulary lists) were excluded. In all, the selection counted 926 texts labeled with the CEFR level of the book in which they appeared.

Lastly, the selected reading materials were converted into plain-text documents with optical character recognition (OCR) software. Hyphenated line breaks, as well as errors induced by the OCR software, were manually corrected. Furthermore, superfluous diacritical marks were removed, such as those commonly used to indicate stress in written Dutch (e.g., *veel*/*véél* 'many'). The mandatory diaereses (e.g., *efficiënt* 'efficient') and accents in loan words (e.g., *café* 'pub') were not removed. In all, the corpus counted almost half a million tokens.

The statistics listed in Table 4.2 show that the corpus of graded materials was relatively small compared to the usual size of a reference language corpus. In particular, the corpus could not achieve strict representativeness and balance. The reasons for this data scarcity were twofold. Firstly, the focus on the CEFR scale ruled out some graded books for Dutch L2 using other difficulty scales.

39 It is probably needless to say that the solutions to these cloze passages were used, not the original fill-in-the-blank exercises.

**Table 4.2**
*Number of Books, Document, and Tokens in the Corpus of CEFR-Graded Reading Materials for Dutch L2*

| Levels | A1 | A2 | B1 | B2 | C1 | All |
|---|---|---|---|---|---|---|
| Books | 5 | 22 | 11 | 6 | 1 | 45 |
| Documents | 53 | 447 | 306 | 110 | 10 | 926 |
| Tokens | 17,878 | 205,035 | 153,537 | 78,439 | 6,199 | 461,088 |

Consequently, the corpus was not representative of all possible graded reading materials for Dutch L2. Secondly, while it was reasonably easy to obtain relevant reading materials for the beginner to low-intermediate proficiency levels, there was a minimal availability of materials for the advanced levels. This was explained by the fact that the Dutch integration and state exams for NT2 learners target the A2, B1, and B2 levels, respectively.[40] As a result, most books covered only the A1, A2, B1, and B2 levels.

*Lemmatization, Part-of-Speech Tagging, and Word-Sense Disambiguation*

Various natural language processing tools were used to tag each token in the corpus with its canonical form (lemma), part of speech, and disambiguated meaning. First, the text documents were analyzed with the Frog memory-based tagger (van den Bosch et al., 2007). Figure 4.2 provides an example output of the tagger. By default, the tagger performs various other linguistic analyses, including morphological segmentation, named entity recognition, and dependency parsing. For developing the resource, only columns two, three, and six were relevant: the token (including single and multi-word units), the lemma, and the part of speech (based on the CGN tagset; Van Eynde, 2004).

WORD-SENSE DISAMBIGUATION The tokenized, lemmatized, and part-of-speech tagged texts were further analyzed with automatic WSD. Several tools

40 An NT2 learner is required to pass the state exams I & II to obtain a junior college (MBO) and college (HBO) degree in the Netherlands. For more information, see https://www.nt2.nl/fr/dossier/staatsexamen_nt2/staatsexamen_nt2_programma_i__niveau_b1 and https://www.nt2.nl/nl/dossier/staatsexamen_nt2/staatsexamen_nt2_programma_ii__niveau_b2

**Figure 4.2**

*Illustration of the Analyses Performed by the Frog Tagger*

```
1  Jan         Jan       [Jan]        SPEC(deeleigen)   1.0  B-PER  B-NP  2  su
2  woont       wonen     [woon][t]    WW(pv,tgw,met-t)  1.0  O      B-VP  0  ROOT
3  in          in        [in]         VZ(init)          1.0  O      B-PP  2  ld
4  Nederland   Nederland [Nederland]  SPEC(deeleigen)   1.0  B-LOC  B-NP  3  obj1
5  .           .         [.]          LET()             1.0  O      O     4  punct
```

*Note.* The columns represent the following information: index in the sentence, token, lemma, morphological segmentation, part of speech, confidence, named entity, phrase chunk, dependency head, dependency relation.

have been made available in the DutchSemCor project (Vossen et al., 2012). Three WSD tools have been evaluated on DutchSemCor: SVM-WSD with domain features, TiMBL-WSD without domain features, and UKB-WSD.[41] Because the SVM-WSD tool developed by Rubén Izquierdo achieved top – if not the best – performance on classifying senses, this system was used in the current study.[42]

The SVM-WSD tool is based on a one-vs.-rest SVM classifier. Each sense identifier for adjectives, nouns, and verbs has a separate model pre-trained with a bag-of-words approach on DutchSemCor. Based on these pre-trained models for each sense, the classifier computes the probability that a word in a given context expresses this particular meaning. The classifier then outputs the sense identifier with the highest probability. Notably, the SVM-WSD tool has already shown its usefulness to improve a lexical simplification system that translates text into pictographs for native speakers of Dutch with cognitive disabilities (Sevens et al., 2016).

Additionally, the SVM-WSD tool includes a dictionary of 92,617 lexemes and 117,225 senses. This dictionary maps a DutchSemCor sense identifier to its corresponding entry in Open Dutch WordNet (cf. *infra*). An example output of the WSD tool is given in Figure 4.3.

OPEN DUTCH WORDNET Two semantic networks have been made available for Dutch: Cornetto (Vossen et al., 2013) and Open Dutch WordNet (ODWN) (Postma et al., 2016). Cornetto is a lexical database including seman-

41 The results can be found on http://wordpress.let.vupr.nl/dutchsemcor/.
42 The tool is available from the following link: http://github.com/cltl/svm_wsd.

**Figure 4.3**

*Illustration of the Output of the Word-Sense Disambiguation Tool*

| | | | | | |
|---|---|---|---|---|---|
| Jan | Jan | SPEC(deeleigen) | - | - | - |
| woont | wonen | WW(pv,tgw,met-t) | r_v-10183 | wonen-v-1 | odwn-10-109912332-v |
| in | in | VZ(init) | - | - | - |
| Nederland | Nederland | SPEC(deeleigen) | - | - | - |
| . | . | LET() | - | - | - |

*Note.* The final three columns represent the following information: DutchSemCor (Cornetto) ID, ODWN sense ID, ODWN synset ID.

tic relations originating from the Dutch WordNet and the *Referentie Bestand Nederlands*. Cornetto also provides the sense identifiers annotated in the DutchSemCor corpus. However, because the database contains information from proprietary sources, it is not freely available.

ODWN was created to substitute Cornetto's proprietary sources with openly licensed material. The advantage of ODWN is that it also includes DutchSemCor identifiers and can therefore be used in conjunction with the SVM-WSD tool cited above. Another advantage of ODWN is that the database is included in Open Multilingual WordNet (OMW).[43] As such, concepts in Dutch can be linked to the same synsets in other languages and, in particular, English. This means that studies using semantic relations such as hypernymy for Dutch could be more easily comparable with previous studies on English, as the OMW uses the same concept (synset) identifiers for all languages.

Some clarifications are perhaps needed to explain the difference between the DutchSemCor and ODWN identifiers illustrated in Figure 4.3. Recall that each token (e.g., *woont*) was labeled with its most probable DutchSemCor sense class (e.g., *r_v-10183*). This sense class was then mapped to the corresponding lexical entry ID and synset ID in ODWN. On the one hand, a lexical entry represents a semasiological entry in ODWN (see Figure 4.4). Each lexical entry can be looked up by its unique ID (e.g., *wonen-v-1*), which connects the entry's headword (lemma) with its grammatical class (part of speech) and its meaning (sense number). A synset (or 'synonym set'), on the other hand, represents an onomasiological entry in ODWN (see Figure 4.5). The synset ID refers to a unique concept that can be expressed by a set of lexical units, which

43 http://compling.hss.ntu.edu.sg/omw/

**Figure 4.4**
*Example of a Lexical Entry in Open Dutch WordNet*

```
<LexicalEntry id="wonen-v-1" partOfSpeech="verb">
  <Lemma writtenForm="wonen" mode="infinitive"/>
  <WordForms/>
  <Morphology/>
  <MorphoSyntax>
    <auxiliaries auxiliary="hebben"/>
  </MorphoSyntax>
  <SyntacticBehaviour valency="di" transitivity="transitive">
    ...
  </SyntacticBehaviour>
  <Sense id="r_v-10183" senseId="1" definition="huizen"
  ↪  synset="odwn-10-109912332-v" provenance="cdb2.2_None">
    <SenseRelations/>
    <Semantics-verb>
      <semanticTypes semanticType="state"/>
    </Semantics-verb>
    <Pragmatics/>
    <SenseExamples>
      <SenseExample id="6495">
        <canonicalForm canonicalform="in een appartement/flat in de stad
        ↪  wonen" phraseType="vp" expressionType="freeCombination"/>
        <Syntax_ex>
          <combiWord lemma="appartement" partOfSpeech="noun"/>
          <combiWord lemma="flat" partOfSpeech="noun"/>
        </Syntax_ex>
        <Semantics_ex/>
        <Pragmatics/>
      </SenseExample>
      <SenseExample id="6496">
        ...
      </SenseExample>
      <SenseExample id="6497">
        ...
      </SenseExample>
      <SenseExample id="6636">
        ...
      </SenseExample>
    </SenseExamples>
  </Sense>
</LexicalEntry>
```

**Figure 4.5**
*Example of a Synset Entry in Open Dutch WordNet*

```
<Synset id="odwn-10-109912332-v">
  <Definitions>
    <Definition gloss="huizen" language="nl" provenance="odwn"/>
    <Definition gloss="houses" language="en"
    ↪  provenance="google-translate"/>
  </Definitions>
  <SynsetRelations>
    <SynsetRelation provenance="pwn" relType="has_hyperonym"
    ↪  target="eng-30-02637202-v"/>
    <SynsetRelation provenance="odwn" relType="involved_location"
    ↪  target="eng-30-03259505-n"/>
    <SynsetRelation provenance="odwn" relType="involved_location"
    ↪  target="eng-30-03546340-n"/>
    <SynsetRelation provenance="odwn" relType="involved_instrument"
    ↪  target="eng-30-03544360-n"/>
  </SynsetRelations>
</Synset>
```

are considered to be synonymous from either an intra-language or an inter-language perspective. For example, the lexical entry *wonen-v-1* is connected to a synset (*odwn-10-109912332-v*) indexed in the Dutch WordNet (v1.0, see prefix) only. The entry can only be linked to intra-lingual synonyms. By contrast, the lexical entry *woning-n-1* is linked to a synset (*eng-30-03259505-n*) originating from the original English WordNet (v3.0, see prefix) and is indexed in the Open Multilingual WordNet (OMW). As a result, the entry can be linked to synonyms in Dutch, English ('home'), French (*maison*, *demeure*), and other available languages.

To sum up, each word-sense disambiguated token was labeled with three different sense identifiers:

1. the sense class (e.g., *r_v-10183*), the output of the WSD classifier;
2. the sense ID (e.g., *wonen-v-1*), referring to the disambiguated lexical entry in ODWN; and
3. the synset ID (e.g., *odwn-10-109912332-v*), referring to the semantic concept in ODWN or OMW.

**Figure 4.6**
*Fallback to Two Types of Sense Identifiers*

| LEMMA | POS | SENSE ID | SYNSET ID |
|---|---|---|---|
| wonen | WW() | wonen-v-1 | odwn-10-109912332-v |
| overduidelijk | ADJ() | obvious.a.01 | eng-30-01618053-a |
| overbodig | ADJ() | d_a-415574 | - |

INCREASING COVERAGE WITH ODWN On closer inspection, it was evident that not all sense classes in the WSD classifier's dictionary were linked to an entry in ODWN. Only 52,430 senses (45%) were mapped to a lexical entry or synset ID. Therefore, it was necessary to increase coverage by correcting all missing information. First, the WSD system was extended to retrieve the sense and synset IDs for all lemmatized tokens with a monosemous lexical entry. Because this entry had only one possible sense, it did not have to be semantically disambiguated. Next, for all semantically ambiguous entries, the system used the most probable sense class determined by the classifier (cf. *supra*). However, in the absence of an ODWN sense or synset ID, the fallback procedure illustrated in Figure 4.6 was applied. First, in the absence of an existing lexical entry in ODWN for a given synset ID, the empty sense ID was instead set to the lexical entry indexed in NLTK's (Bird et al., 2009) OMW (e.g., *overduidelijk*, *obvious.a.01*). Next, in the absence of any lexical entry in ODWN, the original DutchSemCor sense class given by the classifier (e.g., *overbodig*, *d_a-415574*) was kept. This was done for the sake of completeness and future compatibility. In total, the system disambiguated 76% of all distinct lexical units (adjectives, adverbs, nouns, and verbs).

*Lexical Entries*

The second step in the resource development process was to define the list of entries to be included in the resource. The following information represented each lexical unit: the lemma; the part of speech; the sense ID, either from ODWN, NLTK's OMW, or DutchSemCor; and the synset ID. The list of lexical

units was corrected and filtered at two levels. The first filtering applied to the lemma.

- All non-alphanumeric tokens (e.g., punctuation marks, Arabic numerals, etc.) were excluded.
- All non-standard forms and abbreviations commonly found in Dutch chatspeak were also ruled out.
- Similar alphanumeric numbers (e.g., *4de, 5de*, '4th, 5th') were simplified to the same lexical entry *[digit]de*.
- All compounds with an optional parenthesized stem were split into two. For instance, *(studie)keuze* ('(study) choice') was tallied as two separate lexemes, namely *keuze* ('choice') and *studiekeuze* ('study choice').
- The omission of shared stems in coordinated compounds was also resolved. For example, a coordinated compound with a shared stem (underlined) *binnen- en buitenland* ('home and abroad') was resolved into *binnenland en buitenland*. These corrections were done with the help of a rule-based compound splitter for Dutch.[44]

The second filtering applied to the part of speech.

- The CGN tagset used by the Frog tagger is quite extensive: the tagset counts over 320 tags, accounting for a number of detailed lexical and morphological attributes. Since it would be irrelevant to keep all of these precise attributes in the resource, the tagset was simplified to 37 tags (see Table 4.3).
- All entries with a part of speech not covered by the simplified tagset (e.g., special symbols) were filtered from the resource.
- The multi-word units detected by Frog were not tagged with a specific part of speech, but with a "multi-tag" part of speech (e.g., *door en door*, `VZ(fin)_VG(neven)_VZ(fin)`, 'through and through'). In this case, each one of the individual tags were also converted according to the simplified tagset.

44 The software is publicly available from the following link: http://ilps.science.uva.nl/resources/compound-splitter-nl/.

**Table 4.3**
*List of Simplified CGN Tags*

| Simplified tag | Part of speech | # 37 |
|---|---|---|
| `N(soort/eigen)` | noun (common/proper) | # 2 |
| `ADJ()` | adjective | # 1 |
| `WW()` | verb | # 1 |
| `TW(hoofd/rang)` | numeral (card./ord.) | # 2 |
| `VNW(...)` | pronoun | # 20 |
| `LID(bep/onbep)` | article (def./indef.) | # 2 |
| `VZ(init/fin/versm)` | preposition (initial/final/fused) | # 3 |
| `VG(neven/onder)` | conjunction (coord./subord.) | # 2 |
| `BW()` | adverb | # 1 |
| `TSW()` | interjection | # 1 |
| `SPEC(deeleigen)` | part of proper noun | # 1 |
| `LET()` | punctuation | # 1 |

*Lexical Frequencies*

The third and final step in the resource development process was to compute frequency statistics for each lexical entry as it occurred across all five CEFR levels in the corpus. The raw frequencies of occurrence of an individual lexical entry at a particular level were adjusted to take into account the degree of dispersion of that entry across documents at that given level.These adjusted (normalized and standardized) frequencies were computed with the formulae proposed by Carroll et al. (1971) and which are given below (Definitions 4.1 to 4.3). In the following formulae, $N_{\text{level}}$ is the number of words in the level and $n_i$ is the number of words in document $i$ of all documents $d$ in that level.

**Definition 4.1: raw frequency**

The raw frequency

$$F_{\text{entry,level}} = \sum \mathbf{f} = \sum_{i=1}^{d} f_i \tag{4.1}$$

is the number of times the entry occurs at a particular level. This amounts to summing the vector $\mathbf{f}$ of the entry's frequency of occurrence $f$ in document $i$ for all $d$ documents in that level.

**Definition 4.2: dispersion**

The dispersion index

$$\text{Carroll's } D = \left[\ln(\sum \mathbf{f}) - \left(\left[\sum_{i=1}^{d} f_i \ln f_i\right] / \sum \mathbf{f}\right)\right] \cdot \frac{1}{\ln d} \tag{4.2}$$

measures the extent to which the entry is scattered across documents at a particular level. The values range from 0 (*undispersed*) to 1 (*fully dispersed*).

**Definition 4.3: normalized frequency**

The adjusted frequency

$$\text{Carroll's } U = \frac{10^6}{N_{\text{level}}}\left[F \cdot D + (1 - D) \cdot \left(\frac{1}{N_{\text{level}}} \sum_{i=1}^{d} f_i \cdot n_i\right)\right] \tag{4.3}$$

normalizes the raw frequencies per one million tokens, based on the degree of dispersion across documents. When $D > 0$, the influence of $F$ increases. When $D = 0$, the raw frequencies are not taken into account. Instead, the normalized frequencies are drawn from a weigthed frequency distribution.

**Definition 4.4: standard frequency index**

The standard frequency index

$$SFI = 10 \cdot \left[\log_{10}(U) + 4\right] \tag{4.4}$$

**Table 4.4**
*Number of Lexical Entries in NT2Lex*

| NT2Lex version | CGN | CGN+ODWN |
|---|---|---|
| All entries | 15,227 | 17,743 |
| Lexical entries | 14,368 | 16,884 |
| Grammatical entries | 400 | 400 |
| Multi-word entries | 459 | 459 |

standardizes the normalized frequencies to a scale of 0 to 100. A value of 100, 90, 80, ..., 40 on the standard scale indicates that the lexical entry occurs, respectively, once every $10^0$, $10^1$, $10^2$, ..., $10^6$ entries.

### 4.2.2 *Resource Description*

The resource development process described in the previous section resulted in creating the NT2Lex resource, compiled in two versions. The first version of the resource (viz., CGN) was similar to the previously developed graded lexicons in that it contained adjusted frequencies computed only for the lemmatized and part-of-speech tagged entries (without considering the disambiguated word senses). The second version of the resource (viz., CGN+ODWN) includes adjusted frequencies computed for the word-sense disambiguated entries. An excerpt from this complete version of the resource is given in Figure 4.7. For a comparative overview of the number of entries in both versions of NT2Lex, see Tables 4.4 and 4.5.

*NT2Lex Version CGN*

The basic version of NT2lex counts 15,227 entries, mostly content words (see Table 4.5). The size in number of total entries is similar to the size of the resources developed for Swedish (SVALex; $N = 15681$) and English (EFLLex; $N = 15281$). As for French, the total number of entries is considerably higher

(FLELex; $N$ = 17871), which is because the resource for French was compiled from a much larger corpus.

Similar comparisons can be made regarding the number of new entries per each CEFR level. Table 4.5 shows that may novel words are introduced at the A1 (100%) and A2 (87%) levels. This number decreases at the intermediate levels: slightly more or less than half of the entries are novel words, both for the B1 (57%) and B2 (45%) levels. At the advanced C1 level, less than one-fourth (22%) of entries are novel words. This decreasing trend is very similar to the resources developed for Swedish and English. For French, however, the decreasing trend is much more marked: from the A2 level onwards, less than half of the entries are novel, with almost no new entries appearing at the C2 level. Again, this difference is likely because the resource for French was compiled on a corpus with more A1-level texts.

However, there is a striking difference in the number of multi-word entries included in NT2Lex and the other resources. Only 459 of the entries are multi-word units, contrary to 2,038 for French and 1,450 for Swedish. The multi-word units that are included in the resource mainly pertain to well-known named entities (e.g., *Olympische Spelen*, 'Olympic Games'), phrasal verbs (e.g., *voorzien van*, 'to provide'), and adverbs (e.g., *om het even*, 'all the same'). This difference could be explained by the fact that the majority of compounds are agglutinative in Dutch (e.g., *afvalverwijderingsstructuur*, 'waste disposal structure'), whereas compound words can be multi-word units in other languages. Among all single-word entries in NT2Lex, 4,431 (31%) are compound words. As for the Swedish language, where the compounding system is similar to Dutch, this disparity could be attributed to the fact that different taggers were used to detect multi-word units.

*NT2Lex Version CGN+ODWN*

The word-sense disambiguated version of NT2Lex includes an extra 2,516 entries, which amounts to 17,743 entries. Table 4.6 shows that 1,454 entries are polysemous (i.e., having at least two senses). The most polysemous entry in the resource is the entry *pakken* (verb, 'to take', 'to grab', 'to defeat', 'to hinder', etc.), which has 10 different senses that are attested in the resource. Although all of these polysemous entries are lexical ones, it should be noted

**Figure 4.7**

*Illustration of Entries in NT2Lex with Adjusted Frequencies (CGN+ODWN Version)*

| word | tag | sense_se-id | sense_sy-id | U@A1 | U@A2 | U@B1 | U@B2 | U@C1 | U@TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| in zwang | VZ(init) N(soort) | in_zwang-n-1 | eng-30-14411884-n | - | - | - | - | 22 | 0 |
| omgangstaal | N(soort) | omgangstaal-n-1 | eng-30-07157123-n | - | - | - | 26 | - | 3 |
| pakken | WW() | pakken-v-1 | odwn-10-101230891-v | 35 | 117 | 101 | 5 | - | 99 |
| pakken | WW() | pakken-v-10 | eng-30-01100145-v | - | 51 | 12 | - | - | 28 |
| zijn | WW() | zijn-v-1 | eng-30-02603699-v | 2094 | 1647 | 1423 | 1253 | 1335 | 1601 |

**Table 4.5**

*Number of Lexical Entries per Level in NT2Lex*

| | CGN | | | | | | CGN+ODWN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Entries | New | New % | Multi-Word | Hapaxes | $F > 10$ | Entries | New | New % | Hapaxes | $F > 10$ |
| A1 | 953 | 953 | 100 % | 70 | 313 | 225 | 1,189 | 1,189 | 100 % | 427 | 228 |
| A2 | 6,220 | 5,383 | 87 % | 1,224 | 2,482 | 1,231 | 7,630 | 6,580 | 86 % | 3,073 | 1,386 |
| B1 | 8,559 | 4,879 | 57 % | 1,997 | 3,936 | 1,081 | 10,160 | 5,571 | 55 % | 4,739 | 1,128 |
| B2 | 8,172 | 3,641 | 45 % | 1,861 | 4,362 | 638 | 9,366 | 3,998 | 43 % | 5,092 | 619 |
| C1 | 1,680 | 371 | 22 % | 252 | 1,127 | 63 | 1,841 | 405 | 22 % | 1,282 | 62 |

**Table 4.6**
*The Number of Word Senses, Polysemes, and Unique Synsets in NT2Lex*

| Levels | A1 | A2 | B1 | B2 | C1 | All |
|---|---|---|---|---|---|---|
| Entries | 1,189 | 7,630 | 10,160 | 9,366 | 1,841 | 17,743 |
| Senses | 849 | 5,705 | 7,272 | 6,517 | 1,302 | 11,999 |
| Polysemes[a] | 139 | 828 | 979 | 771 | 118 | 1,451 |
| Synsets | 658 | 4,450 | 5,465 | 4,936 | 1,046 | 8,934 |

[a] entries with more than one sense

that some multi-word entries have also been disambiguated for word senses. Nevertheless, none of these multi-word units were polysemous. Furthermore, the word-sense disambiguated version of NT2Lex contains many semantic concepts, with 8,934 distinct synsets out of 11,999 word senses.

### 4.2.3 *Web-Based Tools for Cross-lingual Search and Lexical Complexity Analysis*

A web-based system was developed for querying both versions of the NT2Lex resource.[45] The system was designed as an Angular Material Design UI connected to an object-oriented RESTful backend developed in the Django framework. Documentation of the primary entry points to the RESTful API is given in Section 4.A.1. Although the system was specifically developed for the first publication of NT2Lex (Tack et al., 2018b), its architecture was independent of the specific resource format. As a result, the resources developed for other languages could be added to the same database, which enabled to perform cross-lingual search queries of an entry in NT2Lex with an entry in the other languages (Figure 4.8).

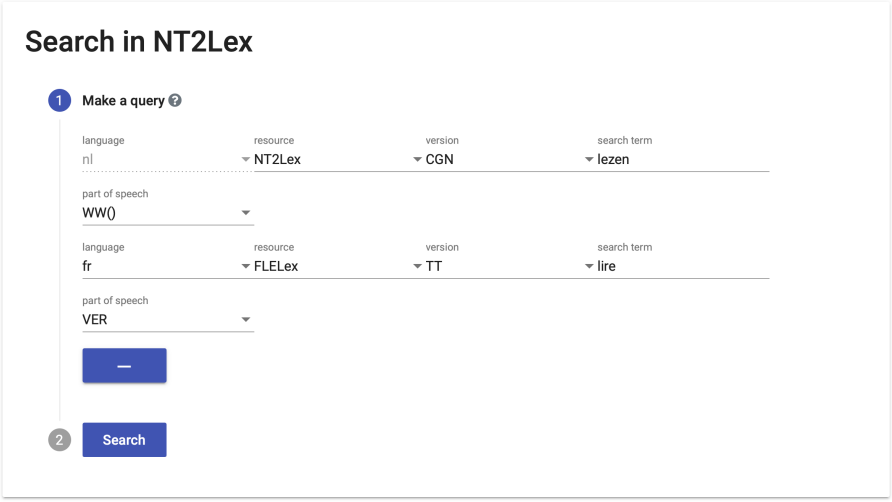
Also, the system included a tool for the annotation of difficulty (Figure 4.9) with the method previously developed for French (Tack et al., 2016a; see Chapter 3). Figure 4.9 gives an illustration of a lexical complexity analysis for the A1 level. All words which were estimated difficult for the A1 level were highlighted. Similarly, all out-of-vocabulary (OOV) words (i.e., which never

45 http://cental.uclouvain.be/nt2lex/

**Figure 4.8**

*Screenshot of an Online Crosslingual Search Query in NT2Lex*

**(a)** *Search Query*




**(b)** *Search Result*

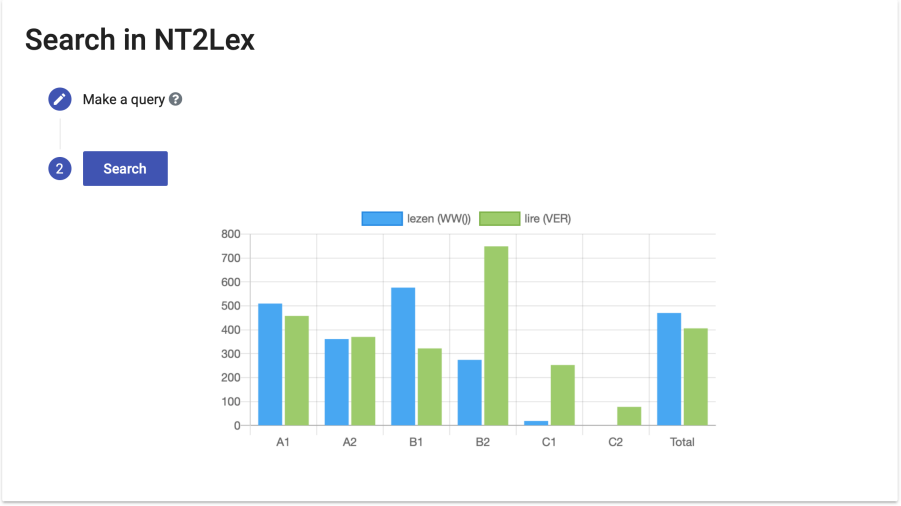
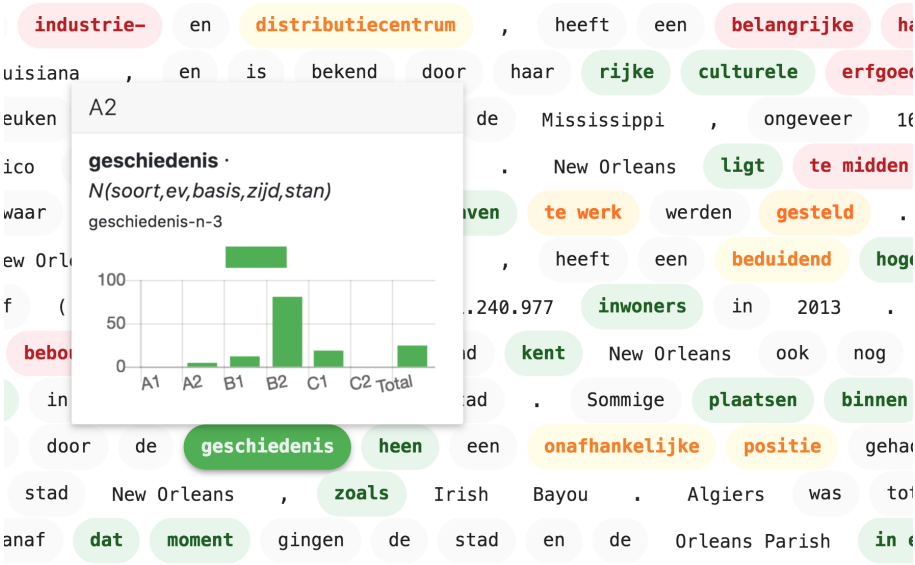


**Figure 4.9**

*Screenshot of a Lexical Complexity Analysis with NT2Lex*




*Note.* In this example, the threshold for identifying complex words was set at the A1 level. All highlighted words were labeled as complex (i.e., occurring in the resource beyond the targeted proficiency level).

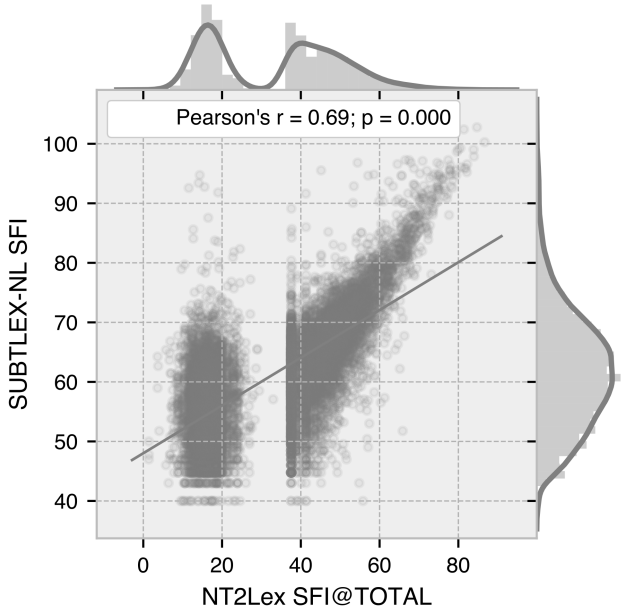
occurred in Dutch L2 graded textbooks and readers) were also considered difficult and highlighted (in red).

## 4.3 A COMPLEXITY ANALYSIS WITH LEXICAL FEATURES AND NORMS

Based on the resource described in the previous section, three analyses were conducted to answer the study's research questions. The first analysis investigated how well the distribution of (novel) entries per each level was associated with other well-known lexical features and norms (RQ4.1). In particular, the following features were examined: frequency and familiarity effects, semantic relations (polysemy and synonymy), and psycholinguistic norms (age of acquisition and concreteness). Unless otherwise specified, the analyses were performed on the most complete version of NT2Lex: the word-sense disambiguated CGN+ODWN version.

**Figure 4.10**

*Comparison of NT2Lex-CGN and SUBTLEX-NL Frequencies*




### 4.3.1 *Lexical Features*

First, the analyses focused on the frequency statistics computed for all lexical entries in NT2Lex, regardless of the level in which they appeared. The standardized frequency[46] values in NT2Lex-CGN[47] were contrasted with those attested in SUBTLEX-NL (Keuleers et al., 2010).[48] Figure 4.10 shows that a significant, large, and positive correlation was observed between the NT2Lex and SUBTLEX frequencies, $r = .69$, $p < .001$. This large correlation suggested that the estimated frequencies were very similar to those of a reference corpus, even though the resource had been compiled from a relatively small corpus.

Figure 4.11 shows the interplay between these standardized frequencies and the degree of dispersion. The density of dispersion values (marginal plot

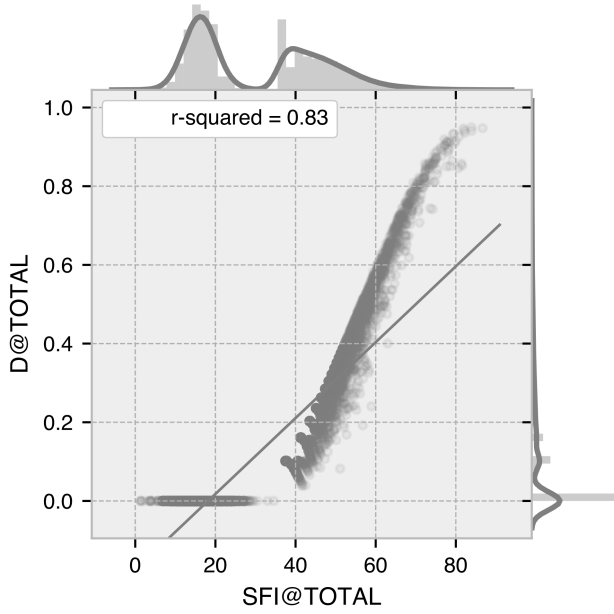
46 Because of the standardized scale, the SFI was preferred over Carroll's U.

47 For reasons of comparability, the non-WSD version of the resource was used because SUBTLEX-NL does not include frequencies for word senses.

48 SUBTLEX is a lexical database with word frequencies estimated from a reference corpus of film subtitles and is frequently used in psycholinguistic research (Brysbaert & New, 2009).

**Figure 4.11**

*The Interplay Between Frequency and Dispersion in NT2Lex*




on the right) shows that many entries (8,936 entries to be precise) had zero dispersion, *D* = 0. This finding indicated that 50% of all entries appeared in one text document and, consequently, could only occur in one CEFR level. What is more, the total number of entries occurring in one level amounted to 60% (10,635 entries) in the WSD version and 62% in the non-WSD version. Among entries appearing in at least two levels, 11% did not occur in consecutive CEFR levels (i.e., there were zero-occurrence gaps between levels). These observations were not only striking but also critical. Because there was a large sparsity in the distribution of lexical entries across CEFR levels, this implied that there would be a large sparsity in the distribution of lexical complexity.

Furthermore, Figure 4.11 shows that the adjusted frequencies for these zero-dispersion entries were drawn from a weighted frequency distribution (around *SFI* < 40, see the Gaussian-like density function in the marginal plot on top). Conversely, for all non-zero dispersion entries, a Zipfian distribution (Zipf, 1949) was observed in the standard frequency index (around *SFI* > 40, see also the marginal plot on top). Most words were located at the lowest frequency

ranks on the standard scale, whereas few words were located at the highest frequency ranks. Another well-known Zipfian effect is the association between word frequency and length. Unsurprisingly, word frequency was negatively correlated with word length: the shorter the word, the higher its frequency, $r = -0.39$, $p < .001$. In sum, the frequency distribution in NT2Lex seemed coherent with what one would expect to obtain from a language corpus. These results were taken as proof of the consistency of the resource.

*Lexical Frequency, Dispersion, and Sophistication per CEFR level*

Next, the analyses focused on the frequency distribution for all (new) entries attested per CEFR level. Figure 4.12 shows the average SUBTLEX-NL frequencies per each level. There was a decreasing trend in frequency from the A1 to the C1 levels, most noticeable for the new entries per level. When all entries were taken into account, the trend reached a plateau starting from A2 to the higher levels, except the C1 level.[49] When only the novel entries were taken into account, the frequency of word occurrence decreased rapidly as the difficulty level increased. Therefore, when learners are reading texts at higher proficiency levels, they encounter new (complex) words that are considerably less frequent in Dutch.

Like the trend observed above, the extent of lexical dispersion also decreased when the proficiency level increased. Figures 4.13a and 4.13b show a decreasing trend in respectively the average SUBTLEX-NL document frequencies (i.e., the number of different films in which the word occurred) and the NT2Lex dispersion indices for novel entries ranging from the basic A1 level to the advanced C1 levels. Therefore, when learners are reading texts at higher proficiency levels, they encounter novel words that are much less widespread.

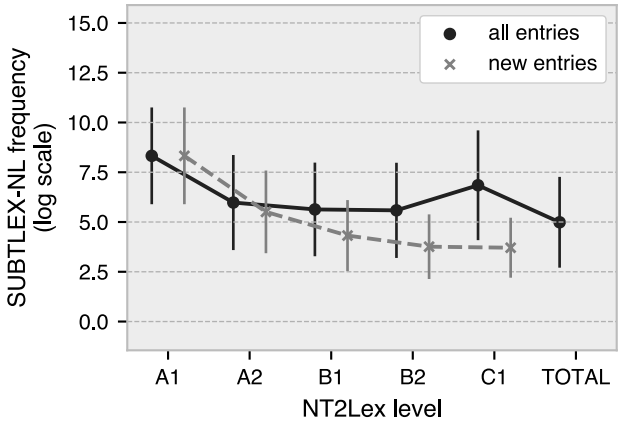
Lastly, lexical sophistication was measured as the absence from a list of 2,000 most familiar Dutch words according to the *Basiswoordenboek Nederlands* (de Kleijn & Nieuwberg, 1993).[50]. At the A1 level, almost 60% of entries appeared in the basic word list. Two ratios (see Lu, 2012) were computed to measure the

49 A possible reason for this increase in frequency at the C1 level is that the subcorpus was the most restricted in size due to the limited availability of C1-level texts. Because of this limitation, the previously introduced (basic and intermediate) entries were probably more predominant.

50 The vocabulary list is available from the following link: http://www.dikverhaar.nl/wp-content/uploads/Basiswoordenlijst_2000_frequente_meest_woorden.pdf It should be noted that 61 of the 2,000 basic word forms were not attested in NT2Lex.

**Figure 4.12**

*Mean SUBTLEX-NL Frequencies per Level in NT2Lex-CGN*




degree of lexical sophistication per level. The first ratio $LS_1 = W_{\text{sophisticated}}/W$ was the number of sophisticated entries over the total number of entries. The second ratio $LS_2 = V_{\text{sophisticated}}/V$ was the number of sophisticated verb entries over the total number of verb entries. There was an increasing trend in lexical sophistication at increasing proficiency levels, as shown in Figure 4.14.

*Semantic Relations*

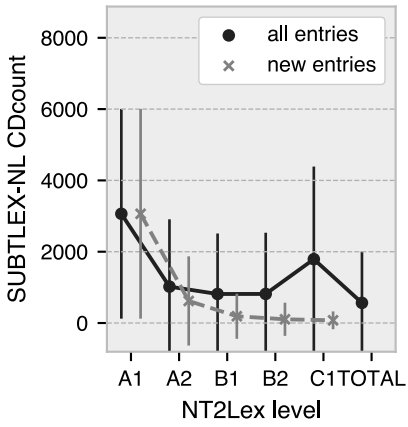
Figures 4.15a and 4.15b show the degree of polysemy and synonymy attested in NT2Lex and ODWN. The degree of polysemy was computed as the number of senses per lexical entry. The degree of synonymy was computed as the number of lexical entries referring to the same synset ID. Pearson product-moment correlation coefficients showed strong and positive associations between the degree of polysemy in NT2Lex and ODWN ($r = 0.66; p < .001$), on the one hand, and between the degree of synonymy in NT2Lex and ODWN ($r = 0.67; p < .001$), on the other hand.

However, NT2Lex displayed a much lower extent of onomasiological variation (i.e., meaning-to-forms mappings) as there were considerably fewer entries referring to the same concept. The limited range of lexicalizations for each concept was due to the specialized nature of the resource. As for

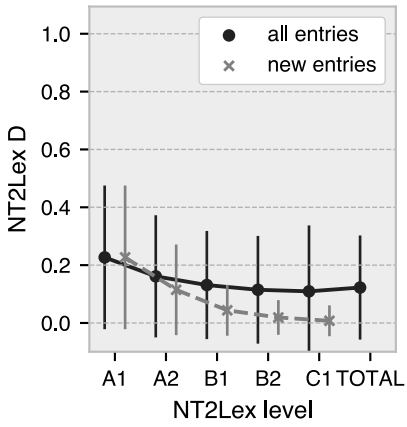
**Figure 4.13**

*Lexical Dispersion per Level in NT2Lex*

**(a)** *SUBTLEX-NL Document Frequency (M and SD)*




**(b)** *NT2Lex Dispersion (M and SD)*




semasiological variation (i.e., form-to-meanings mappings), Figure 4.15c shows a decreasing trend in the degree of polysemy per level. At the elementary levels, the novel lexical stock contained more polysemous, ambiguous entries. In contrast, at the advanced levels, new entries displayed a low dispersion and frequency of occurrence and seemed to have more well-defined meanings.

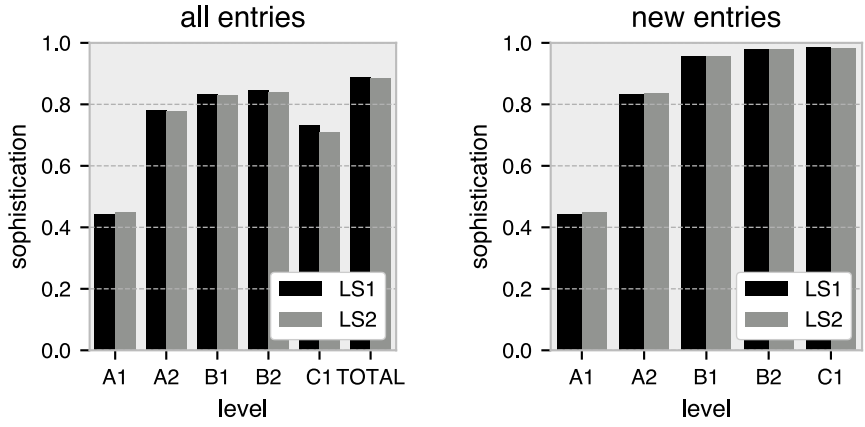
### 4.3.2 *Psycholinguistic Norms*

The analyses also examined whether the degree of difficulty in NT2Lex was associated with several psycholinguistic norms for Dutch. The analyses focused on the age of acquisition (AoA)[51] and concreteness norms established by Brysbaert et al. (2014).

51 It should be noted that the age of acquisition norms do not give a measure of the actual age of acquisition at which words are acquired by native speakers. Instead, the values represent the average age at which Dutch L1 speakers believe the word to be acquired.

**Figure 4.14**

*The Degree of Lexical Sophistication Across Levels in NT2Lex*




*Age of Acquisition*

Figure 4.16a shows that highly-dispersed words in L2 reading materials were also acquired earliest by Dutch L1 speakers. A Pearson product-moment correlation showed a moderate and negative association between dispersion and age of acquisition, $r = -0.47$, $p < .001$. In other words, words acquired earliest were also present in most texts intended for L2 readers, all proficiency levels combined. Regarding the differences between proficiency levels, Figure 4.17 shows that early-acquisition words were much more prevalent at elementary levels, while late-acquisition words were more prevalent at advanced levels. The novel words in A1 and A2-level reading materials were acquired earliest, respectively, between five and seven and a half. Conversely, new words at the intermediate (B1/B2) and advanced (C1) levels were acquired later, approximately at the ages of 10 and 11. Moreover, Figure 4.16b shows that early-acquisition words were significantly more concrete, while late-acquisition words were significantly more abstract, $r = -0.41$, $p < .001$.

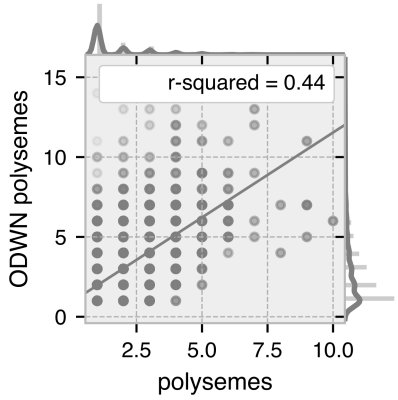
*Concreteness*

Figure 4.18 shows the distribution of concreteness norms across NT2Lex levels, ranging from 1 (*very abstract*) to 5 (*very concrete*). The highest levels (B2/C1)

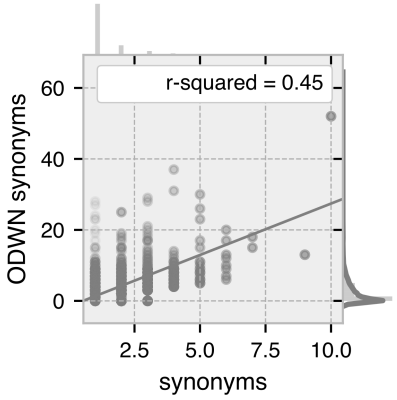
**Figure 4.15**

*Polysemy and Synonymy in NT2Lex*

**(a)** *Polysemes per Entry*




**(b)** *Synonyms per Entry*




**(c)** *Number of Polysemes per Entry and per Level (M and SD)*

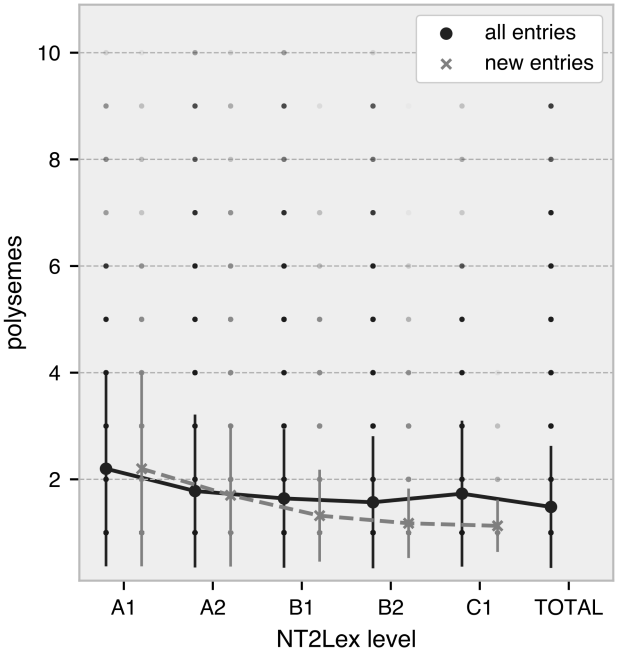
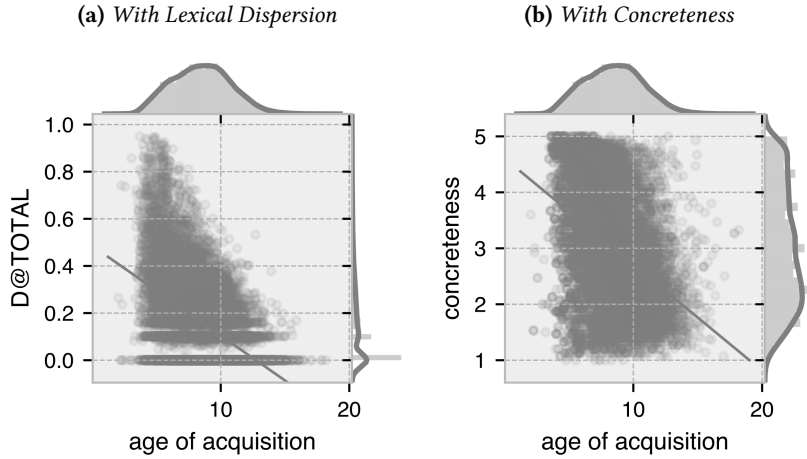


**Figure 4.16**

*Age of Acquisition in NT2Lex*




contained a considerably higher proportion of abstract or less concrete words. New words in reading materials intended for more proficient learners were less concrete and more abstract. By contrast, the most basic levels contained a higher proportion of concrete words. However, the differences between levels seemed less clear-cut than the differences observed for AoA.

*Conclusion for RQ4.1*

In sum, the results showed that the degree of difficulty attested in NT2Lex was coherent with some well-established effects. Firstly, the decreasing trend of frequency at increasing difficulty levels was coherent with general frequency effects observed in L2 research (Ellis, 2002). Secondly, the decreasing trend of polysemy at increasing difficulty levels seemed to reflect the general finding that lower latencies in form recognition tend to be observed for more polysemous words (Millis & Button, 1989). Finally, the decreasing trend of word concreteness at increasing difficulty levels also reflected previous findings on L2 speaker's language development. For instance, Crossley et al. (2009), showed that productive vocabulary became more abstract (as measured from

**Figure 4.17**

*Density of Age of Acquisition per CEFR Level in NT2Lex*

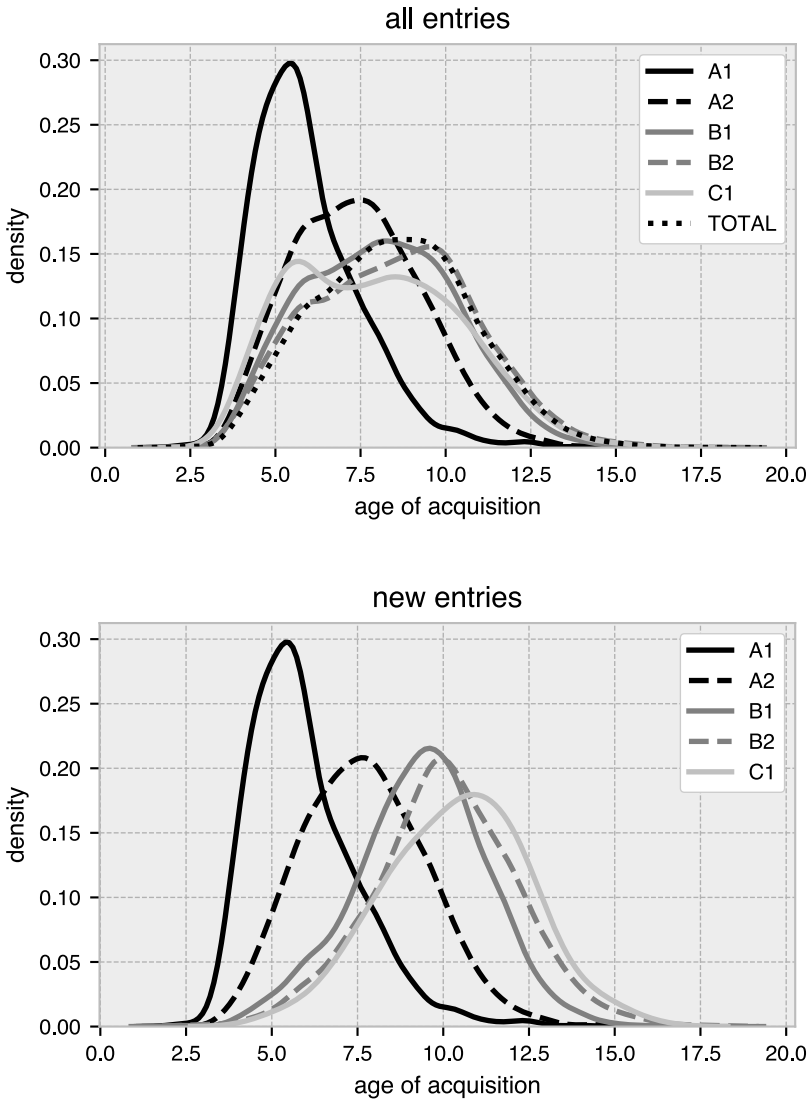
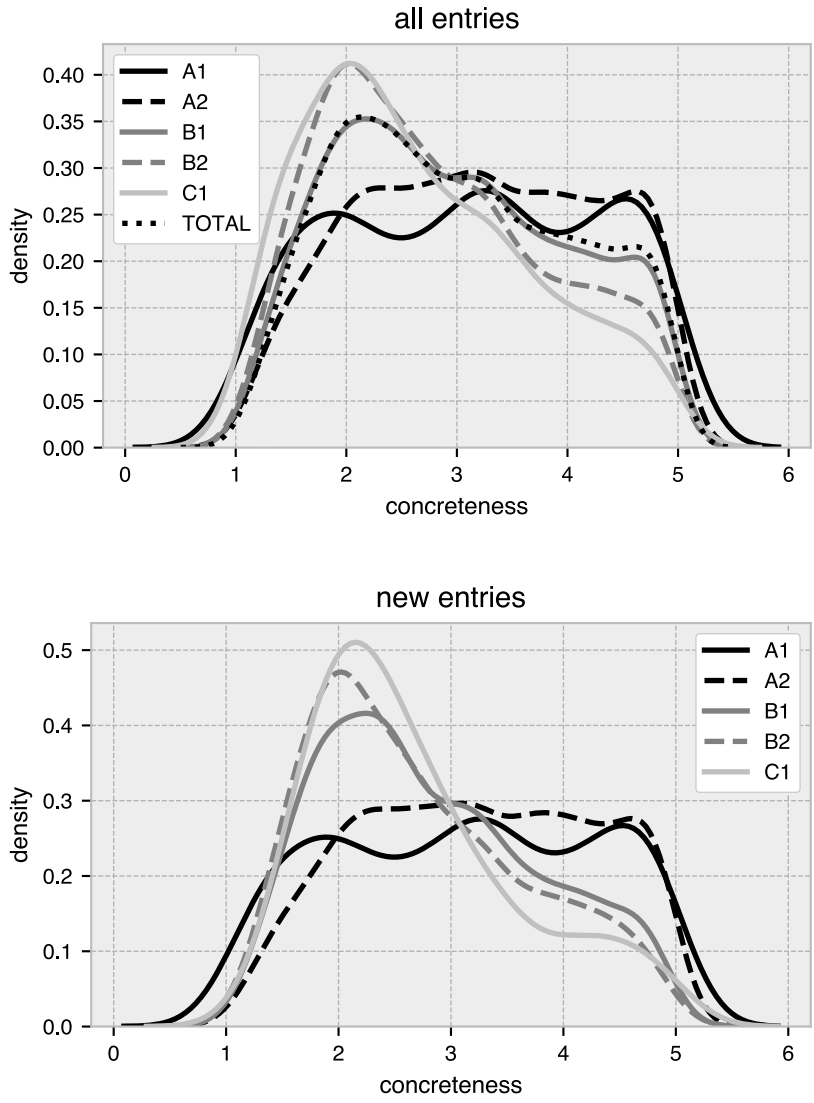


**Figure 4.18**

*Density of Word Concreteness per CEFR Level in NT2Lex*



MRC concreteness norms) as learners spent more time learning. Hypothesis H4.1, therefore, seemed to be confirmed.

## 4.4 AVERAGED AND DISAMBIGUATED SEMANTIC COMPLEXITY

The second analysis investigated whether word-sense disambiguation enhanced the computation of semantic complexity (RQ4.2). The analysis focused on the WSD version of NT2Lex and one semantic feature, namely the word's rank in the WordNet hypernymy hierarchy. To improve the computation of this semantic feature, the analysis compared word-sense averaged and disambiguated features to determine which lead to (a) more substantial discrimination between CEFR levels and (b) a more accurate computation.

### 4.4.1 *Rank in the WordNet Hypernymy Tree*

As introduced in Section 4.1.2, the WordNet hypernymy hierarchy, illustrated in Figure 4.19, provides an essential measure of semantic complexity. In this hypernymy tree, the various concepts $c$ (or synset nodes in WordNet) are structured according to their depth $d$ and height $h$ in the hierarchy. The tree's root concept (i.e., *entity.n.01*), which has the lowest depth $d(c_{entity.n.01}) = 0$ and the highest height $h(c_{entity.n.01}) = 19$, corresponds to the most all-encompassing, superordinate hypernym. Conversely, the leaf nodes correspond to the most subordinate concepts or hyponyms such as *gig.n.04*, which has the lowest possible height $h(c_{gig.n.04}) = 0$. A measure of semantic complexity can then be computed by looking at the rank of a concept (or synset node) in the longest path in the tree, which goes either from a particular node to the root node. Besides this standard hypernymy rank, two other ranks are proposed as well: the word's hyponymy rank and relative hypernymy rank.

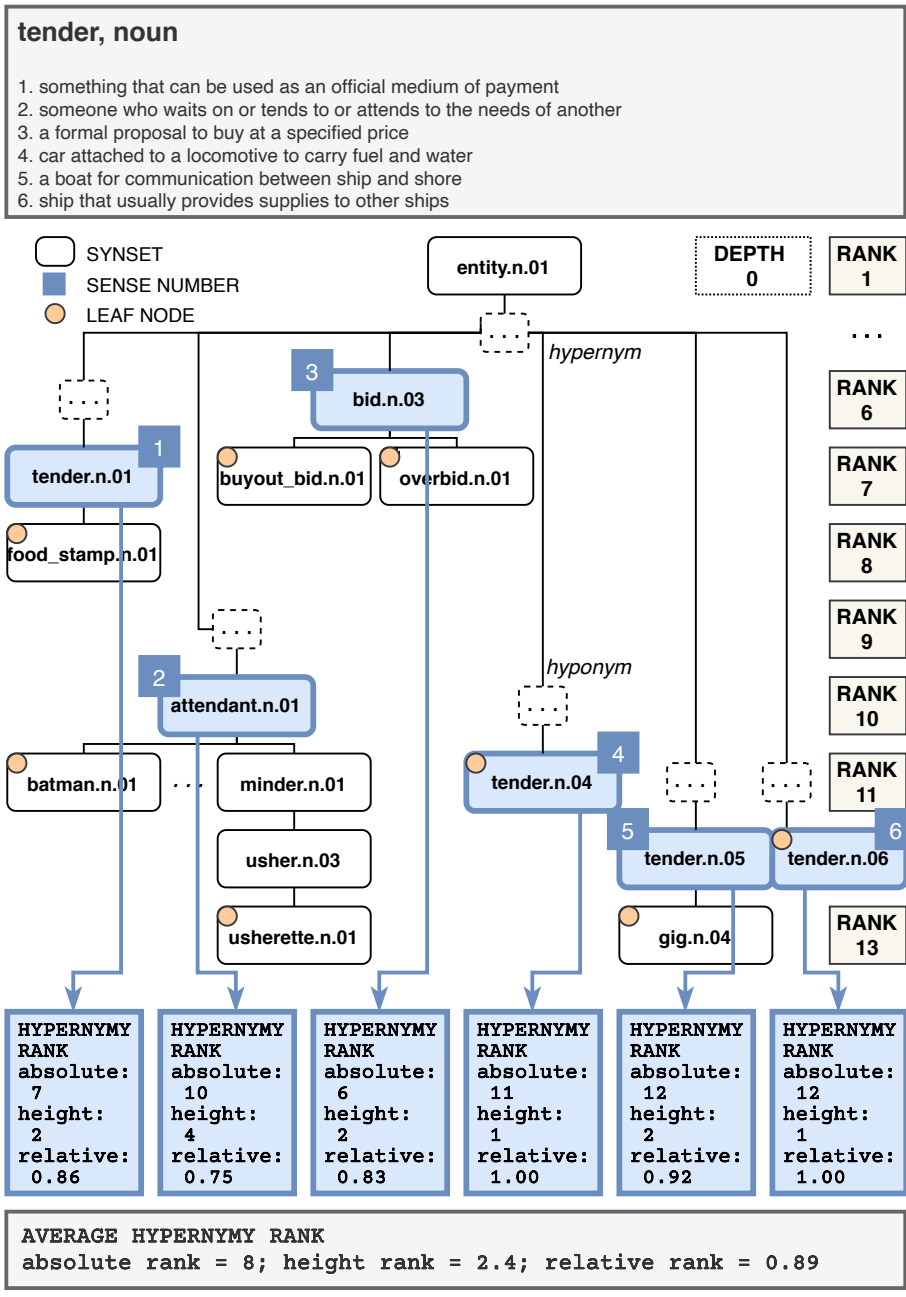
*Hypernymy Rank*

The hypernymy rank of a concept

$$HR(c) = d(c) + 1 \tag{4.5}$$

**Figure 4.19**

*Computation of Absolute, Relative, and Averaged WordNet Hypernymy Ranks*



was computed based on its absolute depth $d$ in the hierarchy (see Figure 4.19). This rank corresponded to the standard hypernymy feature used in Coh-Metrix (see Crossley et al., 2009; Crossley et al., 2007). Following Crossley et al. (2009, p. 318), the root hypernym *entity.n.01* was given the highest rank $HR(c_{entity.n.01}) = 1$, whereas more subordinate nodes located further down the tree (e.g., *gig.n.04*) were given a much lower rank, $HR(c_{gig.n.04}) = 13$. The interpretation of these ranks was as follows: the higher-rank nodes represented the most general concepts (i.e., basic-level categories), whereas the lower-rank nodes represented the least general concepts. In what follows, this rank will be referred to as the *absolute hypernymy rank*.

*Hyponymy (or Height) Rank*

As the name suggests, the height rank of a concept

$$HR_h(c) = h(c) + 1 \tag{4.6}$$

was computed based on its absolute height $h$ in the hierarchy. It was the opposite of the absolute hypernymy rank: the root hypernym *entity.n.01* was given the lowest rank $HR_h(c_{entity.n.01}) = 20$ whereas its subordinate, leaf hyponym *gig.n.04* was given the highest rank $HR_h(c_{gig.n.04}) = 1$. The interpretation of these ranks was as follows: the higher-rank nodes represented the most specific concepts, whereas the lower-rank nodes represented the least specific concepts. So, whereas the absolute hypernymy rank reflected varying degrees of conceptual *genericity*, the height or hyponymy rank was reflected varying degrees of conceptual *specificity*.

It should be noted that these two ranks are not commensurate with each other. In Figure 4.19, for instance, the least specific sense of the word *tender* is *attendant.n.01* according to its height in the hypernymy tree, but conversely it is not the most generic sense according to its depth in the hypernymy tree. Therefore, hypernymy and hyponymy rank cannot be used interchangeably.

*Relative Hypernymy Rank*

To take into account both genericity and specificity, that is, both the depth and height in the hypernymy tree, the relative hypernymy rank

$$HR_r(c) = \frac{d(c)}{d(c) + h(c)} \tag{4.7}$$

was also examined in this study. The relative rank value was computed by normalizing the position (absolute depth) of the node in the full hypernymy path ($d + h$), which ranged from the longest path to the root hypernym (depth $d$) to the longest path to a leaf hyponym (height $h$). The root hypernym *entity.n.01* was given the lowest possible value $HR_r(c_{entity.n.01}) = 0.0$, whereas its subordinate, leaf hyponym *gig.n.04* was given the highest possible value $HR_r(c_{gig.n.04}) = 1.0$. The interpretation of these values was as follows: nodes with a value close to 0 represented the most general concepts, whereas nodes with a value close to 1 represented the most specific concepts. Isolated nodes (i.e., with no hypernyms and no hyponyms) received the middle rank of 0.5.

*Averaging vs. Disambiguation*

The computation of these three ranks in the hypernymy tree has an important condition: the specific sense of a word needs to be known beforehand. However, this is not always possible for most complexity analyses. A solution is to average the rank over the set of all possible concepts $C$ for word $w$

$$HR_a(w) = \frac{1}{|C_w|} \sum_{i}^{|C_w|} HR(c_i) \tag{4.8}$$

but there may be a drawback to this word-sense averaging.

In Figure 4.19, for instance, the average hypernymy rank would probably overestimate the complexity of the word *tender* when it is used in its more usual sense of a bid. Because *bid.n.03* has the highest rank $HR(c_{bid.n.03}) = 6$, it is semantically more generic and hence hypothetically easier than the other five meanings of the word. Therefore, using the average rank $HR_a(w_{tender}) \approxeq 8$ would underestimate the concept's genericity and, consequently, overestimate its complexity.

On the other hand, the averaged hypernymy rank would underestimate the complexity of the word *tender* when it refers to a specific type of ship (i.e., sense numbers five and six). Because both concepts *tender.n.05* and *tender.n.06* have the lowest hypernymy rank $HR(c_{tender.n.05}) = HR(c_{tender.n.06}) = 12$, they are semantically less generic and hence hypothetically more difficult than the other meanings of the word. Therefore, using the average rank $HR_a(w_{tender}) \approxeq 8$ would overestimate this concept's genericity and, consequently, underestimate its complexity.

In sum, it is essential to ask whether the application of word-sense disambiguation leads to a more precise measure of semantic complexity.

### 4.4.2 *Semantic Complexity Across Levels in NT2Lex*

Multiple statistical analyses were run to compare hypernymy ranks computed with word-sense averaging and word-sense disambiguation. The three hypernymy ranks described above (viz., absolute hypernymy rank, hyponymy rank, and relative hypernymy rank) were computed for all new entries occurring at each level in NT2Lex (WSD version). For each entry with a disambiguated sense, the rank for the entry's precise sense was compared with the average rank for all possible senses of that entry. Figure 4.20 shows the distribution of these averaged and disambiguated hypernymy ranks.

The statistical analyses examined the following basic assumption: new words occurring in reading materials at higher difficulty levels should display a greater semantic complexity. This increase in semantic complexity was expected to occur as follows:

1. Concept genericity (i.e., hypernymy rank) should decrease from levels A1 to C1.

2. Concept specificity (i.e., hyponymy rank) should increase from levels A1 to C1.

3. The relative position in the hierarchy should evolve from 0 (*most generic*) at the A1 level towards 1 (*most specific*) at the C1 level.

**Figure 4.20**

*Wilcoxon Pair-Wise Comparisons of Hypernymy Ranks in NT2Lex Before and After Disambiguation*

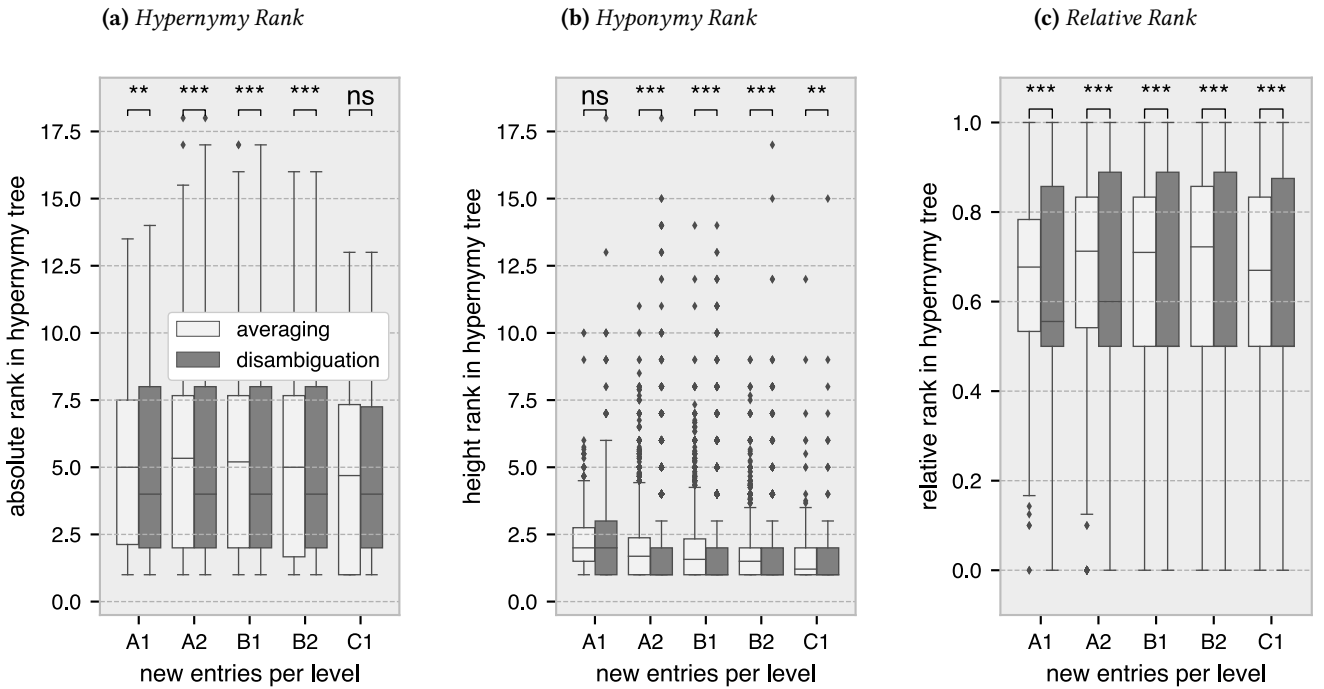


**Table 4.7**

*Median Values of Hypernymy Ranks for Novel Entries at Each Level in NT2Lex*

| | | Hypernymy | | Hyponymy | | Relative | |
|---|---|---|---|---|---|---|---|
| Level | *N* | AVG | WSD | AVG | WSD | AVG | WSD |
| A1 | 693 | 5.00 | 4.00 | 2.00 | 2.00 | 0.68 | 0.56 |
| A2 | 4,344 | 5.33 | 4.00 | 1.69 | 1.00 | 0.71 | 0.60 |
| B1 | 3,149 | 5.20 | 4.00 | 1.57 | 1.00 | 0.71 | 0.50 |
| B2 | 2,024 | 5.00 | 4.00 | 1.50 | 1.00 | 0.72 | 0.50 |
| C1 | 208 | 4.69 | 4.00 | 1.21 | 1.00 | 0.67 | 0.50 |

AVG = word-sense averaging WSD = word-sense disambiguation

Table 4.7 lists the median hypernymy, hyponymy, and relative ranks per each level in NT2Lex.[52] Two critical observations were made. Firstly, the results were contrary to what was expected for the absolute hypernymy rank: there was no apparent decrease in genericity from A1 to C1. Conversely, there was an increase in specificity from A1 to C1 for hyponymy rank and, to a lesser degree, relative rank. At first sight, these results seemed to corroborate the assumption that advanced-level texts contained more specific vocabulary than basic-level texts. Secondly, while there was a distinct increase in semantic complexity for averaged ranks, the same trend could no longer be observed when disambiguated ranks were considered. The significance of these initial observations was further examined with statistical testing.

*Averaged Semantic Complexity Features*

Significant differences were observed between CEFR levels for all three ranks in the hierarchy, as assessed by multiple Kruskal-Wallis tests (Table 4.8). Pairwise two-sample Wilcoxon comparisons between levels were computed with the DSCF test (Critchlow & Fligner, 1991; Dwass, 1960; Steel, 1960). The

52 Because there was no difference in median ranks for word-sense disambiguation, the mean ranks were also computed. However, seeing that the distributions were very skewed (see Figure 4.20) and, consequently, the normality assumption was violated, it would have been misleading to compare these mean values statistically. Therefore, the mean ranks were included in the Appendix (Section 4.A.2) for documentation purposes.

**Table 4.8**

*Kruskal-Wallis Tests of Differences in Hypernymy Ranks Across CEFR Levels*

| | $\chi^2$ | df | $p$ | $\epsilon^2$ |
|---|---|---|---|---|
| WORD-SENSE AVERAGING | | | | |
| Hypernymy rank | 13.7 | 4 | .008** | 0.001 |
| Hyponymy rank | 205 | 4 | <.001*** | 0.020 |
| Relative rank | 24.7 | 4 | <.001*** | 0.002 |
| WORD-SENSE DISAMBIGUATION | | | | |
| Hypernymy rank | 7.37 | 4 | .118 | 0.0007 |
| Hyponymy rank | 95.3 | 4 | <.001*** | 0.009 |
| Relative rank | 9.70 | 4 | .046* | 0.0009 |

* $p < .05$ ** $p < .01$ *** $p < .001$

standardized Wilcoxon statistic $W$ and $p$-values for each comparison are presented in Table 4.9.

For the absolute hypernymy rank, the expected decrease in genericity was only observed between the A1 (*Mdn* = 5.00) and the A2 (*Mdn* = 5.33) levels, but this difference was not significant, $W = 1.12$, $p = .93$. However, for all other levels, the reverse trend was observed: the degree of genericity increased as the difficulty level increased. This difference was only significant between the A2, B2, and C1 levels: novel words were considerably more generic at the B2 (*Mdn* = 5.00) and C1 (*Mdn* = 4.69) levels than at the A2 level, respectively, $W = -3.97$, $p = .040$, and $W = -3.87$, $p = .049$. In other words, the degree of concept genericity first slightly decreased from A1 to A2 but then increased again from A2 onwards. These results were contrary to the expectations that concept genericity would decrease at higher difficulty levels. For this reason, it was concluded that the standard absolute hypernymy rank could not significantly distinguish between increasing difficulty levels.

Conversely, for the hyponymy (height) rank, DSCF comparisons corroborated the expected increase in specificity, which was significant for all levels except for the B2-C1 levels. At the A1 level, the vocabulary used in the reading materials was least specific (*Mdn* = 2.00). Next, new words introduced at the A2 level became significantly more specific (*Mdn* = 1.69), $W = -10.9$, $p < .001$.

**Table 4.9**

*DSCF Tests for Multiple Comparisons of Hypernymy Ranks Across CEFR Levels*

| Rank | Level | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|---|
| | | | WORD-SENSE AVERAGING | | | |
| Hypernymy | A1 | | 1.115 | −0.202 | −1.674 | −3.235 |
| | A2 | .934 | | −2.078 | −3.972 | −3.865 |
| | B1 | 1.000 | .583 | | −1.940 | −3.026 |
| | B2 | .761 | .040* | .646 | | −2.154 |
| | C1 | .149 | .049* | .203 | .548 | |
| Hyponymy | A1 | | −10.87 | −13.81 | −17.64 | −11.09 |
| | A2 | *** | | −7.02 | −13.24 | −6.95 |
| | B1 | *** | *** | | −6.37 | −4.47 |
| | B2 | *** | *** | *** | | −2.07 |
| | C1 | *** | *** | .014* | .588 | |
| Relative | A1 | | 6.034 | 5.774 | 6.640 | 1.145 |
| | A2 | *** | | −0.100 | 2.002 | −1.839 |
| | B1 | *** | 1.000 | | 1.897 | −1.713 |
| | B2 | *** | .618 | .665 | | −2.256 |
| | C1 | .928 | .691 | .745 | .501 | |
| | | | WORD-SENSE DISAMBIGUATION | | | |
| Hyponymy | A1 | | −6.98 | −9.58 | −11.65 | −8.67 |
| | A2 | *** | | −5.01 | −8.23 | −6.00 |
| | B1 | *** | *** | | −3.61 | −4.40 |
| | B2 | *** | *** | .080 | | −3.07 |
| | C1 | *** | *** | .016* | .190 | |
| Relative | A1 | | 3.762 | 3.397 | 3.803 | 0.115 |
| | A2 | .060 | | −0.557 | 0.384 | −1.992 |
| | B1 | .115 | .995 | | 0.854 | −1.816 |
| | B2 | .056 | .999 | .975 | | −2.151 |
| | C1 | 1.000 | .622 | .701 | .549 | |

*Note.* The upper triangles contain the standardized Wilcoxon statistic $W$ and the lower triangles contain the p-values.

* $p < .05$ ** $p < .01$ *** $p < .001$

This significant increase in specificity continued from A2 to B1 (*Mdn* = 1.57), $W = -7.02$, $p < .001$, and from B1 to B2 (*Mdn* = 1.50), $W = -6.37$, $p < .001$. Between B2 and C1 (*Mdn* = 1.21), the increase in specificity was no longer significant, $W = -2.07$, $p = .59$. In sum, the hyponymy rank could account for a significant increase in semantic complexity (i.e., conceptual specificity), particularly between reading materials at basic to the high-intermediate levels.

Furthermore, a Pearson product-moment correlation showed a significant association between hyponymy rank and word concreteness, $r(9160) = -.10$, $p < .001$. The higher the concreteness rating for a word, the higher its hyponymy rank (i.e., the lower its rank number). The correlation analysis showed that more specific words tended to be more concrete, but this association was negligible. By contrast, Crossley et al. (2009, p. 322) observed a significant positive correlation between hypernymy and concreteness: the higher the word's degree of abstractness (i.e., the lower its concreteness rating), the higher its hypernymy rank (i.e., the lower its rank number), $r(97) = .62$, $p < .001$.

A similarly increasing trend in specificity was observed for the relative position in the hypernymy path. Nevertheless, DSCF comparisons showed that this difference was only significant between, on the one hand, the A1 (*Mdn* = .68) level and, on the other hand, the A2 (*Mdn* = .71), B1 (*Mdn* = .71), and B2 (*Mdn* = .72) levels, respectively, $W = 6.03$, $p < .001$; $W = 5.77$, $p < .001$; and $W = 6.64$, $p < .001$. Again, the relative hypernymy rank showed an increase in conceptual specificity from the basic to the high-intermediate levels, but, contrary to the hyponymy rank, the effect was less clear-cut.

### *Disambiguated Semantic Complexity Features*

When the word-sense disambiguated hypernymy ranks were compared across CEFR levels, only the degree of specificity (hyponymy rank) could significantly discriminate between difficulty levels, as assessed by multiple Kruskal-Wallis tests (Table 4.8) and post hoc comparisons (Table 4.9). The concepts introduced at each increasing difficulty level became more specific, which was significant from the basic towards the low-intermediate difficulty levels (A1 < A2 < B1). Nevertheless, compared to the small effect size for the indices of semantic complexity averaged over all possible senses of a word ($\epsilon^2 = .020$, see Table 4.8), the discriminative effect of disambiguated indices of semantic complexity was

negligible ($\epsilon^2$ = .009, see Table 4.8). Consequently, it can be concluded that averaged features enabled to better discriminate between different difficulty levels than indices computed for the specific sense of the word.

However, whether this implied that the computation of semantic complexity was consequently less accurate was less clear. Figure 4.21b on page 182 shows that, when the hyponymy rank was averaged over all possible senses of the word, conceptual specificity was significantly underestimated (i.e., the rank numbers were significantly higher) for all levels except the most basic A1 level. This underestimation may have lead to believe that there were significant differences in conceptual specificity at higher proficiency levels. It may just have been that the effect of specificity was a more crucial determinant of differences in readability at lower proficiency levels, but not at higher proficiency levels, where the precise meanings of the newly introduced words already displayed the highest possible degree of specificity. Word-sense averaging may therefore have lead to observe unduly distinctive and significant differences at higher proficiency levels. Indeed, for the absolute hypernymy rank, some significant differences were observed between higher proficiency levels (A2 > B1 > B2), but these effects were counterintuitive.

*Conclusion for RQ4.2*

In conclusion, the answer to the second research question was negative: word-sense disambiguation could not significantly lead to better discrimination of semantic complexity. The answer to whether the computation was consequently more accurate was not clear-cut. One could argue for both. On the one hand, the analysis could indicate whether a particular index of complexity was appropriate using a stricter, semantically disambiguated measure. Through word-sense disambiguation, the current study unearthed that, contrary to the usual hypernymy rank, the hyponymy rank may be a better indicator of semantic complexity in graded reading materials, both when one considers either all meanings of the word or only its precise meaning.

On the other hand, the accuracy of these disambiguated indices of semantic complexity depends on practical matters. The WSD system could attribute at least one sense to only three out of four (76%) lexical entries that bore a

whole meaning (i.e., adjectives, adverbs, nouns, and verbs). This low coverage means that if we were to integrate (supervised) word-sense disambiguation in our predictions of complexity and difficulty, we would not obtain a precise index of semantic complexity for a considerable proportion of content words. Performing such partial-coverage analyses of complexity would, of course, have significant repercussions on the ultimate performance that a predictive system could achieve. For this reason, it may be more advisable to resort to other, more high-coverage computational approaches that also rely on contextualized representations. We will come back to this issue in Chapter 7.

## 4.5 A CROSS-LINGUAL COMPARISON OF COGNATES

The third and final analysis examined to what extent learner characteristics were accounted for in NT2Lex. In particular, the analysis focused on the word's cognate status in the learner's L1 and compared one pair of languages, namely Dutch and French. On the one hand, the analysis examined the distribution of Dutch words in NT2Lex that were cognates for native speakers of French. On the other hand, the analysis examined the French words in FLELex that were cognates for native speakers of Dutch as a means of comparison. The analysis examined whether there was (a) a significant trend in the occurrence of cognate words across increasing difficulty levels and (b) a strong association between the levels of difficulty at which these Dutch-French cognates first occur in both resources.

### 4.5.1 *Etymological Relatedness and Translation Equivalence*

The cognate status of a word was defined on two levels. First, a broad definition of cognateness was adopted: two words were considered cognates if they shared at least one etymological root. This definition was the most general interpretation of cognateness (see Section 4.1.3) and was therefore an upper bound on the number of cognate words that could be identified. Second, the set of etymological cognates were further restricted to translation equivalents. This stricter definition was used to investigate the well-established effect that semiotically transparent cognate words will be easier to read and learn (see

Section 4.1.3). The procedure used to identify the cognate status of a word is summarized in Figure 4.21 and further detailed below.

*Etymological Cognates*

The etymological relation between two words was determined with Etymological WordNet (de Melo, 2014).[53] Etymological WordNet is a machine-readable database extracted from Wiktionary, which contains information on the etymological origins of words. For a word form in a given language (e.g., the French word *abricot*, 'apricot'), the database gives information on its ancestors and descendants and other mechanisms of word formation such as derivation.

The database was queried to generate an ancestry tree for all words in the vocabulary, which included all lemmata attested in NT2Lex and FLELex. The longest path was constructed from the word to its etymological root(s) in the ancestry tree. For example, the following path was constructed for *abricot*:

```
fra: abricot
├── cat: abrecoc
└── spa: albaricoque
        ├── ara: البرقوق
        └── ara: برقوق
```

Next, this ancestry tree was used to retrieve, for all Dutch and French words, all cognate words in the other language with the same etymological root.[54]

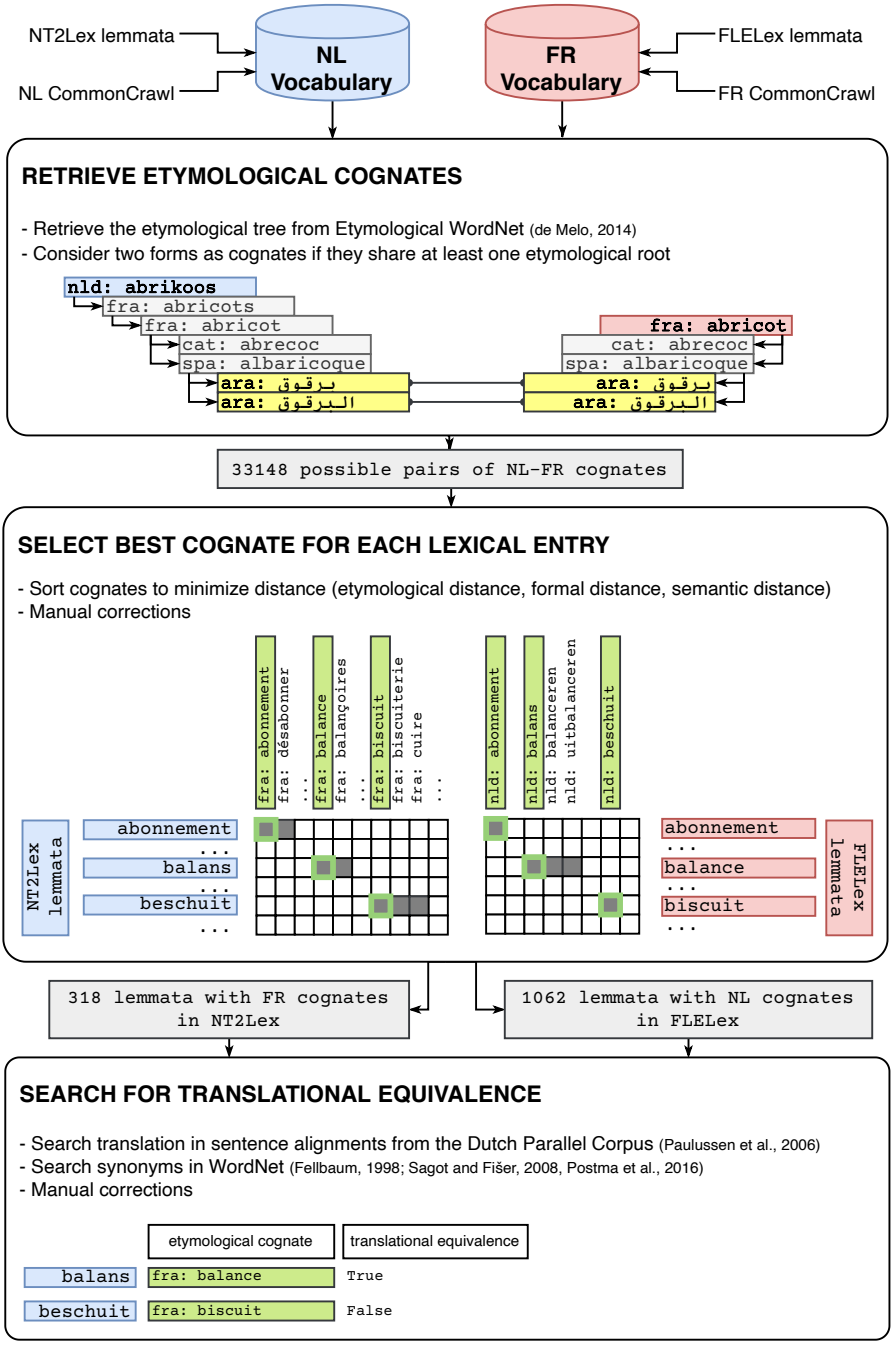
However, the Dutch cognate for *abricot* was not included in the NT2Lex resource. Restricting the search space to these lemmata would therefore have lead to various missing links between both languages. The lemmata were supplemented with the FastText vocabularies extracted from the Dutch and French Common Crawl corpora (Grave et al., 2018). For example, the following ancestry tree was constructed for the word *abrikoos*,

53 In this way, the study adopted a methodology similar to Rabinovich et al. (2018).

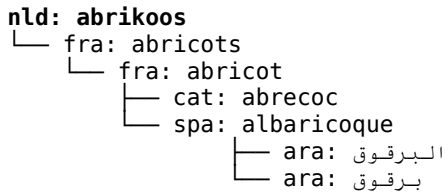
54 Because the etymological network was a directed graph, the network needed to be recursed in two directions in order to retrieve all ancestors. For a given node, all related words were identified either (a) via a link from the ancestor to the node (`etymological_origin_of` and `has_derived_form`) or (b) via a link from the node to the ancestor (`etymology`, `etymologically`, `is_derived_from`, and `derived`).

**Figure 4.21**

*Identification of Cognates in NT2Lex and FLELex*



which was missing from NT2Lex, but attested in the Dutch Common Crawl:




As a result, for each lemma in NT2Lex and FLELex, a list of etymologically related words was retrieved, empty if no shared etymological origin could be found. Now, in cases where a lemma had more than one related word, the closest cognate needed to be found.

For each lemma, the list of etymologically related words were sorted based on three types of distances: the (a) etymological, (b) formal, and (c) semantic distance between the related word form and the lemma. Firstly, the etymological distance between two words was computed with tree-based distance measures, including the (normalized) tree edit distance (TED) (Pawlik & Augsten, 2015, 2016).[55] Secondly, the formal distance between two words was computed with standard measures (Mitkov et al., 2007). The basic assumption was that, because two cognate words were derived from the same origin, they would also display a certain degree of formal similarity. The following distance measures were used: the truncation method[56] (Simard et al., 1993), the Dice coefficient of character n-grams (Brew et al., 1996; Church, 1993), the longest common subsequence (LCS) ratio (Melamed, 1999), and the normalized edit distance (NED) (Inkpen & Frunza, 2005). Thirdly, the semantic distance between two related words was computed with the cosine of the two FastText vectors. The basic assumption was that, because two cognate words were derived from the same origin, they would also dis play a certain degree of semantic similarity. With all these distance measures combined, the least distant cognate word was retrieved for each lemma in NT2Lex and FLELex. This retrieval was manually verified and corrected in terms of number agreement (singular vs. plural) and part of speech (nominalizations).

55 For instance, the TED between *abrikoos* and *abricot* is two: two edit operations (two deletions or two insertions) have to be performed to change the one into the other. In this case, the distance is fairly low, indicating that both words are etymologically very similar.

56 This measure considers two words as (dis)similar if the four leading characters are (dis)similar.

At this stage, it should be reiterated that this was an upper bound on the degree of cognateness. Indeed, among the set of Dutch-French cognates identified, there were pairs of strongly cognate words, such as the exact formal and semantic equivalents *abonnement* 'subscription'. However, there were also pairs of related words which were only partially equivalent. Some word pairs were related because they shared a cognate morpheme, such as the Dutch *badkamer* 'bathroom' and the French *chambre* 'room'. Furthermore, other pairs were etymologically and formally related, but did not convey the same meaning. For instance, although the Dutch *beschuit* was both etymologically and formally cognate with the French *biscuit* 'biscuit', it was the semantic equivalent of the French *biscotte* 'rusk'. An even better pair of more semantically opaque etymological cognates were the French *août* 'August' and the Dutch *oogst* 'harvest' (vs. the Dutch *augustus* 'August'), originating from the Latin *augustus*. In short, there was much variation in the degree of cognateness identified so far, ranging from strictly similar cognates to derivationally related cognates and partially equivalent forms and meanings. As a result, the analysis focused on translation equivalents as a lower bound on cognateness.[57]

*Translation Equivalents*

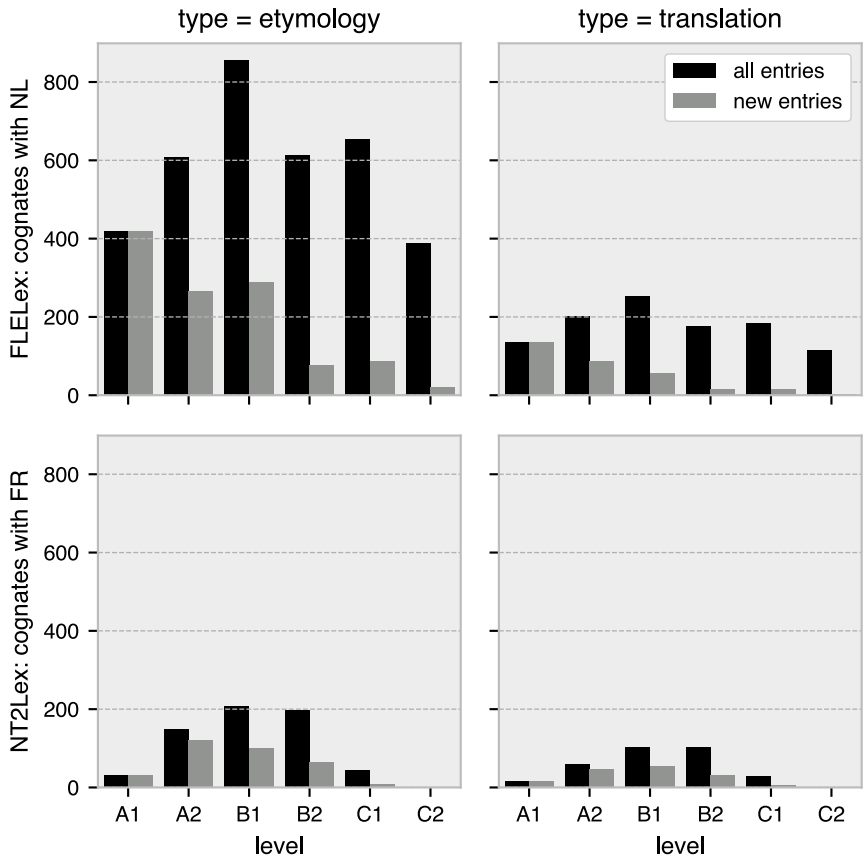
Previously, semantic similarity was assessed by means of the cosine distance between two distributional word vectors, as used by Mitkov et al. (2007) to distinguish between true cognates and false friends. However, this method did not enable to rule out word pairs that were not translation equivalents. Although the cosine distance could detect varying degrees of semantic similarity, it could not explicitly say whether two words were exact translations.

Translation equivalence was therefore determined in multiple ways. First, it was assessed whether a pair of cognate words were synonymous according to the semantic relations attested in the OMW. Next, this resource-based assessment was supplemented with a corpus-based evaluation of word translations. The Dutch-French sentence alignments of the Dutch Parallel Corpus (DPC)

57 It should be noted that translation equivalence was not the lowest possible bound on cognateness. The lowest possible bound would be to only consider exact formal and semantic equivalents (e.g., *abonnement*) (cf. Duyck et al., 2007). However, this lowest bound would probably have been too stringent. It would have ruled out semiotically transparent cognates such as *balans* and *balance*.

**Figure 4.22**

*Number of Dutch-French Cognates per Level in FLELex and NT2Lex*




(Paulussen et al., 2006) were consulted to determine whether a pair of words were translationally equivalent,. A pair of words were considered exact translations if they occurred in at least one pair of translated sentences. Finally, the list of cognate words was manually revised to correct or add translations.

### 4.5.2 *Cognates in NT2Lex and FLELex*

Figure 4.22 shows the distribution of Dutch-French cognates identified in NT2Lex and FLELex. The figure shows a decreasing trend in the number of

**Table 4.10**

*Chi-Squared Tests of Trends in Proportions of Dutch-French Cognates Among New Entries per Level in NT2Lex and FLELex*

| | A1 | A2 | B1 | B2 | C1 | C2 | $\chi^2$ | *p* |
|---|---|---|---|---|---|---|---|---|
| | | | | ETYMOLOGICALLY RELATED | | | | |
| FLELex | 10.2 % | 9.86% | 7.26% | 5.93% | 5.23% | 4.23% | 72.0 | <.001 |
| NT2Lex | 3.25% | 2.25% | 2.07% | 1.76% | 2.43% | — | 5.22 | .022 |
| | | | | TRANSLATION EQUIVALENT | | | | |
| FLELex | 3.32% | 3.19% | 1.38% | 1.15% | 0.84% | 0.60% | 64.5 | <.001 |
| NT2Lex | 1.68% | 0.87% | 1.11% | 0.91% | 1.89% | — | 0.030 | .86 |

newly introduced cognate words at each increasing difficulty level. Reading materials at the basic levels contained a higher number of new word forms with a transparent meaning based on the learners' L1. On the other hand, these initial results also showed no strict one-to-one, cross-lingual correspondence in the distribution of cognates across difficulty levels. The decreasing trend was evident in the case of FLELex, but not in the case of NT2Lex. Words introduced in French L2 reading materials at the A1 level contained many cognates for Dutch-speaking learners, whereas cognates for French-speaking learners were much less prevalent in Dutch L2 reading materials at the A1 level. To further substantiate these initial observations, two statistical analyses were run: (a) chi-squared tests of trends in proportion of cognates across increasing difficulty levels and (b) a correlational analysis of the correspondence between the levels at which Dutch-French cognates occurred in NT2Lex and FLELex.

The results of the first analysis are given in Table 4.10. In FLELex, there was a significant decreasing trend in proportions of Dutch cognates among new words occurring at increasing difficulty levels, both in overall etymological relatedness (upper bound) and strict translation equivalence (lower bound). In the basic-level reading materials, 10% of new words were etymologically related to Dutch, and 3% were semantically transparent cognates. In the advanced-level reading materials, 4% of new words were etymologically related to Dutch, and 0.6% were semantically transparent cognates. This result indicated that native speakers of Dutch starting to learn French would

**Table 4.11**

*Spearman and Kendall Correlations of the Difficulty Levels at which Dutch-French Cognates are First Introduced in FLELex and NT2Lex*

| Type | $N$ | $\rho$ | $p$ | $\tau_b$ | $p$ |
|---|---|---|---|---|---|
| Etymologically related | 600 | .075 | .066 | .063 | .066 |
| Translational equivalent | 162 | .21 | .006 | .18 | .007 |

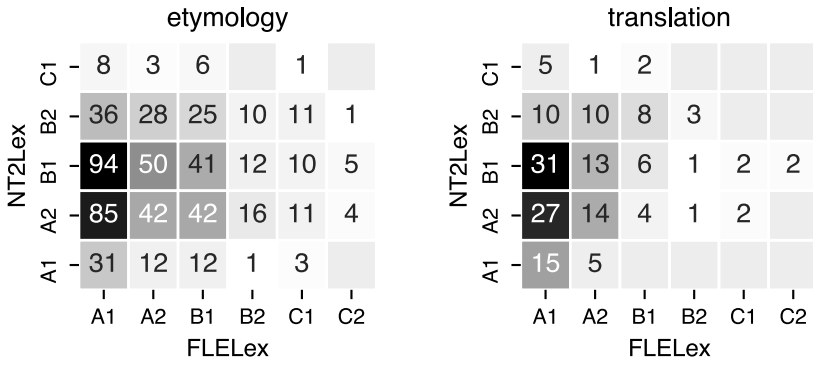
probably have less difficulty because they could capitalize more on prior L1 lexical knowledge.

Conversely, the prevalence of cognates with French remained generally low ($\approx$ 2%) in NT2Lex. This result indicated that native speakers of French starting to learn Dutch would not profit as much from their prior L1 knowledge while reading. The answer to the first question of this section was mixed: there was a significant decreasing trend in the prevalence of cognates across increasing difficulty levels, but this trend was not retained cross-lingually.

The second analysis focused only on the pairs of Dutch-French cognates attested in both FLELex and NT2Lex. Multiple correlation analyses were run to examine whether there was a significant association between the difficulty levels at which these cognates were first introduced. The results of these analyses are listed in Table 4.11. This table shows that the only significant association was found for semantically transparent cognates. This result was expected and underscored the importance and relevance of the criterion of translation equivalence for assessing L2 lexical difficulty. Nevertheless, the association between difficulty levels remained poor (Spearman's $\rho < .4$; Kendall's $\tau_b < .2$). This poor correlation can be seen more clearly in Figure 4.23. Some cognates words that first occurred in FLELex at A1 were only introduced in NT2Lex at B2 (e.g., *citer*/*citeren*, 'to cite') or C1 (e.g., *activer*/*activeren*, 'to activate'). Conversely, some cognate words that first occurred in NT2Lex at A2/B1 were only introduced in FLELex at B2 (e.g., *menu*/*menu*, 'menu'; *burgemeester*/*bourgmestre*, 'mayor'), C1 (e.g., *flexibel*/*flexible*, 'flexible') or C2 (e.g., *flexibiliteit*/*flexibilité*, 'flexibility'). In sum, the results were negative: there was no strong correspondence between the levels of difficulty at which Dutch-French cognate words were first introduced in FLELex and NT2Lex.

**Figure 4.23**

*Confusion Matrix of the Difficulty Levels at which Dutch-French Cognates are First Introduced in FLELex and NT2Lex*




*Conclusion for RQ4.3*

In sum, the results could not prove that learner characteristics such as cognateness were consistently accounted for in a theoretical measure of lexical difficulty such as word occurrence in CEFR-graded reading materials. Even more so, it appeared that prior L1 lexical knowledge and cognateness would have an impact on how effective these theoretical measurements are. Using the same CEFR-graded lexicon threshold as a measure of difficulty would very likely miscalculate the fact that a more significant proportion of vocabulary at the basic-level reading materials will be semantically transparent for certain learner profiles (e.g., Dutch-speaking learners of French) but not others (e.g., French-speaking learners of Dutch). What is more, the fact that there was no solid one-to-one correspondence between the levels at which translationally equivalent cognates (e.g., *flexibel* [A2] / *flexible* [C1]) first occurred showed that there was much non-systematicitiy in this theoretical measure of difficulty. Consequently, it may be more advisable to shift the focus away from a theoretical measure of general lexical difficulty towards an empirical measure of personal lexical difficulty.

## 4.6 DISCUSSION

The three studies presented in Sections 4.3 to 4.5 examined the use of a primary theoretical measure of lexical difficulty for L2 readers. This measure was derived from word occurrence in a corpus of reading materials intended for specific proficiency levels. The central premise was that words newly introduced at a specific proficiency level would be unfamiliar and difficult. In contrast, words occurring at the same level but introduced at lower levels would be more familiar and less difficult.

Now, there is, of course, a need to critically reflect on this basic premise and ask the question of what knowledge on difficulty we can and cannot gather from this measure. Therefore, the study provides answers to three analytic questions.

The first question is concerned with whether a measure of lexical complexity can be extracted from the resource. The results show that the graded distributions are associated with various standard lexical features and norms. On the one hand, the A1 to C1 levels display a decreasing frequency, dispersion, polysemy, and concreteness. Words that occur in basic-level textbooks are more frequent, more dispersed, and more concrete than words in advanced-level textbooks. On the other hand, the A1 to C1 levels display an increasing trend in sophistication, hyponymy, and age of acquisition. Words that occur in advanced-level books are more sophisticated, represent more specific concepts, and are learned much later by native speakers than words occurring in basic-level materials. Importantly, this increasing trend in complexity is only clearly visible when considering only those words newly introduced at each level. Because words that are first introduced at the basic levels (such as auxiliary verbs) also frequently reoccur in advanced-level textbooks, the inclusion of these "familiar" words would underestimate the lexical complexity at the advanced levels. Moreover, the results also show a more rapid, almost exponential increase in complexity from the A1 to the A2 and B1 levels. As a result, it can be inferred that the resource provides a reasonable measure of lexical complexity, which increases more rapidly in the early acquisition stages and which is most prevalent for new words at each increasing difficulty level.

The second question is whether a more accurate measure of difficulty can be obtained by resorting to word-sense disambiguation. There is a significant difference in semantic complexity per difficulty level when computed for the particular word sense or averaged over all possible word senses. There is a significant increase in semantic specificity (i.e., hyponymy rank) from A1 to A2 and from A2 to B1. This difference reflects the findings of Crossley and Salsbury (2010) that words (verbs) produced at the earliest acquisition stages are significantly less specific. This increase remains significant from the B1 level onwards for word-sense averaging.

However, this increase is no longer significant when the particular sense of the word is considered. This result may indicate that averaging the hyponymy rank for all senses underestimates the degree of specificity at higher difficulty levels. Words at these levels may be so specific already that concept specificity no longer becomes a distinguishing or criterial factor. The underestimation of specificity from word-sense averaging would then lead to observing undue significant differences between more advanced levels.

The idea that there may not be as much of a difference in terms of vocabulary at higher CEFR levels makes sense if we reconsider some insights from the reference level descriptors (RLDs). Not all currently available RLDs go as far as the C1/C2 levels.[58]. Moreover, there is some debate as to whether it is possible to determine the vocabulary for the C1/C2 levels (Marello, 2012). Therefore, an avenue for research is to contrastively analyze the level at which words are introduced in CEFR-graded materials with the CEFR level at which a word should be introduced according to the RLD (see Pintard & François, 2020).

A further incidental finding of the word-sense disambiguation analysis is that hyponymy rank is perhaps a better indicator of semantic complexity than hypernymy rank. Contrary to the clear increasing trend in specificity (i.e., hyponymy rank), there is no decreasing trend in genericity (i.e., hypernymy rank) in reading materials at increasing difficulty levels. This finding is contrary to the significant trend observed for productive vocabulary use Crossley et al. (2009). Therefore, it may be that hypernymy rank is a better predictor of productive vocabulary than reading vocabulary. Crossley et al.

58 The English Vocabulary Profile was completed with material for the C1/C2 levels (Capel, 2012). The *Référentiels* for French are available for levels A1 to B2. Riba (2010)'s doctoral dissertation aimed to supplement the *Référentiels* with new descriptors for the C1/C2 levels.

(2007) did not find a significant difference in hypernymy between two levels of textual difficulty, and Crossley (2013) also noted an absence of bidirectionality between hyponymy and hypernymy in a semantic priming task. The genericity of a concept may be less of a defining factor than the degree of specificity. It may be more advisable to look at semantic complexity from the point of view of concept *specificity* (or hyponymy) rather than from the point of view of concept genericity (or hypernymy).

It should be noted that conceptual genericity and specificity are not the same as word abstractness and concreteness. It may seem contradictory that the A1 to C1 levels are characterized by a decrease in concreteness (see Section 4.3) and an increase in specificity (see Section 4.4). However, the distinction between concept genericity and specificity is not equal to the distinction between word abstractness and concreteness. For instance, the average concreteness for the more generic concept *hond* 'dog' is 5 (*concrete*), whereas the average concreteness for the more generic concept *doctrine* 'doctrine' is close to 1 (*abstract*). Conversely, the average concreteness for the more specific concept *keeshond* 'Keeshond' is approximately 4 (*more concrete than abstract*), whereas the average concreteness for the more specific concept *calvinisme* 'Calvinism' is close to 1 (*abstract*). Consequently, it is not problematic to conclude that words introduced at the A1 levels are both concrete and generic and that words introduced at the C1 levels are both abstract and specific.

The third question is concerned with whether this theoretical measure accounts for learner characteristics such as cognate status. The results show a cognate effect in some specific cases: a considerable proportion of basic-level entries in FLELex are cognates with Dutch. However, the results also show no one-to-one correspondence between the CEFR level at which these cognates first occur in reading materials for various target languages. The underlying reasons for this apparent non-systematicity may be similar to what was previously observed for vocabulary size and depth (Milton, 2010; Milton & Alexiou, 2009).

> The evidence we have from vocabulary size tests (...) suggests the actual volumes of vocabulary that are associated with each CEFR level. This information should be very useful to learners, teachers and other users of the CEFR is helping to link language perfor-

> mance to the CEFR levels. The evidence also appears to suggest that vocabulary breadth may vary from one language to another but it is not yet clear whether this reflects differences between the languages themselves, or differences in the construction of the corpora from which vocabulary size tests are derived. (Milton, 2010, p. 211)

Similarly, it is not clear whether the lack of correspondence between NT2Lex and FLELex is due to differences between French and Dutch or whether it simply reflects a non-systematicity in how the reading materials were included in the graded textbooks. On the one hand, it may be that there are more cognates in French for Dutch speakers because of the historical influence of the French language (e.g., see Deneckere, 1954). On the other hand, it may also very well be that the reading texts were selected without rigorous criteria that specifically targeted the lexical level and which were coherent with the CEFR descriptors. Indeed, other studies have also underscored this issue. Pintard and François (2020) highlighted merely 25% correspondence between the level (A1-B2) introducing a new word in FLELex and when this word should have been introduced according to the French RLD. Similarly, Graën et al. (2020) observed the need to correct the distributions for making cross-lingual comparisons possible. In other words, the lack of systematicity may be indicative of some incongruities in the way reading materials were graded with CEFR levels, which may call for a more critical reflection.

Lastly, it is currently unsure whether there is a better way to measure lexical complexity than by looking at the first occurrence level. Previously, Gala et al. (2013) examined whether the application of "a more complex function to transform the frequency distributions might produce a better classification [than the level of first occurrence]" (p. 137). However, because this more complicated transformation rule did not improve the complexity measure, the authors stuck to the first level of occurrence. Similarly, Alfter and Volodina (2018) examined relabeling the first-occurrence level with the CEFR level located at a specific occurrence percentile. However, this method would only work for 40% of entries in NT2Lex since the other 60% only occur in one level (cf. Section 4.3) and cannot be relabeled. Recently, Pintard and François (2020) attempted to establish a transformation rule with a FLELex boosting classifier

optimized to predict RLD levels. Although the classifier outperformed the first-level method for the A1-B2 levels, achieving 54% accuracy on predicting *Beacco* CEFR levels, this transformation rule could also potentially relabel a word with a predicted RLD level at which the word never occurred in the resource. Consequently, such a relabeling would be misleading if the objective is to establish a difficulty measure from *actual* word occurrence in graded reading materials, as is the case in this study. Therefore, further research is needed to determine whether one could extract a more accurate measure of lexical difficulty from graded frequency distributions.

## 4.7 CONCLUSION

Based on a corpus of graded reading materials, it is possible to derive a theoretical measure of lexical difficulty for foreign language readers. All lexical entries introduced in materials at a given level are assumed to be unfamiliar and labeled as complex target vocabulary. Although this method is an easy way to estimate lexical difficulty, it nevertheless corresponds to what one would expect. Novel lexical entries at the beginner levels are more frequent, more dispersed, more concrete, less specific, and acquired earlier by native speakers. Novel lexical entries at the more advanced levels are less frequent, less concrete, more specific, and acquired later by native speakers.

At the same time, this theoretical measure is not without concerns. First of all, because half of the lexical entries occur at one difficulty level only, the measure is characterized by a substantial degree of sparsity in the distribution of lexical difficulty across levels. Moreover, there is a lack of one-to-one correspondence in the distribution of translation equivalents across languages. Several Dutch-French semantically transparent cognate words such as *flexibel* and *flexible* occur at disparate difficulty levels in the NT2Lex and FLELex resources.

In sum, there is still a non-negligible lack of systematicity in this theoretical measure of difficulty, despite a clear increasing trend observed for the complexity of novel entries. Consequently, it is uncertain whether we should consider this measure as ground truth for the automated prediction of lexical difficulty.

## 4.A APPENDIX

### 4.A.1 *RESTful API*

*Base URL for the API*

```
$ API="https://cental.uclouvain.be/cefrlex/api/v1"
```

*Retrieve a Resource*

- List all resources:

```
$ curl "${API}/resources/"
```

- Retrieve a list of resources, filtered by the query parameters:

```
$ curl "${API}/resources/?language=fr"
$ curl "${API}/resources/?type=productive"
```

- Retrieve a resource based on a unique resource identifier:

```
# either the name is sufficient
$ curl "${API}/resources/SVALex/"
# or if multiple versions exist, add the version name
$ curl "${API}/resources/NT2Lex/"
{"detail":"Multiple versions of the resource were found. Please use a
↪  unique resource identifier instead."}
$ curl "${API}/resources/NT2Lex-CGN+ODWN-01"
# or if multiple versionings exist, add the version number
$ curl "${API}/resources/NT2Lex-CGN+ODWN-01"
```

*Search a Resource*

- Retrieve a single entry based on a lemma, pos, and/or sense

```
$ curl "${API}/resources/NT2Lex-CGN/search/?lemma=lezen"
$ curl "${API}/resources/NT2Lex-CGN/search/?lemma=lezen&pos=WW()"
$ curl
↪  "${API}/resources/NT2Lex-CGN+ODWN/search/?lemma=lezen&sense=lezen-v-1"
```

- Retrieve a list of entries, filtered by query parameters[59]:

59 See Django's syntax for field lookups: https://docs.djangoproject.com/en/3.0/ref/models/querysets/#field-lookups.

```
$ curl "${API}/resources/NT2Lex-CGN/search/?lemma__value__startswith=lees"
```

*Analyze a Text with a CEFR-based Lexical Complexity Threshold*

```
$ curl -H "Content-Type: application/json" -X POST -d '{"level":"A1","text":"Dit
↪  is een testanalyse voor de lexicale moeilijkheid op een A1 niveau."}'
↪  "${API}/resources/NT2Lex-CGN/analyze/"
$ curl -H "Content-Type: application/json" -X POST -d '{"level":"A2","text":"Dit
↪  is een testanalyse voor de lexicale moeilijkheid op een A2 niveau."}'
↪  "${API}/resources/NT2Lex-CGN+ODWN/analyze/"
```

### 4.A.2 *Hypernymy Ranks*

Median Hypernymy Ranks for Novel Entries at Each Level in NT2Lex

| rank | levels | absolute | relative | height |
|---|---|---|---|---|
| AVG | A1 | 5.000000 | 0.677083 | 2.000000 |
| | A2 | 5.333333 | 0.712500 | 1.687500 |
| | B1 | 5.200000 | 0.709821 | 1.571429 |
| | B2 | 5.000000 | 0.722222 | 1.500000 |
| | C1 | 4.690476 | 0.669823 | 1.208333 |
| WSD | A1 | 4.000000 | 0.555556 | 2.000000 |
| | A2 | 4.000000 | 0.600000 | 1.000000 |
| | B1 | 4.000000 | 0.500000 | 1.000000 |
| | B2 | 4.000000 | 0.500000 | 1.000000 |
| | C1 | 4.000000 | 0.500000 | 1.000000 |

Mean Hypernymy Ranks for Novel Entries at Each Level in NT2Lex

| rank | levels | absolute | relative | height |
|---|---|---|---|---|
| AVG | A1 | 5.071804 | 0.664576 | 2.217935 |
| | A2 | 5.224993 | 0.698167 | 1.962091 |
| | B1 | 5.161824 | 0.700400 | 1.914020 |
| | B2 | 5.041988 | 0.707945 | 1.789462 |
| | C1 | 4.698157 | 0.681935 | 1.781925 |
| WSD | A1 | 5.206349 | 0.636691 | 2.275613 |
| | A2 | 5.318370 | 0.666652 | 1.969383 |
| | B1 | 5.233725 | 0.669058 | 1.889489 |
| | B2 | 5.163043 | 0.673548 | 1.748518 |
| | C1 | 4.750000 | 0.650383 | 1.644231 |

CHAPTER 5

# *A Posteriori Knowledge Of Difficulty* PERCEIVED LEXICAL DIFFICULTY IN A NATURAL READING TASK

**Abstract** This chapter presents an empirical study on lexical difficulty as subjectively assessed by French L2 learners when reading. The chapter describes two trials during which 56 learners highlight words they perceive as unknown in a natural reading task. The entire data set comprises a total of 262,054 observations of perceived lexical difficulty. The data set presents two advantages. Firstly, the data accounts for individual differences between learners. Unlike previous data sets where different participants read different contexts, every learner in the same group reads the same texts. Secondly, the data is contextualized and provides evidence for the impact of high and low contextual constraints, which are contexts where a masked word can or cannot be predicted by a state-of-the-art BERT language model.

Because of unresolved issues in establishing a theoretical measure of lexical difficulty (Chapter 4), the doctoral study's focus shifted towards establishing lexical difficulty based on empirical learner data. As noted in Chapter 3, this empirical measure can be of two types. On the one hand, we may resort to indirect measures of how unknown words are processed while reading, as extracted from eye movements or brain potentials. On the other hand, we may use direct measures of lexical difficulty established from how non-native readers themselves judged (or perceived) the difficulty of words in context.

Two essential aspects characterized the empirical studies reviewed in Chapter 3. The first aspect was the considerable individual differences observed when non-natives processed words while reading. However, this between-

learner variance was only observed in studies using indirect difficulty measures. Conversely, previous machine learning studies, which used direct difficulty measures as training data, did not account for individual differences; instead, personal difficulty judgments were aggregated into a general tendency of difficulty. The second aspect was concerned with contexts that posed high or low semantic constraints. A context had high semantic constraints if an unknown (masked) word could be easily predicted. On the other hand, a context had low semantic constraints if an unknown (masked) word could not be predicted. Again, the effect of high and low semantically constrained contexts was only examined on indirect difficulty measures.

This chapter presents the collection of a new direct measure of lexical difficulty for French L2 readers. The study reported in this chapter adopts the same methodology as previous machine learning studies. At the same time, the study improves the previous methodology in two ways. The first improvement is accounting for individual differences in a direct difficulty measure, specifically in how learners perceived lexical difficulty while reading. The second improvement is investigating the effect of high and low semantic constraints in this direct difficulty measure. The following two questions guided the study:

RQ5.1 To what extent do learners (with the same proficiency level and L1) vary in how they perceive lexical difficulty when reading the same text? Would it be reliable to aggregate this data into a single ground truth of difficulty?

RQ5.2 Is there an effect of high and low semantically constrained contexts on how learners perceive lexical difficulty?

A standard method of determining high and low contextual constraints is the cloze procedure. In a cloze procedure, a sample of subjects is given a text (or a context) in which one or more words are left out. The cloze procedure computes the degree of word predictability by tallying the number of subjects inferring the word(s) correctly, divided by the total number of subjects. Taylor (1953) first proposed this method as a measure of text readability. Since then, cloze probability has often been used in psycholinguistic studies to define high and low contextual constraints (e.g., see Parviz et al., 2011; Smith & Levy, 2011).

Another way to compute cloze probabilities is to use a computational language model. The idea of using language models for predicting readability and difficulty has been a long-standing research topic (e.g., Collins-Thompson & Callan, 2005; Schwarm & Ostendorf, 2005) and continues to date (e.g., Benzahra & Yvon, 2019). Moreover, research on the development of computational language models has also been expanding in recent years, introducing new state-of-the-art methods with deep contextualized word representations (Peters et al., 2018). One model that has recently gained popularity is the BERT transformer (Devlin et al., 2019), trained on a large body of English texts to predict – among other things – a masked word from the surrounding sentence context. Similar pre-trained models have been made available for other languages, including French (CamemBERT, Martin et al., 2019; FlauBERT, Le et al., 2019).[60]

The chapter is structured as follows. Section 5.1 describes the collection of two data sets measuring how French L2 learners individually perceive lexical difficulty in a natural reading task. Section 5.2 presents two analyses. The first analysis (Section 5.2.1) examines the extent of learner variance in the collected data. The second analysis (Section 5.2.2) looks into the effect of high and low semantic constraints computed with state-of-the-art BERT language models for French. Section 5.3 concludes the chapter.

## 5.1 LEARNER DATA COLLECTION

The study adopted a methodology similar to previous shared tasks (Paetzold & Specia, 2016a; Yimam et al., 2018), albeit with some crucial differences:

- Instead of distributing different texts between participants, all learners in the same (proficiency) group read the same texts.
- The data set included unique and anonymized learner identifiers.
- Instead of sentence or paragraph contexts, the study focused on a more natural reading task with texts presented in their entirety.

60 Pre-trained language models are available from the *transformers* library (Wolf et al., 2020), available at https://huggingface.co/transformers/.

Two data collection trials were conducted. During the first trial (see Section 5.1.1), participants read more extensively and consequently judged a more extensive number of words. Because the reading materials contained several graded readers for the A1–B1 levels, the trial focused on participants having an A2 or B1 proficiency level. However, a practical concern was the excessive time required to conduct this reading task and, consequently, the limited availability of participants. For this reason, the second trial (see Section 5.1.1) focused on a larger sample of learners participating in a shorter reading task (1-hour sessions). The second trial targeted four different proficiency levels: the A2, B1, B2, and C1 levels. For each of these proficiency levels, a sample of five short texts (100 – 500 words) was selected from a corpus of publicly available texts.

### 5.1.1 *Trial 1*

#### *Materials*

The reading materials came from a corpus of simplified French texts (Brouwers et al., 2014), which included either tales simplified for native speakers (A1–B1) or encyclopedic articles simplified for children (Vikidia). A set of 51 texts were selected from this corpus, based on a procedure described in Tack et al. (2016a). In short, the selection procedure aimed to minimize the lexical overlap between texts and to maximize the overall lexical richness. The surface-level characteristics of these materials are described in Table 5.1. On average, the texts counted 413 tokens, with 188 distinct types and 151 different lemmas per text. The average readability was *standard* ($M_{RE}$ = 65) based on Flesch's (1951) Reading Ease score (see the Appendix, Section 5.A.1), with a standard deviation from *fairly difficult* (RE = 52.6) to *fairly easy* (RE = 77.4). The reading materials were suitable for general audiences and secondary school students (12–18 years). The surface-level characteristics and readability scores for each text are given in Table 5.14 in the Appendix (Section 5.A.2).

TARGET VOCABULARY Contrary to the standard methodology in studies on L2 vocabulary (cf. Chapter 2), no selection procedure was used to define a set of target words. Instead, all individual words attested in the reading

**Table 5.1**

*Reading Materials in Trial 1*

| | *M* | *SD* | min | *Q*1 | *Mdn* | *Q*3 | max |
|---|---|---|---|---|---|---|---|
| Sentences | 30.7 | 50.2 | 4 | 10 | 16 | 21 | 275 |
| Tokens | 413 | 716 | 101 | 140 | 221 | 267 | 4,694 |
| Types | 188 | 223 | 58 | 89 | 130 | 161 | 1,515 |
| Lemmas | 151 | 162 | 52 | 75.5 | 107 | 135 | 1,115 |
| RE | 65.0 | 12.4 | 34.9 | 55.8 | 66.0 | 72.9 | 91.5 |
| RE KM | 81.6 | 11.1 | 55.8 | 73.5 | 82.7 | 88.8 | 105 |

RE = Reading Ease (de Landsheere, 1963; Flesch, 1951)
KM = Kandel and Moles (1958)

materials were also targeted in the study. The list of individual words was identified with a standard tokenization procedure: all alphanumeric word units identified with TreeTagger (Schmid, 1994) were considered relevant; all non-alphanumeric tokens (e.g., digits, symbols, and punctuation marks) were considered irrelevant. In total, the materials included 21,048 relevant tokens.

This target vocabulary was compared with the vocabulary attested in a reference corpus of L2 reading materials. The FLELex resource (Francois et al., 2014) was used to indicate word frequency in reading materials for French learners. Table 5.2 shows the CEFR proficiency level at which the words were introduced in reading materials. In total, 65% of the relevant target items appeared in L2 reading materials (all levels taken together) and occurred once every $10^3$ words ($Mdn_{SFI}$ = 69.9). This coverage was relatively low, but it should also be noted that the list of target items comprised some (3.77%) proper nouns not included in FLELex. Furthermore, the majority of words were introduced in materials intended for the A1 to B1 levels. This corresponded to the A2–B1 levels of the learners participating in this trial. For the A2 participants, 59% of the target words were assumed to be familiar because they previously appeared in the A1 level. For the B1 participants, 63.2% of the target words were assumed to be familiar because they previously appeared in the A1 and A2 levels.

**Table 5.2**

*FLELex Vocabulary in Trial 1*

| | | | *SFI* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Level | *N* | % | *M* | *SD* | min | *Q*1 | *Mdn* | *Q*3 | max |
| A1 | 12,423 | 59.0 | 69.7 | 13.9 | 32.3 | 59.8 | 72.2 | 83.0 | 86.3 |
| A2 | 683 | 3.24 | 44.7 | 7.45 | 31.1 | 37.8 | 45.9 | 50.9 | 60.2 |
| B1 | 440 | 2.09 | 43.1 | 6.74 | 30.1 | 36.9 | 44.7 | 48.9 | 56.8 |
| B2 | 89 | 0.42 | 42.3 | 3.74 | 35.3 | 39.6 | 41.0 | 43.6 | 53.9 |
| C1 | 82 | 0.39 | 48.5 | 3.10 | 45.4 | 46.2 | 47.8 | 50.8 | 57.2 |
| C2 | 27 | 0.12 | 52.6 | 2.42 | 46.0 | 51.1 | 52.9 | 52.9 | 57.5 |
| Total | 13,744 | 65.3 | 68.4 | 14.3 | 19.5 | 58.8 | 69.9 | 82.1 | 87.4 |

A cloze procedure was used to determine the degree of contextual constraint for each word in the reading materials. Each word was masked and predicted from its preceding or surrounding context with a state-of-the-art contextualized language model for French (viz., CamemBERT; Martin et al., 2019). The cloze procedure was repeated for each word embedded in its sentence context. Only one word was masked at a time. Two implementations of the language model were used: a "masked" language model, which predicted the masked word from its surrounding (left and right) context; and a "causal" language model, which predicted the masked word from its preceding (left) context only. If the language model predictions were correct, the context was considered to be highly constrained. If the language model predictions were incorrect, the context was considered to be weakly constrained. Various thresholds $k$ were explored to determine whether a word could be correctly predicted from its context. These thresholds ranged from considering only the most probable prediction ($k = 1$) to allowing for the 20 most probable predictions ($k = 20$).

The outcome of this cloze procedure is given in Table 5.3. The table shows that only a tiny percentage (7%) of words could be immediately predicted from their antecedent contexts, that is, with the causal language model. This causal language model simulated incremental word processing: it quantified how likely it was to read the next word in a single forward pass. Consequently, it was hypothesized that this set of highly predictable words would be processed

**Table 5.3**

*Cloze Procedure for Identifying High and Low Constraint Contexts in Trial 1*

| | | $k$ = 1 | | $k$ = 5 | | $k$ = 10 | | $k$ = 20 | |
|---|---|---|---|---|---|---|---|---|---|
| Predictions | | *N* | % | *N* | % | *N* | % | *N* | % |
| Masked | 1 | 11,158 | 53 | 14,543 | 69 | 15,310 | 73 | 15,905 | 76 |
| | 0 | 9,890 | 47 | 6,505 | 31 | 5,738 | 27 | 5,143 | 24 |
| Causal | 1 | 1,481 | 7 | 5,282 | 25 | 7,423 | 35 | 9,400 | 45 |
| | 0 | 19,567 | 93 | 15,766 | 75 | 13,625 | 65 | 11,648 | 55 |

*Note.* The cloze predictions were obtained from a state-of-the-art pre-trained language model for French, namely the CamemBERT model (Martin et al., 2019). Two implementations of CamemBERT were used: the masked language model (which predicts the masked word from its surrounding [left and right] context) and the causal language model (which predicts the masked word from its preceding [left] context only).
1 = correct prediction (high-constraint context)
0 = incorrect prediction (low-constraint context)
$k$ = number of most probable predictions

with much greater ease and would not, therefore, be perceived as difficult. When considering both the left and right context windows, slightly more than half (53%) of the words could be immediately predicted. By contrast, one out of four (24%) words were challenging to predict from the surrounding context, even when the 20 most probable predictions were considered. Therefore, it was assumed that, for half of the words in the reading materials, the meaning would be congruently construed from the surrounding context, whereas the meaning would *not* be congruently construed from the surrounding context for about one out of four words in the reading materials.

*Participants*

The participants ($N$ = 9, see Table 5.4) attended French classes in Dutch-speaking education or were enrolled in a compulsory (extra-curricular) language class at a French-speaking university in Belgium. They were either at the elementary (A2, $n$ = 5) or the intermediate (B1, $n$ = 4) proficiency levels. The participants' proficiency level was determined based on the CEFR level targeted in the language course. The participants were native speakers of Chinese ($n$ = 1), Dutch ($n$ = 4), Japanese ($n$ = 1), or Spanish ($n$ = 3). They

**Table 5.4**
*Participants in Trial 1*

| Participant | Level | Language | ISCED 2011 |
|---|---|---|---|
| a2-ch-01 | A2 | zho | 6: Bachelor or equivalent |
| a2-es-01 | A2 | spa | 6: Bachelor or equivalent |
| a2-es-02 | A2 | spa | 8: Doctoral or equivalent |
| a2-nl-01 | A2 | nld | 2: Lower secondary education |
| a2-nl-02 | A2 | nld | 3: Upper secondary education |
| b1-es-01 | B1 | spa | 6: Bachelor or equivalent |
| b1-ja-01 | B1 | jpn | 6: Bachelor or equivalent |
| b1-nl-01 | B1 | nld | 3: Upper secondary education |
| b1-nl-02 | B1 | nld | 6: Bachelor or equivalent |

were either in secondary education ($n$ = 3), in higher education ($n$ = 5), or non-native scientific staff ($n$ = 1).

*Procedure*

The materials were presented to the participants via a web-based reading interface. Unique anonymized login credentials were generated beforehand and distributed to the participants to log into the system. After login, the interface showed a page with guidelines. The participants were instructed to read the guidelines and try out the judgment task on a dummy text. After reading the guidelines, the participants clicked to move on to another page where they started reading.

The guidelines for the difficulty judgment task were as follows. The participants were told that the study aimed to evaluate the appropriateness of reading materials for learners. In particular, they were told that their feedback was needed to judge whether the vocabulary was difficult or not. They were instructed to read the texts and highlight the words of which they did not know the meaning. Some guidelines were given to give learners an idea of what difficulty entailed:

> *I don't remember having seen this word before.*
> *I have seen this word before, but I don't know what it means.*

*I can't find a synonym/explanation for this word.*
*I can't translate this word in my native language.*
*I need to use a dictionary to understand the word.*

The participants needed to click in order to highlight a word. They could click as many times as needed to highlight/unhighlight/re-highlight a specific word. When they finished reading, they clicked on a button at the bottom of the page to move on to the next text.

The reading task was self-paced, and no time constraints were imposed. The participants could pause and resume their reading when needed. Each participant read all 51 texts and spent four hours on average to complete the task. The texts were presented in a random order to avoid order effects. The random seed generator was set to the participant's unique identifier. Before participation, all participants signed a consent form granting permission to use their data for research purposes. Participants also received a remuneration (€20) for their work.

### 5.1.2 *Trial 2*

#### *Materials*

The reading materials were drawn from a corpus of open-licensed texts of three genres, namely the descriptive (Wikivoyage), informative (Wikinews), and narrative (Project Gutenberg; *Jeunesse* and *Contes*) genres. A set of texts were selected from this corpus. Because the allotted time for the reading task was restricted to half an hour, the set of texts did not count more than 2,400 words in total. This estimation was done to allow for a comfortable reading speed (i.e., 80 words per minute).[61] The corpus was searched to retrieve texts counting between 100 and 500 words. Remaining texts were ranked according to Flesch's (1951) Reading Ease score (see Section 5.A.1). A separate set of texts was selected for each proficiency level, based on the readability rank and

61 The optimal reading rate is estimated at 300 words per minute when reading in the native language (Carver, 1982). However, because reading speed can be at least 50% slower in a non-native language (Fraser, 2007), Carver's (1982) lower bound of 80 words per minute was used to ensure the participants would not run into time constraints that would force them to scan the text instead of performing a regular reading.

the adequacy with the CEFR level.[62] As in the previous trial, the texts were selected to achieve a minimal content overlap between texts.

As a result, five texts were selected per each proficiency level. Each set included two narrative texts, two descriptive texts, and one informative text. The surface-level characteristics of these materials are described in Table 5.5. On average, the texts counted between 348 and 393 tokens, with 168 – 203 distinct types and 148 – 178 different lemmas per text. The median reading ease was *easy* ($Mdn_{RE}$ = 80.2) for the A2-level participants, *fairly easy* ($Mdn_{RE}$ = 73.2) for the B1-level participants, and *standard* for the B2-level participants ($Mdn_{RE}$ = 69.7) and the C1-level participants ($Mdn_{RE}$ = 65.2).[63] The surface-level characteristics and readability scores for each text are given in Table 5.15 in the Appendix (Section 5.A.2).

TARGET VOCABULARY Like in the previous trial, no selection procedure was used to define a set of target words; instead, all relevant words attested in the reading materials were also targeted in the study. Relevant lexical units were identified with the LGTagger (Constant & Sigogne, 2011) to detect multi-word units. All other non-alphanumeric tokens (e.g., digits, symbols, and punctuation marks) were irrelevant. In total, the reading materials included a set of 6,196 relevant tokens.

The distribution of FLELex vocabulary (Table 5.6) was also similar to the previous trial. For all levels taken together, approximately 57% of words appeared in a reference corpus of L2 reading materials, excluding 7% of relevant items that were proper nouns. On average, these words occurred once every $10^3$ words ($M_{SFI} \approx 70$) in L2 reading materials. Again, the majority of these words were introduced in materials intended for the A1 to B1 levels. Concerning the individual proficiency levels targeted in the trial, the texts for the A2–B1 levels contained relatively more basic-level words, whereas the texts for the B2–C1 levels contained relatively more advanced-level words. However, this difference was not very clear-cut.

62 The A2-level participants were given short tales for children, whereas the C1-level participants were given the original *Fables* by de La Fontaine.

63 Because Flesch's (1951) formula was developed to assess the readability of written prose, the reading ease scores were probably incorrectly estimated (RE > 90) for the two C1-level texts written in verse (Fables by de La Fontaine). The median reading ease was probably lower in reality (*fairly difficult*).

**Table 5.5**
*Reading Materials Trial 2*

| | | *M* | *SD* | min | *Q*1 | *Mdn* | *Q*3 | max |
|---|---|---|---|---|---|---|---|---|
| A2 | Sentences | 21.6 | 10.3 | 12 | 16 | 17 | 25 | 38 |
| | Tokens | 348 | 96.7 | 217 | 282 | 374 | 424 | 445 |
| | Types | 168 | 42.2 | 105 | 144 | 191 | 194 | 205 |
| | Lemmas | 148 | 36.2 | 94 | 127 | 168 | 172 | 178 |
| | RE | 74.3 | 10.9 | 58.3 | 68.0 | 80.2 | 82.4 | 82.9 |
| | RE KM | 89.3 | 10.2 | 73.6 | 84.5 | 95.3 | 96.5 | 96.8 |
| B1 | Sentences | 26.2 | 10.0 | 13 | 21 | 28 | 29 | 40 |
| | Tokens | 388 | 118 | 252 | 273 | 428 | 487 | 500 |
| | Types | 192 | 38.3 | 148 | 154 | 211 | 223 | 226 |
| | Lemmas | 170 | 32.9 | 132 | 136 | 186 | 194 | 200 |
| | RE | 73.5 | 12.2 | 57.9 | 70.3 | 73.2 | 74.0 | 91.9 |
| | RE KM | 88.8 | 11.0 | 75.1 | 85.7 | 87.6 | 90.0 | 106 |
| B2 | Sentences | 32.2 | 19.9 | 13 | 21 | 21 | 46 | 60 |
| | Tokens | 393 | 93.0 | 258 | 351 | 400 | 472 | 484 |
| | Types | 203 | 41.6 | 155 | 176 | 194 | 231 | 258 |
| | Lemmas | 178 | 29.6 | 139 | 165 | 172 | 196 | 216 |
| | RE | 72.4 | 14.0 | 57.5 | 61.4 | 69.7 | 83.0 | 90.5 |
| | RE KM | 88.1 | 12.9 | 74.5 | 78.3 | 85.2 | 97.9 | 105 |
| C1 | Sentences | 30.4 | 27.3 | 9 | 15 | 20 | 31 | 77 |
| | Tokens | 383 | 129 | 268 | 276 | 387 | 399 | 588 |
| | Types | 198 | 64.9 | 152 | 162 | 182 | 184 | 312 |
| | Lemmas | 174 | 50.1 | 136 | 139 | 162 | 174 | 259 |
| | RE | 71.9 | 20.5 | 48.8 | 58.5 | 65.2 | 93.3 | 93.6 |
| | RE KM | 87.2 | 19.1 | 65.5 | 75.3 | 80.4 | 107 | 107 |

RE = Reading Ease (de Landsheere, 1963; Flesch, 1951)
KM = Kandel and Moles (1958)

**Table 5.6**

*FLELex Vocabulary in Trial 2*

| | | | | SFI | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Level | *N* | % | *M* | *SD* | min | *Q*1 | *Mdn* | *Q*3 | max |
| A2 | A1 | 727 | 51.5 | 69.9 | 13.3 | 36.9 | 60.6 | 72.2 | 83.0 | 86.3 |
| | A2 | 31 | 2.19 | 44.5 | 6.96 | 31.1 | 37.9 | 45.7 | 49.4 | 54.5 |
| | B1 | 33 | 2.34 | 40.1 | 5.01 | 33.1 | 36.7 | 38.1 | 45.1 | 51.4 |
| | B2 | 2 | 0.14 | 41.9 | 3.22 | 39.6 | 40.7 | 41.9 | 43.0 | 44.2 |
| | C1 | 6 | 0.42 | 48.3 | 4.82 | 47.8 | 47.8 | 47.8 | 47.8 | 50.8 |
| | C2 | 2 | 0.14 | 52.7 | 0 | 52.7 | 52.7 | 52.7 | 52.7 | 52.7 |
| | Total | 801 | 56.7 | 68.2 | 14.5 | 21.9 | 59.2 | 69.0 | 80.7 | 87.4 |
| B1 | A1 | 808 | 52.5 | 68.4 | 14.5 | 32.3 | 59.0 | 70.3 | 83.0 | 86.3 |
| | A2 | 34 | 2.21 | 44.0 | 8.15 | 31.1 | 37.3 | 42.7 | 52.1 | 58.0 |
| | B1 | 41 | 2.66 | 43.4 | 6.33 | 32.7 | 37.4 | 45.5 | 47.9 | 57.5 |
| | B2 | 5 | 0.32 | 46.2 | 5.65 | 45.4 | 45.8 | 47.8 | 49.3 | 53.0 |
| | C1 | 4 | 0.26 | 47.5 | 2.40 | 45.4 | 46.0 | 47.0 | 48.5 | 50.8 |
| | C2 | 0 | 0.0 | | | | | | | |
| | Total | 892 | 57.9 | 67.6 | 14.1 | 17.9 | 57.3 | 67.6 | 80.0 | 87.4 |
| B2 | A1 | 794 | 47.9 | 69.5 | 14.1 | 36.9 | 59.0 | 72.9 | 83.0 | 86.3 |
| | A2 | 84 | 5.07 | 44.1 | 7.05 | 32.7 | 37.9 | 40.8 | 49.8 | 59.6 |
| | B1 | 43 | 2.60 | 42.5 | 7.01 | 30.9 | 36.7 | 38.9 | 47.6 | 57.5 |
| | B2 | 11 | 0.66 | 43.0 | 3.58 | 39.1 | 40.3 | 41.2 | 45.4 | 49.6 |
| | C1 | 10 | 0.60 | 47.1 | 1.48 | 45.4 | 46.2 | 47.0 | 47.8 | 50.1 |
| | C2 | 2 | 0.12 | 48.6 | 3.57 | 46.0 | 47.3 | 48.6 | 49.8 | 51.1 |
| | Total | 944 | 57.0 | 67.0 | 14.9 | 23.4 | 56.5 | 68.4 | 80.0 | 87.4 |
| C1 | A1 | 754 | 48.6 | 69.2 | 14.3 | 32.3 | 57.9 | 71.3 | 83.4 | 86.3 |
| | A2 | 71 | 4.57 | 44.7 | 7.36 | 32.9 | 37.9 | 46.0 | 50.3 | 58.5 |
| | B1 | 43 | 2.77 | 42.8 | 5.63 | 34.9 | 37.7 | 43.9 | 47.6 | 54.2 |
| | B2 | 10 | 0.64 | 42.1 | 2.00 | 39.1 | 40.0 | 43.6 | 43.6 | 43.6 |
| | C1 | 6 | 0.39 | 45.4 | 0 | 45.4 | 45.4 | 45.4 | 45.4 | 45.4 |
| | C2 | 1 | 0.06 | 46.0 | 0 | 46.0 | 46.0 | 46.0 | 46.0 | 46.0 |
| | Total | 885 | 57.0 | 67.3 | 14.8 | 23.4 | 56.7 | 67.8 | 81.3 | 87.4 |

**Table 5.7**

*Cloze Procedure for Identifying High and Low Constraint Contexts in Trial 2*

| | | | $k$ = 1 | | $k$ = 5 | | $k$ = 10 | | $k$ = 20 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Predictions | | *N* | % | *N* | % | *N* | % | *N* | % |
| A2 | Masked | 1 | 807 | 57 | 1,006 | 71 | 1,044 | 74 | 1,084 | 77 |
| | | 0 | 606 | 43 | 407 | 29 | 369 | 26 | 329 | 23 |
| | Causal | 1 | 108 | 8 | 389 | 28 | 527 | 37 | 658 | 47 |
| | | 0 | 1,305 | 92 | 1,024 | 72 | 886 | 63 | 755 | 53 |
| B1 | Masked | 1 | 851 | 55 | 1,111 | 72 | 1,161 | 75 | 1,201 | 78 |
| | | 0 | 689 | 45 | 429 | 28 | 379 | 25 | 339 | 22 |
| | Causal | 1 | 129 | 8 | 393 | 26 | 548 | 36 | 687 | 45 |
| | | 0 | 1,411 | 92 | 1,147 | 74 | 992 | 64 | 853 | 55 |
| B2 | Masked | 1 | 828 | 50 | 1,066 | 64 | 1,123 | 68 | 1,176 | 71 |
| | | 0 | 829 | 50 | 591 | 36 | 534 | 32 | 481 | 29 |
| | Causal | 1 | 112 | 7 | 390 | 24 | 519 | 31 | 653 | 39 |
| | | 0 | 1,545 | 93 | 1,267 | 76 | 1,138 | 69 | 1,004 | 61 |
| C1 | Masked | 1 | 684 | 44 | 933 | 60 | 1,001 | 64 | 1,053 | 68 |
| | | 0 | 868 | 56 | 619 | 40 | 551 | 36 | 499 | 32 |
| | Causal | 1 | 87 | 6 | 304 | 20 | 448 | 29 | 556 | 36 |
| | | 0 | 1,465 | 94 | 1,248 | 80 | 1,104 | 71 | 996 | 64 |

1 = correct prediction (high-constraint context)
0 = incorrect prediction (low-constraint context)
$k$ = number of most probable predictions

A more clear-cut difference was observed in the proportions of high and low contextual constraints across the various sets of texts. Table 5.7 shows that 6 – 8% of words could be immediately predicted from their antecedent contexts. This result was similar to the first trial. Again, approximately 7% of words were highly predictable. The A2-level texts contained more (57%) words immediately predicted from the surrounding context and, on the other hand, contained fewer (23%) words extremely difficult to predict, even considering the top 20 most probable predictions. Conversely, the C1-level texts contained fewer (44%) words immediately predicted from the surrounding context and, on the other hand, contained more (32%) words extremely difficult to predict.

**Table 5.8**
*Participants in Trial 2*

| Level | Language | ISCED 2011 | Age |
|---|---|---|---|
| A2 | nld | 2: Lower secondary education | 13–14 years |
| B1 | nld | 6: Bachelor or equivalent | 18–19 years |
| B2 | nld | 6: Bachelor or equivalent | 18–19 years |
| C1 | nld | 7: Master or equivalent | 21–22 years |

*Participants*

The participants ($N$ = 47, see Table 5.8) were Dutch-speaking learners attending French classes in Dutch-speaking education in Belgium. They were either secondary school students (A2, $n$ = 10), freshman undergraduate students enrolled in a legal French course (B1, $n$ = 17), freshman undergraduate students majoring in French (B2, $n$ = 12), or graduate students majoring in French (C1, $n$ = 8).

*Procedure*

The procedure for the second trial was the same as the first trial: the same interface, guidelines, and methodology were used. The participants also signed a consent form and were remunerated (cinema ticket) for their work. However, the main difference with the first trial was that a 30-minute vocabulary post-test followed the 30-minute reading task. The data and analyses for this vocabulary test are described in more detail in Vanhauwaert (2017); they were not used in this study. In brief, the aim was to examine how the participants noticed difficulty on polylexical units such as *prophète de malheur* 'prophet of doom' and whether there was an agreement between their subjective judgments and their results on the vocabulary test.

## 5.2 LEARNER DATA ANALYSIS

Multiple analyses were performed on the data collected in the first and second trials to answer the study's two research questions. Concerning the first

research question (RQ5.1), a series of descriptive and inter-rater agreement analyses were performed to examine differences in the learner-specific distributions of difficulty. Concerning the second research question (RQ5.2), a series of statistical tests and mixed-effects models (with *lme4*, Bates et al., 2015) were run to investigate the effect of high and low contextual constraints.

### 5.2.1 *Learner-Specific Distributions of Difficulty*

In previous studies (Paetzold & Specia, 2016a; Yimam et al., 2018), individual judgments were aggregated into a single distribution of difficulty. This distribution of difficulty (5.1) can be defined as a binomial distribution with two parameters:

$$Y \sim B(N, p) \tag{5.1}$$

where

$N$ is the number of words (or observations) read

$p$ is the probability that a word (or observation) was found difficult

Paetzold and Specia (2016a) collected a total of 232,481 observations, with an average of 581 words read per 400 participants[64] and with 2.7% of difficult words. Yimam et al. (2018) collected four data sets: one for English, with 34,879 observations and with 41.4% of difficulty; one for Spanish, with 17,605 observations and with 39.8% of difficulty; one for German, with 7,905 observations and with 41.4% of difficulty; and one for French, with 2,251 observations and with 29% of difficulty.[65]

In comparison, Table 5.9 shows the number of observations and percentages of difficulty for the data collected in this study. The data for the first trial included a total of 189,423 observations, with an average of 21,009 observations per participant and 5.3% of difficult words. The data for the second trial

64 Because the data sets in Paetzold and Specia (2016a) and Yimam et al. (2018) did not provide a unique identifier for each participant, the actual average number of observations per participant could not be computed. Instead, the total number of observations was divided by the total number of participants. In Yimam et al. (2018), the total number of participants was not known.

65 The percentages of difficult words were high in Yimam et al. (2018) because the authors filtered many tokens (stopwords).

**Table 5.9**

*Number of Observations and Percentages of Difficulty in Trials 1 & 2*

| | | | Per Individual | | | Per Individual | | | |
|---|---|---|---|---|---|---|---|---|---|
| Trial | | *N* | *M* | *SD* | *p* | *M* | *SD* | min | max |
| 1 | All | 189,084 | 21,009 | 71.9 | .053 | .053 | .034 | .014 | .12 |
| | A2 | 104,964 | 20,993 | 50.4 | .061 | .061 | .039 | .014 | .12 |
| | B1 | 84,120 | 21,030 | 96.9 | .043 | .044 | .029 | .024 | .087 |
| 2 | All | 72,970 | 1,553 | 83.6 | .040 | .041 | .031 | .010 | .15 |
| | A2 | 14,230 | 1,423 | 0 | .073 | .073 | .043 | .037 | .15 |
| | B1 | 26,248 | 1,544 | 0 | .039 | .039 | .027 | .010 | .10 |
| | B2 | 19,980 | 1,665 | 0 | .027 | .027 | .010 | .012 | .047 |
| | C1 | 12,512 | 1,564 | 0 | .024 | .024 | .010 | .011 | .040 |

*Note.* Although the participants in the first trial received the same texts, the number of observations slightly varied because some participants also highlighted difficulty in the title, whereas others skipped the titles altogether.
*N* = number of observations
*p* = percentage/probability of difficulty

included a total of 72,970 observations, with an average of 1,553 observations per participant and 4% of difficult words. Contrary to the studies cited above, both trials comprised many more observations per participant. Because the data for both trials included unique identifiers per participant, it was possible to derive a distribution of difficulty for each learner $i$ (5.2):

$$Y_i \sim B(N_i, p_i) \tag{5.2}$$

where

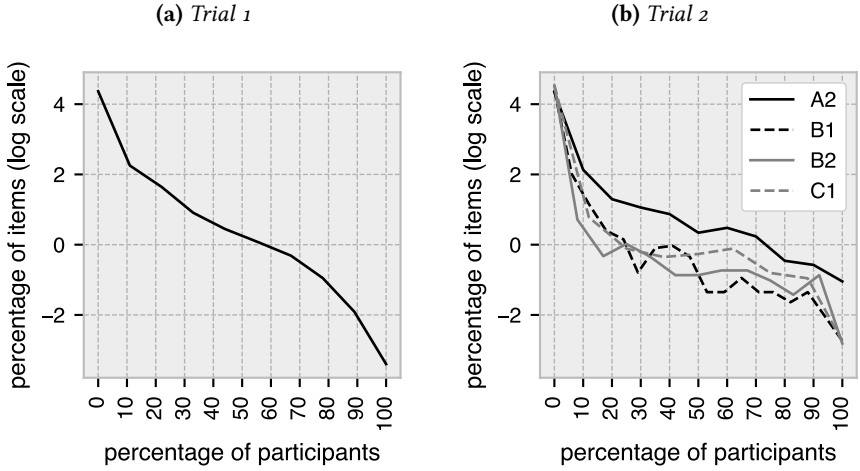
$N_i$ is the number of words read by the learner

$p_i$ is the probability that the learner found a word difficult

The various learner-specific distributions of difficulty were characterized by two sources of variance, shown in Table 5.9. Firstly, some learners read more or less extensively (parameterized as $N$) than others. Of course, the difference in the number of words read per participant was controlled by the

**Figure 5.1**

*Percentages of Participants Judging a Word as Difficult in Trials 1 & 2*




data collection procedure: the participants in Trial 1 were given more reading materials than those in Trial 2. It is conceivable to imagine, however, that there would be an even higher variance in the number of words read per individual if the reading task were uncontrolled and unconstrained, that is, when the learners had a free choice in the type and number of texts they read.

Secondly, Table 5.9 shows that some learners experienced more or less difficulty (parameterized as $p$) than others. In both trials, the average learner found 4 – 5% of words difficult, with a minimum of 1% and a maximum of 15%. Moreover, the percentage of difficult words also decreased noticeably as the proficiency level increased. The A2-level participants in both trials experienced the largest extent of difficulty ($M_{\text{Trial 1}}$ = 6.1%, $M_{\text{Trial 2}}$ = 7.3%), whereas the C1-level participants in Trial 2 experienced the least amount of difficulty ($M_{\text{Trial 2}}$ = 2.4%). Importantly, the extent of difficulty varied between proficiency levels and within proficiency levels. Indeed, Table 5.9 shows that learners did not display the same extent of difficulty, even when they read the same number of texts, had the same proficiency level, and had the same L1 (Trial 2).

The same observation was made when looking at the percentage of learners perceiving the same item as difficult. Figure 5.1 shows that less than 1%

**Table 5.10**

*Agreement Between Participants in Trials 1 & 2*

| | | Trial 1 | | | Trial 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | | All | A2 | B1 | A2 | B1 | B2 | C1 |
| Perfect[a] | % | 78.9 | 81.7 | 89.1 | 77.2 | 81.7 | 92.7 | 93.7 |
| | $n_1$ | 7 | 22 | 92 | 5 | 1 | 1 | 1 |
| Partial[b] | % | 11.6 | 7.6 | 3.5 | 14.4 | 10.8 | 5.3 | 4.1 |
| No[c] | % | 9.5 | 10.7 | 7.4 | 8.4 | 7.5 | 2.0 | 2.2 |
| Krippendorff's $\alpha$ | | .26 | .23 | .30 | .40 | .36 | .51 | .47 |
| ICC | | .12 | .15 | .08 | .09 | .13 | .04 | .05 |

$n_1$ = number of words found difficult
[a] words found (non-)difficult by all participants
[b] words found difficult by at least two participants
[c] words found difficult by one participant

(log < 0) of items were found difficult by all (100%) learners, whereas most items (78.9% in Trial 1; 76.8 – 93.7% in Trial 2) were not found difficult by any learner (0%). The agreement scores presented in Table 5.10 further corroborated this finding: although there was a high agreement between learners, this was mainly concerned with non-difficulty rather than difficulty. In other words, there was only a small set of words that were perceived as difficult by all learners. This set of words can be found in Tables 5.16 and 5.17 in the Appendix (Section 5.A.2).

Krippendorff's (2011) $\alpha$ coefficient of inter-rater reliability was also computed. Table 5.10 shows a low agreement between participants ($\alpha$ = .26) in Trial 1, with a higher agreement between B1 learners than between A2 learners. These low agreement scores were similar to those of Paetzold and Specia (2016a). In Trial 2, there was a slightly higher overall agreement, with a lower agreement for the A2 ($\alpha$ = .40) and B1 ($\alpha$ = .36) levels and a higher agreement for the B2 ($\alpha$ = .51) and C1 ($\alpha$ = .47) levels. Still, all $\alpha$ coefficients remained below the accepted threshold of inter-rater reliability $\alpha$ = .8 (Artstein & Poesio, 2008, p. 576). However, this result did not necessarily show that the individual judgments and, by extension, the data sets were unreliable. A low $\alpha$ coefficient

seems to indicate that it would be unreliable to aggregate these individual judgments into a single ground truth of difficulty based on which a single explanatory or predictive model would be trained. Consequently, the result underscored the necessity to take into account individual differences in the modeling process.

The intra-class correlation coefficient was used to evaluate the possibility of taking into account the individual learner in the statistical analyses. The ICC was computed with the *lme4* library in R (Bates et al., 2015) by fitting a null mixed-effects logistic regression model (`lme4::glmer`) with random effects for individuals. The ICCs for each trial are given in Table 5.10. The results showed that subject-specific effects accounted for 12 – 13% of the variance in both trials combined. Based on the general rule of thumb that individual effects could be accounted for if the ICC is greater than 5% (Heck & Thomas, 2008), it was concluded that a mixed-effects approach could be adopted.

### 5.2.2 *High and Low Semantically Constrained Contexts*

The second research question was whether the degree of semantic constraints interacted with how foreign language learners perceived lexical difficulty in reading. The distinction was made between two types of contextual constraints. On the one hand, high semantic constraints were observed in sentences where an unknown word (i.e., a masked word) could be easily predicted from the context. On the other hand, low semantic constraints were observed in sentences where an unknown word could not be predicted from the context. In this study, two types of high and low semantic constraints were determined with the predictions of a state-of-the-art language model.

The first type of constraint was determined by predicting the next word from the antecedent context. It was hypothesized that words predicted as the most probable next word (i.e., highly contextually constrained) would also be read with greater ease (i.e., would not be perceived as difficult). However, the results showed that these causal language model predictions interacted only marginally with how learners noticed difficult words. Table 5.11 shows a significant but negligible negative association between the degree of constraints and the degree of difficulty, both in the first trial ($\phi = -.029$) and in

**Table 5.11**

*Strength of Association Between High & Low Contextual Constraints and Difficult & Non-Difficult Words*

| Cloze | $k$ | $\chi^2$ | df | $p$ | $V$ | Effect | $\phi$ | Correlation |
|---|---|---|---|---|---|---|---|---|
| | | | | TRIAL 1 | | | | |
| Masked | 1 | 5,801 | 1 | <.001 | .18 | small | −.18 | negligible |
| | 5 | 7,786 | 1 | <.001 | .20 | small | −.20 | weak |
| | 10 | 8,167 | 1 | <.001 | .21 | small | −.21 | weak |
| | 20 | 8,067 | 1 | <.001 | .21 | small | −.21 | weak |
| Causal | 1 | 159 | 1 | <.001 | .029 | no | −.029 | negligible |
| | 5 | 1,285 | 1 | <.001 | .082 | no | −.082 | negligible |
| | 10 | 2,165 | 1 | <.001 | .11 | small | −.11 | negligible |
| | 20 | 3,055 | 1 | <.001 | .13 | small | −.13 | negligible |
| | | | | TRIAL 2 | | | | |
| Masked | 1 | 1,461 | 1 | <.001 | .14 | small | −.14 | negligible |
| | 5 | 2,109 | 1 | <.001 | .17 | small | −.17 | negligible |
| | 10 | 2,278 | 1 | <.001 | .18 | small | −.18 | negligible |
| | 20 | 2,400 | 1 | <.001 | .18 | small | −.18 | negligible |
| Causal | 1 | 63.9 | 1 | <.001 | .030 | no | −.030 | negligible |
| | 5 | 311 | 1 | <.001 | .065 | no | −.065 | negligible |
| | 10 | 555 | 1 | <.001 | .087 | no | −.087 | negligible |
| | 20 | 743 | 1 | <.001 | .10 | small | −.10 | negligible |

$k$ = number of most probable predictions
$\chi^2$ = chi-squared test of independence, without continuity correction
$V$ = Cramer's $V$
$\phi$ = Pearson's Phi coefficient of binomial correlation

the second trial ($\phi$ = −.030). In other words, there was no strong association between words that could be easily predicted from the preceding context and words that were not noticed as difficult.

The second type of constraint was determined by predicting a masked word from the surrounding context. It was hypothesized that words that were among the k most probable words (i.e., highly contextually constrained) would also be perceived as non-difficult. The results showed that these masked language model predictions correlated more with how learners noticed difficult words than the causal language model predictions. Table 5.11 shows an overall

**Table 5.12**

*Percentages of Low Semantic Constraints Per Difficulty & Non-Difficulty*

| | | Low-Constraint Contexts (%) | | | | |
|---|---|---|---|---|---|---|
| Perfect Agreement | *n* | $k = 1$ | $k = 5$ | $k = 10$ | $k = 20$ | |
| | | TRIAL 1 | | | | |
| 100% difficult | 7 | 100 | 100 | 100 | 100 | |
| 0% difficult | 16,532 | 38.5 | 22.3 | 19.0 | 16.7 | |
| | | TRIAL 2 | | | | |
| 100% difficult | 5 | 80 | 60 | 60 | 60 | A2 |
| 0% difficult | 1,083 | 34.9 | 22.3 | 19.7 | 17.9 | |
| 100% difficult | 1 | 100 | 100 | 100 | 100 | B1 |
| 0% difficult | 1,257 | 39.0 | 21.0 | 18.2 | 15.9 | |
| 100% difficult | 1 | 100 | 100 | 100 | 100 | B2 |
| 0% difficult | 1,534 | 46.9 | 31.9 | 28.2 | 24.9 | |
| 100% difficult | 1 | 100 | 100 | 100 | 100 | C1 |
| 0% difficult | 1,453 | 53.3 | 36.5 | 31.9 | 28.3 | |

*Note.* The low contextual constraints were defined as the words that could not be predicted with a masked language model.
$n$ = number of items
$k$ = number of most probable predictions

significant but weak negative correlation between the degree of constraints and the degree of difficulty. In Trial 1, the most significant association was found between words perceived as difficult and not be predicted from the $k = 10$ most probable words ($\phi = -.21$). In Trial 2, the most significant association was found between words perceived as difficult and not be predicted from the $k = 20$ most probable words ($\phi = -.18$).

Further exploration investigated the effect of semantic constraints on the items that achieved a perfect between-learner agreement of (non-)difficulty. Table 5.12 shows that the set of words perceived as difficult by all participants in the same trial almost solely appeared in weakly constrained contexts. An exception was two words that were easy to predict from the surrounding context but which were still found difficult by all A2 learners in Trial 2:

- "(...) un site du patrimoine[66] mondial de l'UNESCO (...)" ('a UNESCO World Heritage Site'), which the masked language model predicted as the most probable word given the context; and

- "(...) il faut emprunter l'autoroute 2 (...)" ('you have to take highway 2'), which the masked language model predicted as the second most probable word (after *suivre* 'follow').

Now, the effect of low-constraint contexts was less clear-cut for words perceived as non-difficult by all learners. A small but non-trivial percentage (16 – 28%) of words not perceived as difficult by all learners appeared in contexts that were only weakly constrained, that is, when the $k = 20$ most probable words were considered.

A final analysis was undertaken to investigate the interaction between the degree of semantic constraints and the between-learner variance observed previously (Section 5.2.1). A single-factor mixed-effects logistic regression model was estimated to predict difficulty for both trials. The models were estimated with maximum likelihood and Nelder-Mead optimization. The models included a single fixed effect for correct cloze predictions (`cloze_k`) and random effects for learners (`sbj_id`), proficiency levels (`pro_level`), and cloze predictions. In Trial 1, the most likely model (GLMM 5.1) was found for cloze predictions determined by the $k = 10$ most probable words. In Trial 2, the most likely model (GLMM 5.2) was found for cloze predictions determined by the $k = 20$ most probable words.

$$\text{difficulty} \sim \text{cloze_10} + (\text{cloze_10}|\text{pro_level:sbj_id}) \qquad \text{(GLMM 5.1)}$$

$$\text{difficulty} \sim \text{cloze_20} + (\text{cloze_20}|\text{pro_level:sbj_id}) \qquad \text{(GLMM 5.2)}$$

The models achieved a substantial explanatory power, both in the first trial, $R^2_{\text{conditional}} = .31$; and in the second trial, $R^2_{\text{conditional}} = .44$. The models' single fixed effect (viz., a high or low contextual constraint) achieved a significantly

66 The observation that *patrimoine* was perceived as difficult by all participants was surprising seeing that the word was etymologically related to the Latin loanword *patrimonium* in Dutch, which was the native language of the participants. Therefore, the difficulty of understanding *patrimoine* was also hindered by the learners' lack of more extensive lexical knowledge in their native language.

large[67] negative effect on difficulty, both in the first trial, $\beta = -1.98$, $SE = 0.12$, $e^{-\beta} = 7.24$, $p < .001$; and in the second trial, $\beta = -2.44$, $SE = 0.14$, $e^{-\beta} = 11.5$, $p < .001$. A likelihood ratio test showed that the part explained by the fixed effect alone was significant in both the first trial, $R^2_{\text{marginal}} = .16$, $\chi^2(1) = 28.8$, $p < .001$; and the second trial, $R^2_{\text{marginal}} = .20$, $\chi^2(1) = 96.1$, $p < .001$. Post hoc Monte Carlo power simulations were performed on both models with the *simr* package in R (Green & MacLeod, 2016). The observed power calculations showed that a sample size of $N_{\text{participants}} = 2$ (in Trial 1) and $N_{\text{participants}} = 5$ (in Trial 2) would have been enough to find a large effect size ($\approx d = .8$) for contextual constraints with at least 80% statistical power (see Figures 5.4 and 5.5 in the Appendix; Section 5.A.3).

Concerning the models' random effects (see Figure 5.2), the between-intercepts variance was relatively similar between the first trial ($\hat{\sigma}^2_{u0} = 0.39$) and the second trial ($\hat{\sigma}^2_{u0} = 0.32$). In other words, there was a similar variance in the extent of difficulty perceived by learners. However, the between-slopes variance was considerably higher in Trial 2 ($\hat{\sigma}^2_{u1} = 0.70$) than in Trial 1 ($\hat{\sigma}^2_{u1} = 0.16$). In other words, there was more individual variance in the effect of low-constraint contexts on difficulty perceived by participants in the second trial. Figure 5.2 shows that this variance was greatest between the A2 and B1 participants.

Instead of defining contextual constraints in a dichotomous manner (i.e., by distinguishing high from low constraints), a more fine-grained analysis may be obtained by looking at how unlikely the masked language model found the word in the given context. Based on the softmax probability attributed by the model, the degree of contextual surprisal was measured as the base-10 logarithm of the reciprocal probability (see Equation (3.6) in Chapter 3). Again, a single-factor mixed-effects logistic regression model (GLMM 5.3) was estimated for both trials (Figure 5.3).

$$\text{difficulty} \sim \text{surprisal} + (\text{surprisal}|\text{pro_level:sbj_id}) \qquad \text{(GLMM 5.3)}$$

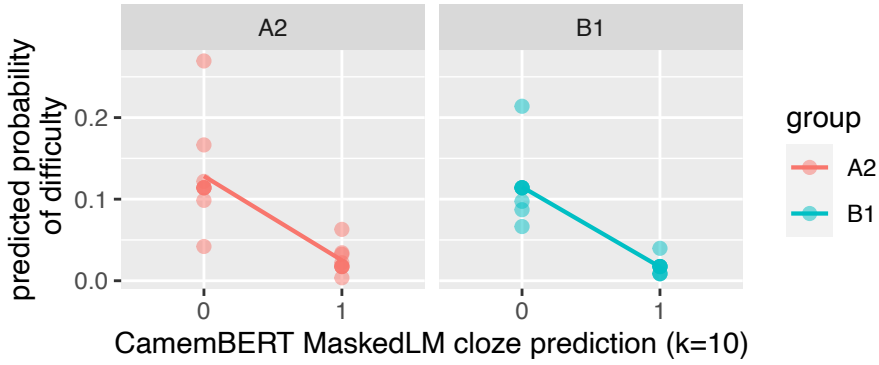
A likelihood ratio test showed that the part explained by the fixed effect was significant for both the first trial, $R^2_{\text{marginal}} = .17$, $\chi^2(1) = 36.8$, $p < .001$; and

67 Effect sizes were determined following Chen et al. (2010), who advanced an interpretation of odds ratios in terms of Cohen's *d* effect sizes.

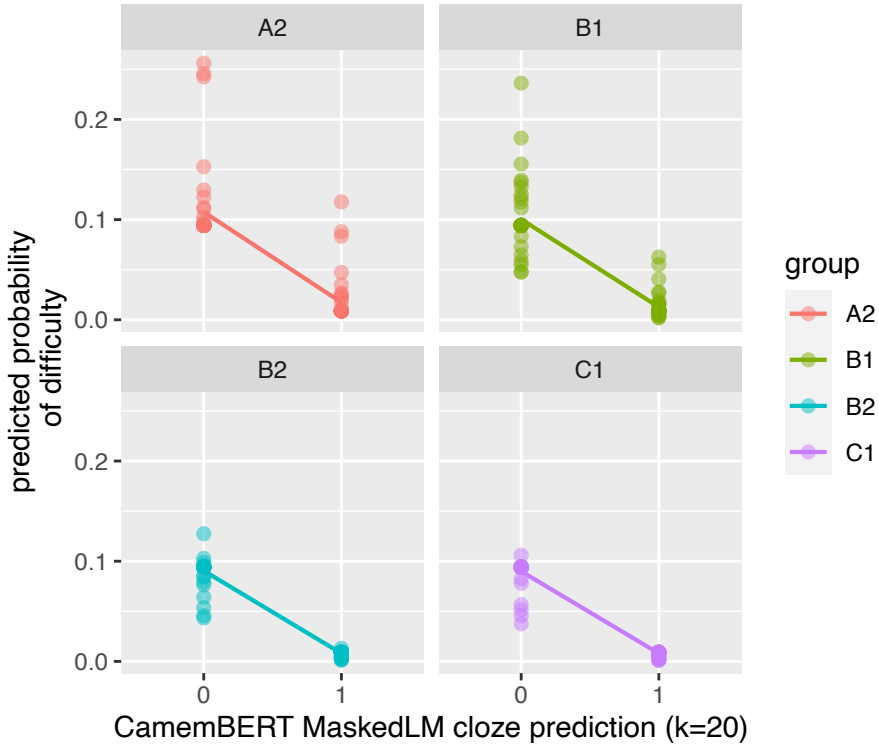
**Figure 5.2**

*Predictions of Difficulty from a Logistic Mixed-Effects Model With Fixed Effects of Low (0) and High (1) Contextual Constraints*

**(a)** *Trial 1*




**(b)** *Trial 2*



the second trial, $R^2_{\text{marginal}} = .19$, $\chi^2(1) = 148$, $p < .001$. However, the degree of contextual surprisal only achieved a small positive effect on difficulty, both in the first trial, $b = 0.59$, $SE = 0.03$, $\beta = 0.89$, $e^{\beta} = 2.44$, $z = 23.0$, $p < .001$; and in the second trial, $b = 0.63$, $SE = 0.02$, $\beta = 1.01$, $e^{\beta} = 2.75$, $z = 31.3$, $p < .001$. In other words, the more surprising a word was in a given context, the higher the likelihood that it was perceived as difficult. However, the effect did not surpass the large effect observed for low contextual constraints. Moreover, the effect of word surprisal did not vary greatly between learners, as manifested in the low between-slopes variance in both the first ($\hat{\sigma}^2_{u1} = 0.006$) and second ($\hat{\sigma}^2_{u1} = 0.01$) trials.

## 5.3 CONCLUSION

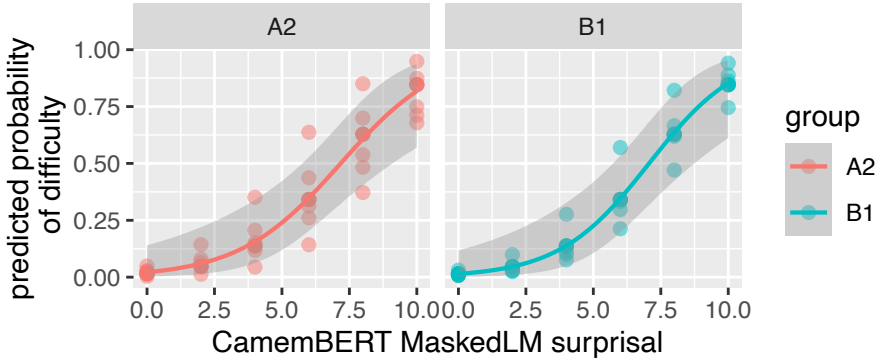
This chapter reported an empirical study of how French learners as a foreign language perceived lexical difficulty when reading. Two trials were conducted with a standard procedure that aimed to obtain a direct measure of difficulty, requiring the participants to highlight unknown words. A total of 262,054 observations were collected. A descriptive analysis of these measurements showed that the distributions of difficulty were strongly imbalanced, with $4 - 5\%$ of words perceived as difficult on average.

Although the procedure was similar to previous benchmarks that also used the same direct measure of difficulty while reading (Paetzold & Specia, 2016a; Yimam et al., 2018), the methodology also presented two important novelties, namely that (a) the observations were comparable because all learners in the same group read the same texts and that (b) each observation was linked to a learner identified by an anonymous ID. As a result, various ($N = 56$) learner-specific distributions of difficulty could be defined. From these individual distributions, a general trend was observed: most words (in a given context) were perceived as non-difficult by all learners, while very few words were perceived as difficult by all learners. Therefore, the results showed a non-negligible variance in how individuals at the same proficiency level and having the same native language noticed difficulty reading the same texts. Moreover, the inter-rater agreement was not strong, as shown by the low to moderate $\alpha$ coefficients ($\alpha \in [.23, .51]$). Consequently, it would not be reliable to model

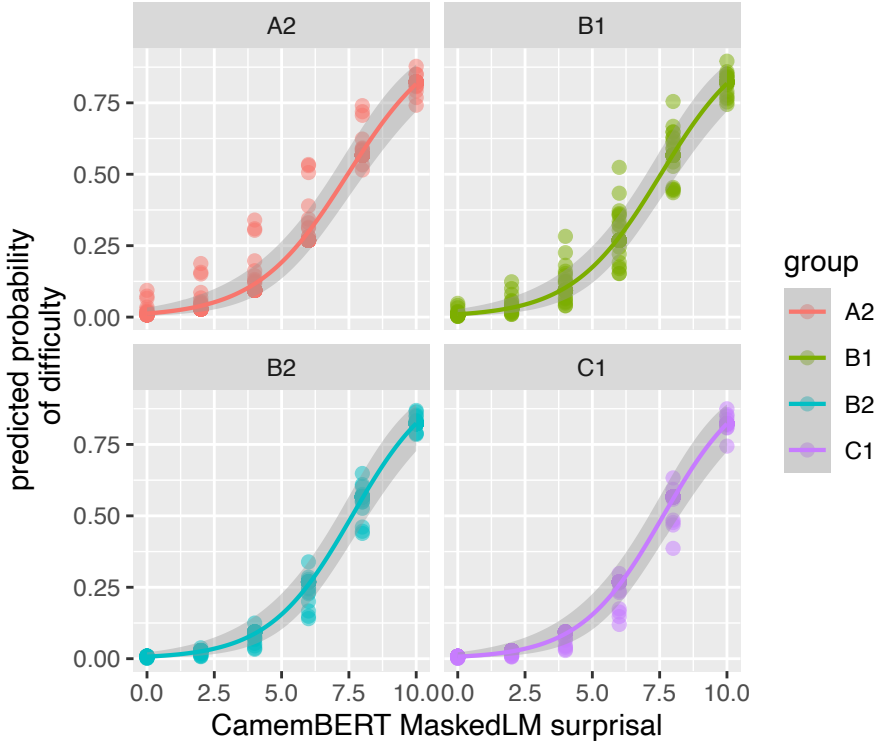
**Figure 5.3**

*Predictions of Difficulty from a Logistic Mixed-Effects Model With Fixed Effects of Contextual Surprisal*

**(a)** *Trial 1*




**(b)** *Trial 2*



difficulty based on a ground truth where these individual measurements are aggregated into a single measure. For this reason, these individual differences will be further explored in explanatory and predictive models of difficulty that will be presented in Part iii.

Furthermore, the analyses showed that the degree of contextual constraint achieved a sizable negative effect on perceived difficulty. Words that occurred in low semantically constrained contexts (i.e., which did not enable correctly predicting the word given the $k$ most probable predictions of a state-of-the-art language model) were more likely to be perceived as difficult. These results were consistent with the well-established effect that was observed in Chapter 2 regarding contextual cues, constraints, and elaborations; and which has been evidenced in various studies on EEG potentials (Chen et al., 2017), lexical inferencing (Hamada, 2015; Kaivanpanah & Rahimi, 2017), and vocabulary learning (Birjandi et al., 2015; Ma et al., 2016; Sun, 2014; Webb, 2008).

Nevertheless, the methodology adopted in this study presented several limitations that require further investigation. On the one hand, the overall sample size was limited to 56 learners due to practical restrictions. In the first trial, it was challenging to obtain extensive reading data ($> 20K$ items) for many participants. Despite this limitation, the observed power calculations showed that the significant effect for high and low contextual constraints achieved sufficient statistical power ($> 80\%$). This was probably because many items and observations were used in the statistical analyses, contrary to most studies reviewed in Chapter 2.

On the other hand, it may be interesting to investigate further the impact of task effects on the measure of lexical difficulty used in this study. It is imperative to contrast these direct measures of perceived lexical difficulty with direct measures of actual difficulty obtained from a vocabulary test. Vanhauwaert (2017) found the post-test results showing a discrepancy between the perceptions of difficulty on multi-word units and the results on the post-test. Although most participants did not systematically highlight these units as difficult, they could not correctly recognize their meaning during the vocabulary post-test. In other words, the participants did not adequately direct their attention towards noticing difficulty on multi-word units. A future study could

look into the effect of attention-drawing techniques, which have shown to increase form recall on formulaic sequences (e.g., Peters, 2012b).

Lastly, it should be noted that the reading task was not followed by a reading comprehension test, as was the case in Paetzold and Specia (2016a) and Yimam et al. (2018), which could be a limitation. Reading comprehension tests are used to ensure the participants are reading for meaning. The reason why no comprehension test was used in this study was twofold: (a) to use a methodology similar to the previous studies and (b) to measure perceptions of difficulty in a more authentic context. For instance, when reading the news or reading a story, the reader is usually not presented with a reading comprehension test afterward. The aim was, therefore, to elicit more natural reading behavior. Nevertheless, it seemed that the effect for high and low semantically constrained contexts persisted even when reading for meaning was not explicitly controlled.

## 5.A APPENDIX

This appendix provides some details on the data collection procedure. The appendix comprises three pieces of information: the Reading Ease formula (Section 5.A.1), an overview of the reading materials used in trials one and two (Section 5.A.2), and the results of the power analyses (Section 5.A.3).

### 5.A.1 *Reading Ease*

The reading ease was computed with the following two formulae proposed for French:

$$\text{RE} = 206.835 - 1.015\,X_1 - 84.6\,X_2 \tag{5.3}$$

$$\text{FL} = 207 - 1.015\,X_1 - 73.6\,X_2 \tag{5.4}$$

where

RE Flesch (1951)'s Reading Ease (proposed by de Landsheere, 1963)

FL *facilité de lecture* 'reading ease' (Kandel & Moles, 1958)

$X_1$ measures the syntactic complexity of the text
(i.e., the average sentence length in number of words per sentence)

$X_2$ measures the lexical complexity of the text
(i.e., the average word length in number of syllables per word)

The original interpretation of Flesch's (1951) reading ease scores is in Definition 5.1.

### 5.A.2 *Reading Materials*

- The materials used in Trial 1 are given in Table 5.14.
- The materials used in Trial 2 are given in Table 5.15.
- The list of words found difficult by all participants in the same trial are given in Tables 5.16 and 5.17.

### 5.A.3 *Power Analyses*

Statistical power analyses were conducted were conducted to examine how many participants were needed. Brysbaert and Stevens (2018), for instance, found that a reaction time study should have at least 1,600 observations (i.e., 40 × 40 participants and items) per experimental condition in order to obtain a properly powered mixed-effects analysis. Moreover, they also showed that fewer participants would be needed if more items were presented (and vice versa). Statistical power was computed with *G*Power* (Faul et al., 2007). For generalized linear mixed-effects models, Monte Carlo power simulations were computed in R with the *simr* package (Green & MacLeod, 2016).

- The output of a post hoc (observed) power analysis for the GLMM 5.1 model is given in Figure 5.4.
- The output of a post hoc (observed) power analysis for the GLMM 5.2 model is given in Figure 5.5.

**Definition 5.1: Flesch Reading Ease Score**

| Score | Description | School Level |
|---|---|---|
| 100 | | |
| | *very easy* | 5th grade (10-11 years) |
| 90 | | |
| | *easy* | 6th grade (11-12 years) |
| 80 | | |
| | *fairly easy* | 7th grade (12-13 years) |
| 70 | | |
| | *plain English* | 8th to 9th grades (13-15 years) |
| 60 | | |
| | *fairly difficult* | 10th to 12th grades (15-18 years) |
| 50 | | |
| 40 | *difficult* | college |
| 30 | | |
| 20 | | |
| | *very difficult* | college graduate |
| 10 | | |
| 0 | | |

**Table 5.14**

*Reading Materials for Trial 1*

| | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|
| *L'auberge* | | 275 | 4,694 | 1,515 | 1,115 | 68.5 | 84.4 |
| *Le secret de maître Cornille* | | 114 | 1,296 | 475 | 353 | 77.6 | 93.1 |
| *La mule du pape* | | 103 | 1,244 | 473 | 380 | 81.5 | 96.4 |
| *Peau d'âne* | | 214 | 2,100 | 615 | 426 | 91.5 | 105 |
| *La chèvre de monsieur Seguin* | | 96 | 1,141 | 409 | 326 | 81.0 | 96.0 |
| *Les souhaits ridicules* | | 72 | 510 | 270 | 204 | 88.2 | 103 |
| *Aalborg* | VK | 5 | 101 | 71 | 63 | 53.1 | 70.6 |
| *Abdomen* | WKP | 26 | 429 | 221 | 189 | 45.9 | 64.8 |
| *Accenteur mouchet* | WKP | 21 | 259 | 170 | 149 | 66.6 | 83.4 |
| *Accident vasculaire cerebral* | VK | 8 | 203 | 118 | 99 | 52.3 | 69.2 |
| *Accordeon* | VK | 13 | 181 | 124 | 108 | 67.9 | 84.3 |
| *Acier* | VK | 11 | 140 | 84 | 72 | 81.5 | 96.3 |
| *Adolescence* | VK | 6 | 105 | 58 | 52 | 56.9 | 74.3 |

**Table 5.14**

*Reading Materials for Trial 1 (Continued)*

| | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|
| *Aeronef* | WKP | 16 | 246 | 133 | 104 | 60.2 | 77.4 |
| *Affluent* | WKP | 7 | 112 | 67 | 59 | 63.7 | 80.4 |
| *Agriculture* | VK | 4 | 105 | 74 | 67 | 54.5 | 71.0 |
| *Aigle royal* | VK | 16 | 129 | 89 | 72 | 82.6 | 97.8 |
| *Aile* | WKP | 6 | 140 | 89 | 75 | 57.5 | 74.0 |
| *Aix-la-Chapelle* | VK | 11 | 177 | 116 | 99 | 46.6 | 65.5 |
| *Albert Einstein* | VK | 6 | 128 | 84 | 78 | 51.7 | 69.2 |
| *Alexandre Millerand* | VK | 19 | 254 | 136 | 114 | 66.0 | 82.7 |
| *Alexandre VI* | VK | 17 | 221 | 132 | 113 | 80.3 | 95.2 |
| *Algue verte* | WKP | 16 | 307 | 167 | 139 | 50.4 | 68.4 |
| *Alimentation* | WKP | 19 | 230 | 142 | 118 | 34.9 | 55.8 |
| *Alligator* | VK | 22 | 240 | 148 | 124 | 61.1 | 78.8 |
| *Amazones* | VK | 17 | 232 | 156 | 129 | 67.9 | 84.3 |
| *Ambulancier* | VK | 13 | 168 | 102 | 84 | 54.7 | 73.0 |

**Table 5.14**

*Reading Materials for Trial 1 (Continued)*

| | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|
| *Amnesty International* | VK | 9 | 105 | 65 | 57 | 52.4 | 71.1 |
| *Amon* | VK | 15 | 233 | 130 | 107 | 69.4 | 85.4 |
| *Anarchie* | VK | 7 | 169 | 115 | 101 | 62.2 | 78.0 |
| *Anarcho-capitalisme* | VK | 10 | 103 | 71 | 60 | 37.0 | 57.9 |
| *Angola* | VK | 20 | 226 | 144 | 121 | 59.1 | 77.0 |
| *Animal de compagnie* | VK | 16 | 231 | 131 | 112 | 61.1 | 78.3 |
| *Anubis* | VK | 10 | 102 | 63 | 52 | 75.4 | 91.3 |
| *Aquarelle* | VK | 10 | 139 | 84 | 65 | 78.9 | 93.9 |
| *Araire* | VK | 15 | 194 | 121 | 100 | 69.0 | 85.4 |
| *Arbousier* | VK | 21 | 222 | 139 | 116 | 72.3 | 88.5 |
| *Arc de triomphe de l'Etoile* | VK | 11 | 132 | 91 | 76 | 65.8 | 82.7 |
| *Archimede* | VK | 12 | 198 | 126 | 106 | 54.2 | 72.0 |
| *Architecture gothique* | VK | 16 | 258 | 151 | 119 | 66.2 | 82.5 |
| *Ardoise* | VK | 14 | 157 | 108 | 89 | 69.9 | 86.4 |

**Table 5.14**

*Reading Materials for Trial 1 (Continued)*

| | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|
| *Armand Jean du Plessis de Richelieu* | VK | 21 | 333 | 185 | 161 | 51.8 | 70.0 |
| *Artemis* | VK | 16 | 249 | 146 | 130 | 62.6 | 79.5 |
| *Artere* | VK | 11 | 142 | 88 | 74 | 79.9 | 94.9 |
| *Arthropode* | VK | 46 | 718 | 328 | 270 | 63.3 | 80.0 |
| *Arthur Rimbaud* | VK | 8 | 118 | 81 | 73 | 78.6 | 93.5 |
| *Athena* | VK | 15 | 211 | 118 | 100 | 73.5 | 89.1 |
| *Atlas mythologie* | VK | 29 | 498 | 250 | 209 | 66.9 | 83.0 |
| *Atrides* | WKP | 52 | 736 | 340 | 280 | 70.3 | 86.3 |
| *Aurore polaire* | VK | 10 | 207 | 121 | 103 | 59.5 | 76.1 |
| *Avalanche* | VK | 20 | 275 | 153 | 117 | 69.5 | 85.7 |

RE = Reading Ease KM = Kandel-Moles VK = Vikidia VK = Wikipedia

**Table 5.15**
*Reading Materials for Trial 2*

| | | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|---|
| A2 | *Oreille* | GUT | 25 | 445 | 205 | 178 | 82.9 | 96.8 |
| | *Yvonne* | GUT | 16 | 282 | 144 | 127 | 82.4 | 96.5 |
| | *Cinéma : plusieurs récents blockbusters vont connaître des suites* | WKN | 12 | 374 | 194 | 172 | 58.3 | 73.6 |
| | *Parc national des Lacs-Waterton* | WKV | 17 | 217 | 105 | 94 | 68.0 | 84.5 |
| | *Sèvres* | WKV | 38 | 424 | 191 | 168 | 80.2 | 95.3 |
| B1 | *La douce bouillie* | GUT | 21 | 487 | 226 | 194 | 73.2 | 87.6 |
| | *L'aïeul et le petit-fils* | GUT | 28 | 273 | 154 | 132 | 91.9 | 105.7 |
| | *Nicolas Sarkozy est élu président de la République française* | WKN | 29 | 500 | 211 | 186 | 57.9 | 75.1 |
| | *Nazare* | WKV | 13 | 252 | 148 | 136 | 70.3 | 85.7 |
| | *Santiago de Cuba* | WKV | 40 | 428 | 223 | 200 | 74.0 | 90.0 |
| B2 | *Le renard et le chat* | GUT | 46 | 472 | 231 | 196 | 83.0 | 97.9 |

**Table 5.15**

*Reading Materials for Trial 2 (Continued)*

| | | | Sentences | Tokens | Types | Lemmas | RE | RE KM |
|---|---|---|---|---|---|---|---|---|
| | *L'hirondelle et les petits oiseaux* | GUT | 60 | 484 | 258 | 216 | 90.5 | 104.7 |
| | *Finlande : les Finlandais appelés aux urnes pour élire leurs députés* | WKN | 13 | 258 | 155 | 139 | 57.5 | 74.5 |
| | *Asie du Sud* | WKV | 21 | 400 | 194 | 172 | 69.7 | 85.2 |
| | *Nebraska* | WKV | 21 | 351 | 176 | 165 | 61.4 | 78.3 |
| C1 | *Le loup et l'agneau* | GUT | 31 | 268 | 162 | 136 | 93.3 | 107.0 |
| | *Simonide préservé par les Dieux* | GUT | 77 | 588 | 312 | 259 | 93.6 | 107.5 |
| | *Suisse : la presse écrite connait un net recul en Romandie* | WKN | 15 | 387 | 182 | 162 | 65.2 | 80.4 |
| | *Tarawa* | WKV | 20 | 399 | 184 | 174 | 58.5 | 75.3 |
| | *Trujillo (Pérou)* | WKV | 9 | 276 | 152 | 139 | 48.8 | 65.5 |

GUT = Project Gutenberg WKN = Wikinews WKV = Wikivoyage

**Table 5.16**

*List of Words Found Difficult by All Learners Participating in Trial 1*

| Level | Text | Context |
|---|---|---|
| All | *Amazones* | Les Amazones pour avoir des enfants s'**unissaient** à des étrangers. |
| | *Amon* | Dans certaines versions égyptiennes de la création du monde, Amon sous la forme d'une **oie** pond l'œuf des origines. |
| | *L'auberge* | Il attendait, l'oreille éveillée aux bruits lointains, frissonnant quand le vent léger **frôlait** le toit et les murs. |
| | | Pendant cinq heures, il monta, escaladant des rochers au moyen de ses crampons, taillant la glace, avançant toujours et parfois **hâlant**, au bout de sa corde, le chien resté au bas d'un escarpement trop rapide. |
| | | Comme il se trouvait trop loin de sa maison pour y rentrer, et trop fatigué pour se traîner plus longtemps, il creusa un trou dans la neige et s'y **blottit** avec son chien, sous une couverture qu'il avait apportée. |
| | | Mais soudain, une voix, un cri, un nom : « Ulrich », secoua son **engourdissement** profond et le fit se dresser. |
| | | D'un bond il rentra dans l'auberge, ferma la porte et poussa les verrous ; puis il tomba **grelottant** sur une chaise, certain qu'il venait d'être appelé par son camarade au moment où il rendait l'esprit. |

**Table 5.17**
*List of Words Found Difficult by All Learners Participating in Trial 2*

| Level | Text | Context |
|---|---|---|
| A2 | *Oreille* | Tiens, voilà ton bien, **décampe**! |
| | *Parc national des Lacs-Waterton* | Le parc national des Lacs-Waterton est un site du **patrimoine** mondial de l'UNESCO situé en Alberta au Canada. |
| | | À partir de Calgary, il faut **emprunter** l'autoroute 2 vers le sud jusqu'à Cardston et ensuite suivre l'autoroute 5 vers l'ouest jusqu'au parc. |
| | *Yvonne* | Elle voulut se retenir à la table; elle se **cramponna** à la nappe, et patatras! |
| | | Médor et Minouche se **lamentèrent** d'abord, puis ils se consolèrent en mangeant sous la table la crème et les gâteaux. |
| B1 | *L'aïeul et le petit-fils* | Ils rappelèrent entre eux l'**aïeul** qui ne quitta plus la table de famille. |
| B2 | *L'hirondelle et les petits oiseaux* | –Prophète de malheur, **babillarde**, dit-on, Le bel emploi que tu nous donnes! |
| C1 | *Simonide préservé par les Dieux* | La louange **chatouille** et gagne les esprits. |

**Figure 5.4**

*Monte Carlo Power Simulations of GLMM 5.1 on Trial 1 as the Number of Participants Increases*

```
library(lme4)
library(simr)

fit <- glmer(difficulty ~ cloze_10 + (cloze_10|pro_level:sbj_id),
data=trial1, family=binomial(link='logit'))

powerSim(fit, nsim=50, seed=0)
## Power for predictor 'cloze_10', (95% confidence interval):
##       100.0% (86.28, 100.0)
##
## Test: Likelihood ratio
##
## Based on 25 simulations, (0 warnings, 0 errors)
## alpha = 0.05, nrow = 188972
##
## Time elapsed: 0 h 57 m 13 s
##
## nb: result might be an observed power calculation

plot(powerCurve(fit, along='sbj_id', breaks=c(1, seq(2, n_sbj, by
    =2)), nsim=25, seed=0))
```

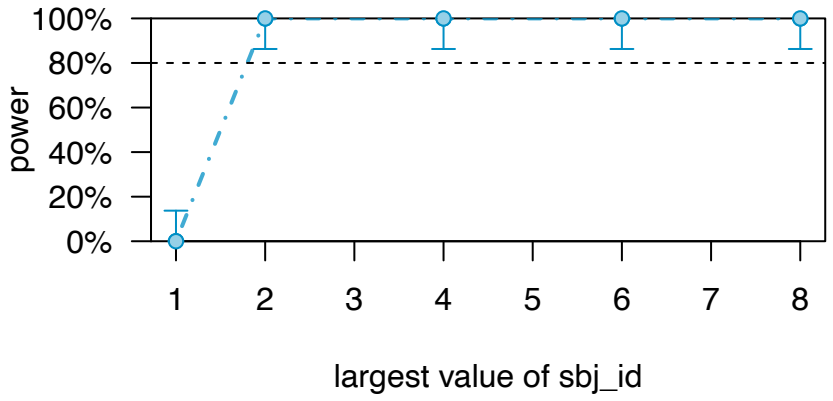


**Figure 5.5**

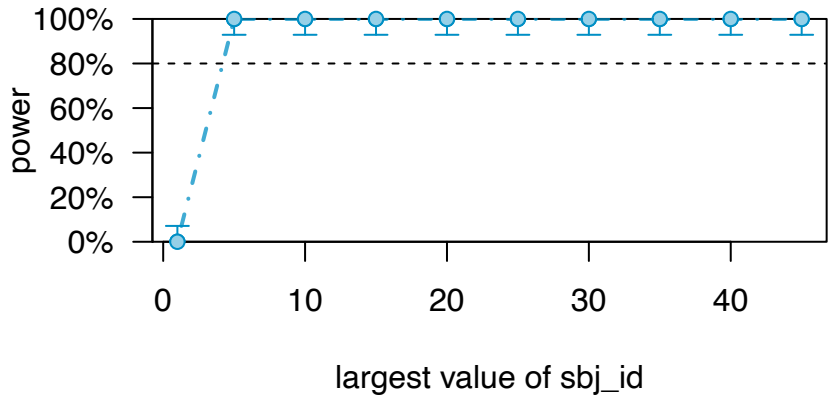
*Monte Carlo Power Simulations of GLMM 5.2 on Trial 2 as the Number of Participants Increases*

```
library(lme4)
library(simr)

fit <- glmer(difficulty ~ cloze_20 + (cloze_20|pro_level:sbj_id),
data=trial2, family=binomial(link='logit'))

powerSim(fit, nsim=25, seed=0)
## Power for predictor 'cloze_20', (95% confidence interval):
##       100.0% (92.89, 100.0)
##
## Test: Likelihood ratio
##
## Based on 50 simulations, (0 warnings, 0 errors)
## alpha = 0.05, nrow = 72610
##
## Time elapsed: 0 h 38 m 55 s
##
## nb: result might be an observed power calculation

plot(powerCurve(fit, along='sbj_id', breaks=c(1, seq(5, n_sbj, by
    =5)), nsim=50, seed=0))
```



PART III

# PREDICTING LEXICAL DIFFICULTY

This is the difference which Plato draws between ευκολος and δυσκολος – the man of *easy*, and the man of *difficult* disposition – in proof of which he refers to the varying degrees of susceptibility which different people show to pleasurable and painful impressions; so that one man will laugh at what makes another despair.

— Arthur Schopenhauer, *Aphorismen zur Lebensweisheit* (1851)

CHAPTER 6

# *Explanatory Factors Of Difficulty*
# A GENERALIZED LINEAR MIXED MODEL WITH FIXED LEXICAL EFFECTS AND RANDOM LEARNER EFFECTS

**Abstract** This chapter presents a mixed-effects modeling approach to the prediction of perceived difficulty. The learner data presented in the preceding chapter (Chapter 5) is used as training data for building a logistic generalized linear mixed model (GLMM). First, various word-level explanatory factors are selected for inclusion in the model, including aspects related to word form, meaning, and use; word exposure; and word etymology. Second, several random learner effects are introduced in the model to make learner-specific predictions of difficulty.

Mixed-effects models (Bates et al., 2015; Pinheiro & Bates, 2000) are statistical models that extend (generalized) linear models by combining fixed and random factors. The models' fixed effects correspond to the independent variables usually included in any linear model, whereas the models' random effects characterize a variance in the linear model's intercept and slope(s). Because such random effects represent a variability between groups, mixed models often have a hierarchical structure with observations nested into one or more higher-order levels. As such, mixed-effects models seem particularly suited for predicting personal lexical difficulty in reading. In this case, observations represent words (level 1) read by a particular learner (level 2), of which the perceived difficulty can be predicted from word-level factors and individual learners differences.

Figure 6.1 illustrates a two-level generalized linear mixed model (GLMM) predicting lexical difficulty perceived by a non-native reader. The lowest level represents a particular word $i$ characterized by linguistic traits (e.g., word surprisal) and other indicators of word processing (e.g., number of previous exposures to the word in the task). This level is nested in a higher-order class representing a particular learner $j$ who is reading word $i$ and has some personal characteristics (e.g., proficiency level). This second-order level comprises some learner-specific coefficients that account for two sources of variance. On the one hand, the random intercept $u_{0j}$ accounts for the variability between learners in the overall extent of difficulty. On the other hand, the random slopes $u_{1j}$ and $u_{2j}$ account for the variability in specific lexical features for different learners. The development and analysis of this mixed model are described in this chapter.

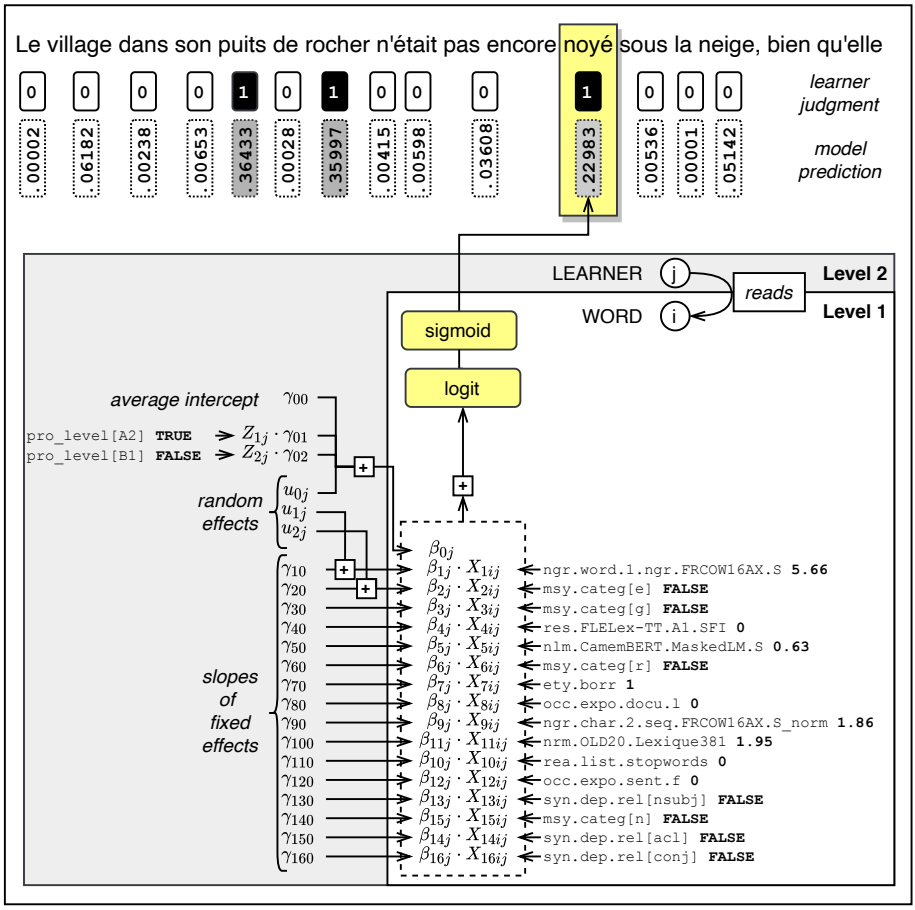
There were two reasons why this doctoral study adopted a mixed-effects modeling approach. The first reason was to identify the most significant factors that explained the data introduced in Chapter 5. Although the previous chapter highlighted contextual constraints on perceived difficulty, the analyses did not compare this particular effect to other explanatory factors. Therefore, a stepwise forward-backward regression was performed to establish the primary factors why learners perceived lexical difficulty in reading. The second reason was to integrate the non-negligible between-learner variance observed in Chapter 5 into the predictions. Because the inter-subject agreement coefficient was below the minimal threshold ($\alpha < .67$), the previous chapter concluded that it would be unreliable to aggregate this learner data into an aggregated measure. Furthermore, because the intra-subject correlation coefficient was above the minimal threshold ($ICC > .05$), the previous chapter noted the necessity of differentiating between learners. Therefore, the study addressed two research questions:

RQ6.1 What are the most significant factors that predict why French L2 learners are triggered to perceive lexical difficulty when reading?

RQ6.2 What is the extent of between-learner variance in explaining and predicting lexical difficulty perceived by non-native readers? Does a model predicting personal difficulty achieve higher performance than a model predicting difficulty for the average learner?

**Figure 6.1**

*Generalized Linear Mixed-Effects Model of Perceived Lexical Difficulty*




The chapter is structured as follows. Section 6.1 describes the set of features considered for inclusion in the model. The features set comprises characteristics of the word's form (Section 6.1.1), meaning (Section 6.1.2), and use (Section 6.1.3); the learner's exposure to the word in the task (Section 6.1.4); and the word's etymology (Section 6.1.5). Next, Section 6.2 describes the development and analysis of the GLMM model. First, the section gives a mathematical formulation of GLM and GLMM models (Section 6.2.1). Second, the section presents the procedure used to select the most significant explanatory factors (Section 6.2.2). Third, the section provides an analysis of the final

**Table 6.1**

*What is Involved in Knowing a Word Receptively? (Nation, 2001, p. 27)*

| | |
|---|---|
| | FORM |
| spoken | What does the word sound like? |
| written | What does the word look like? |
| word parts | What parts are recognizable in this word? |
| | MEANING |
| form and meaning | What meaning does this word form signal? |
| concept and referents | What is included in the concept? |
| associations | What other words does this make us think of? |
| | USE |
| grammatical functions | In what patterns does the word occur? |
| collocations | What words or types of words occur with this one? |
| constraints on use | Where, when, and how often would we expect to meet this word? |

model (Section 6.2.3). Finally, Section 6.3 concludes the results regarding the dissertation's two research aims.

## 6.1 LEXICAL FEATURES

Ideally, the study should have considered all possible predictors previously examined in the literature (see Chapters 2 and 3). However, the study could not examine all these predictors for several practical reasons. The most apparent reason was that certain variables depended on unavailable empirical measures such as eye-movement characteristics. Moreover, other factors required using different experimental conditions and depended on several unavailable technologies (e.g., computerized dynamic assessment). Instead, the set of features was limited to predictors of word form, meaning, use, and etymology and of how the word occurs in the reading task. The distinction between word form, meaning, and use was inspired by Nation's (2001) categorization of receptive vocabulary knowledge (see Table 6.1).

### 6.1.1 *Form*

Comprehending a word entails knowing both the spoken and written form of the word. However, it may be that this comprehension is inhibited by the inherent complexity of the word form. A complex word form is considerably long, composed of various morphemes, has various unlikely character combinations, and lacks proximity with other word forms. This study examined the following features: word length, character n-grams, and orthographic norms.

#### *Word Length*

As previously mentioned in Chapter 3, word length is a traditional and frequently used variable in complexity and readability research. The study focused on the following length-based features: word length in the number of characters, letters (i.e., ignoring digits and other non-alphabetic characters), graphemic vowels ($v \in \{a, e, i, o, u\}$), syllables, and phonemes; the ratio of phonemes to graphemes; and the length of the word's stem (as computed with the Snowball algorithm; Porter, 2001). The expected effect of these features was as follows: the longer the word, the more likely the difficulty. A notation is given in Definition 6.1.

**Definition 6.1: word length features**

| | |
|---|---|
| `len.char` | number of characters |
| `len.let` | number of letters |
| `len.vowg` | number of (graphemic) vowels |
| `len.phon.espeak` | number of phonemes (with eSpeak) |
| `len.ph:gr.espeak` | number of phonemes to number of graphemes |
| `len.stem.snowball` | number of characters in stem (with Snowball) |

| | |
|---|---|
| `len.syll.hunspell` | number of syllables (with Hunspell) |
| `rea.syll.gunning.hunspell` | Gunning's polysyllable threshold |

*Character N-Grams*

The second indicator of form complexity was the likelihood (or, conversely, the surprisal) of observing a series of character *n*-grams in the word. The word's character *n*-gram likelihood was computed by moving a sliding window over *n* consecutive characters in the word and multiplying the probability (or summing the log-probability) of each sequence of *n* characters. The degree of surprisal was computed by taking the negative of this sequential log-likelihood. The sequential log-likelihood or surprisal was normalized to remove the influence of word length, dividing the value by the number of sliding windows. For example, the word *Archiépiscopal* had a high sequential surprisal value, both in the case of character bigrams (*{Ar,rc,ch,hi,ié,ép,pi,is,sc,co,op,pa,al}*, $S$ = 20.2, $S_{\text{norm}}$ = 1.44) and character pentagrams (*{Archi,rchié,chiép,hiépi,iépis,épisc,pisco,iscop,scopa,copal}*, $S$ = 14.98, $S_{\text{norm}}$ = 1.07). Therefore, the expected effect of this variable was as follows: the less likely or more surprising the sequence of character n-grams, the more likely the difficulty.

In some papers on complex word identification (CWI) (cf. Chapter 3), character bigram and trigram frequencies and probabilities were integrated as features into a classifier of word complexity (Alfter & Pilán, 2018; Bingel et al., 2016; Choubey & Pateria, 2016; Gooding & Kochmar, 2018; Konkol, 2016; Popović, 2018; Zampieri et al., 2016). In these papers, however, it was not explicit nor thoroughly examined to what extent character bigrams and trigrams contributed to predicting perceived lexical difficulty. This study aimed to shed light on their potential effect, or lack thereof. In this study, character n-gram language models were estimated with modified Kneser-Ney smoothing (KenLM; Heafield et al., 2013) on the FRCOW16AX gigaword web corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012). Three features were computed:

- the log-likelihood of the sequence of character n-grams

- the surprisal of the sequence of character n-grams
- the average surprisal of character n-grams

A notation of these features is given in Definition 6.2.

**Definition 6.2: character n-gram features**

| | | |
|---|---|---|
| `ngr.char.`$n$`.seq.`$uri$`.`$f$ | | character $n$-gram sequence probability with modified Kneser-Ney smoothing |
| | where | $n \in \{1, 2, 3, 4, 5\}$ |
| | | *uri* is the identifier of the training corpus |
| | | $f \in \{\ell, S, S_{\text{norm}}\}$ |

*OLD20 Norm*

Yarkoni et al. (2008) first proposed the orthographic Levenshtein distance 20 (OLD20) norm as a predictor of latencies in visual word recognition. They defined words with a high OLD20 value as being "orthographically sparse" and those with a low OLD20 value as being "orthographically dense" (p. 972). The norm had a significant positive correlation with reaction time measures: the higher the orthographic sparsity, the higher the latencies in speeded pronunciation ($r = .59$) and lexical decision ($r = .61$). For this reason, the OLD20 norm was another crucial indicator of form complexity.

The OLD20 norm was computed as follows. First, each target word form was compared to a finite and representative list of orthographic word forms. This study focused on the list of 142,694 word forms attested in Lexique3 (New et al., 2007). Next, the words in this finite vocabulary were ranked in function of their Levenshtein distance with the target word form. Only the top 20 nearest neighbors were kept and the individual Levenshtein distances were averaged. A notation of this feature is given in Definition 6.3.

**Definition 6.3: OLD20 feature**

| | |
|---|---|
| `nrm.OLD20.Lexique381` | average Levenshtein distance with the 20 closest orthographic forms in Lexique3.81 |

It should be noted that the feature was initially computed with a publicly available Python script that aimed a faster computation of the OLD20 norm.[68] The computed values were verified for each form in Lexique3. However, they were noticeably different from the pre-computed values provided in Lexique3. For this reason, a standard sorting algorithm (viz., Timsort) was used to find the nearest orthographic neighbors. Although it took a considerably longer time to compute, this algorithm produced the same OLD20 values as those listed in Lexique3.

*Other Variables Related to Form*

Besides word length, character n-grams, and the OLD20 norm, other pre-computed features were extracted from Lexique3 (New et al., 2007). These are listed in Definition 6.4.

**Definition 6.4: Lexique3 features**

| | |
|---|---|
| `res.Lexique381.nbhomogr` | number of homographs |
| `res.Lexique381.nbhomoph` | number of homophones |
| `res.Lexique381.nbmorph` | number of morphemes |
| `res.Lexique381.nombre_s` | is a singular form |
| `res.Lexique381.nombre_p` | is a plural form |
| `res.Lexique381.genre_f` | is a feminine form |
| `res.Lexique381.genre_m` | is a masculine form |
| `res.Lexique381.voisorth` | number of orthographic neighbors |

68 The code was retrieved from the following website: http://crr.ugent.be/programs-data/fast-computation-of-average-levenshtein-distances-in-python-including-old20.

| | |
|---|---|
| `res.Lexique381.voisphon` | number of phonological neighbors |
| `res.Lexique381.puorth` | orthographic unicity point |
| `res.Lexique381.puphon` | phonological unicity point |
| `res.Lexique381.old20` | average Levenshtein distance with the 20 closest orthographic forms |
| `res.Lexique381.pld20` | average Levenshtein distance with the 20 closest phonological forms |

### 6.1.2 *Meaning*

Comprehending a word entails knowing the word's meaning and its associations with other words. This knowledge is a network of lexical units interconnected via various semantic and other associative links. Learning vocabulary then entails learning this target language lexical network. This study examined several features of the reference lexico-semantic network WordNet (Fellbaum, 1998), specifically, its French equivalent (Sagot & Fišer, 2008).

The Open Multilingual WordNet (OMW) was searched to retrieve entries corresponding to the target word's lemma and part of speech. Each part of speech was converted to the category used in WordNet: *a* for adjectives, *r* for adverbs, *n* for nouns, and *v* for verbs. The degree of polysemy was computed by counting the number of senses or concepts linked to the target word. The degree of synonymy was computed by counting the number of lemmas linked to each sense of the target word. Furthermore, the average number of interlingual lexicalizations was computed by counting the number of lemmas in other languages that were linked to the target word's meaning(s). The total number of semantic relations and the hypernymy relations described in Chapter 4 were computed as well. A notation of these features is given in Definition 6.5.

**Definition 6.5: WordNet features**

| | |
|---|---|
| `own.lexi` | (average) number of interlingual lexicalizations |
| `own.hype` | (average) hypernymy rank |
| `own.hypo` | (average) hyponymy rank |
| `own.hypr` | (average) relative hypernymy rank |
| `own.rela` | number of semantic relations |
| `own.sens` | number of senses |
| `own.syno` | number of synonyms |

### 6.1.3 *Use*

Comprehending a word entails having knowledge of the word's function and use in the target language. This study examined three fundamental functional aspects: the word's frequency, the context surrounding the word, and its grammatical function.

*Frequency, Prevalence, and Commonness*

Because frequency effects are well-established in L2 research (see Ellis, 2002), this study examined three indicators of word frequency. A first indicator was obtained from general corpus frequencies, which represent the frequency of the word in the target language. These general corpus frequencies were extracted from the Lexique3 database, which provides the word's frequency of occurrence in films or books (New et al., 2007). A second indicator was obtained from specialized, graded word frequencies. These measures were described in more detail in a previous chapter (Chapter 4). Two types of graded word frequencies were considered: (a) the frequency of the word in graded textbooks for French elementary school children (Lété, 2004) and (b) the frequency of the word in graded textbooks for learners of French (Francois et al., 2014). A notation of these three indicators of frequency is given in Definition 6.6.

**Definition 6.6: word frequency features**

| | | |
|---|---|---|
| `res.Lexique381.`$f$ | | frequency in films or books |
| | where | $f$ is *freqfilms2*, *freqlemfilms2*, *freqlivres*, or *freqlemlivres* |
| `res.Manulex.`$l$`.`$f$ | | frequency in elementary school textbooks |
| | where | $l \in$ {G1, G2, G3-5, G1-5}<br>$f \in$ {*D*, *U*, *SFI*} |
| `res.FLELex-TT.`$l$`.`$f$ | | frequency in L2 textbooks and readers |
| | where | $l \in$ {A1, A2, B1, B2, C1, C2, Total}<br>$f \in$ {*U*, *SFI*} |

The notion of word prevalence refers to the percentage of people in the target population who know the word. Word prevalence not only predicts reaction time measures in lexical decision tasks, but also complements the effect of word frequency (Brysbaert et al., 2018; Keuleers et al., 2015). For French, the previously mentioned Lexique3 database provides a similar crowdsourced word prevalence norm, which is defined in Definition 6.7.

**Definition 6.7: Lexique3 prevalence feature**

| | |
|---|---|
| `res.Lexique381.deflem` | prevalence, or the percentage of people who know the word (lemma) |

The last indicator related to both frequency and prevalence is the notion of word commonness, which can be obtained by looking up the presence of a word in a basic word list. As was previously mentioned in Chapter 3,

the presence or absence from such a basic vocabulary list is a fundamental indicator of lexical complexity in readability research. The commonness of a word was measured by looking up the presence of the word in two lists: (a) a list of stop words[69] and (b) the Gougenheim et al. (1964) 2.00 list, including the sublists for the basic word categories *maison*, *cuisine*, *vetements*, *animaux*, *parties du corps*, *meubles*, and *moyens de transport*. A notation of this feature is given below in Definition 6.8.

**Definition 6.8: basic word list feature**

| | |
|---|---|
| `rea.list.`*uri* | presence in a vocabulary list |

*Contextual Likelihood and Similarity*

Although the previously mentioned indicators of frequency, prevalence, and commonness indicate word usage, they disregard its use in its surrounding context. As previously highlighted in Chapters 2, 3 and 5, the surrounding context is an important explanatory variable. This study, therefore, examined various indicators of word usage in context on top of the traditional frequency effects. This section describes two such indicators: the word's likelihood of occurrence in the given context and its semantic similarity with the surrounding words. In the following section, two other indicators will be presented: the word's morphosyntactic function and its syntactic dependency relation.

The word's likelihood of occurrence was computed with a probabilistic language model. The simplest language model is the $n$-gram language model, which gives the probability of the word occurring after the $n-1$ preceding words. Like the character $n$-gram language models mentioned previously, these word $n$-gram models were estimated with modified Kneser-Ney smoothing (KenLM; Heafield et al., 2013) on the FRCOW16AX gigaword web corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012). Three types of probabilities were computed:

- the $n$-gram probability (i.e., the probability of the target word occurring after the $n-1$ preceding words),

69 https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt

- the sequence probability (i.e., the joint $n$-gram probabilities of all words in the sentence leading up to the target word), and
- the sentence probability (i.e., the joint $n$-gram probabilities of all words in the sentence).

A notation of these features is given in Definition 6.9.

**Definition 6.9: n-gram likelihood features**

| | | |
|---|---|---|
| `ngr.word.`$n$`.ngr.`$uri$`.`$f$ | | word $n$-gram probability $P(w_t\|h)$ of the target word $w_t$ given the preceding words $h$ |
| `ngr.word.`$n$`.seq.`$uri$`.`$f$ | | word $n$-gram sequence probability $\prod_{i=1}^{t} P(w_i\|h)$ from the first word in the sentence to $w_t$ |
| `ngr.word.`$n$`.sen.`$uri$`.`$f$ | | word $n$-gram sentence probability $\prod_{i=1}^{N} P(w_i\|h)$ from the first word to the length $N$ of the sentence |
| | where | $n \in \{1, 2, 3, 4, 5\}$ |
| | | $uri$ is the identifier of the training corpus |
| | | $f \in \{\ell, S, S_{\text{norm}}\}$ |

The word's contextual likelihood was also computed with the pre-trained CamemBERT model (Martin et al., 2019), a state-of-the-art neural language model used in the previous chapter (Chapter 5). Whereas an $n$-gram model only takes into account the $n-1$ preceding words in the sentence, the CamemBERT model computes the word's contextual likelihood by taking into account the entire surrounding (viz., MaskedLM) or preceding (viz., CausalLM) context. As such, its computations are more deeply contextualized than those of the $n$-gram model. The study examined the same features described in Chapter 5, namely the degree of word surprisal and the cloze predictions. A notation of these features is given in Definition 6.10.

**Definition 6.10: BERT language model features**

| | | |
|---|---|---|
| `nlm.`*uri*`.`*m*`.`*f* | | neural language model probability |
| | where | *uri* is CamemBERT |
| | | $m \in$ {MaskedLM, CausalLM} |
| | | $f \in$ {*S*, cloze1, cloze5, cloze10, cloze20} |

Lastly, the word's semantic similarity with the surrounding words was computed with FastText word embeddings (Bojanowski et al., 2017; Grave et al., 2018). Each target word vector was compared with either the preceding word vector or the aggregate vector of all preceding words in the sentence. The following metrics were computed: the cosine distance or similarity and the angular distance or similarity. Additionally, the ratio of the angular similarity of the previous word over the angular similarity for the current word was also computed to capture an incremental change in semantic (dis)similarity from the previous word to the current word. A notation of these features is given in Definition 6.11.

**Definition 6.11: word embedding features**

| | | |
|---|---|---|
| `emb.cc.sim.prev.`*f* | | (dis)similarity between the vector of the target word $\vec{w}_t$ and the vector of the previous word $\vec{w}_{t-1}$ |
| `emb.cc.sim.seq.`*f* | | (dis)similarity between the vector of the target word $\vec{w}_t$ and the aggregate vector of the previous words in the sentence $\sum_{i=1}^{t-1} \vec{w}_i$ |
| | where | `cc` means the vector was trained on CommonCrawl |

| | |
|---|---|
| | $f$ is the cosine distance/similarity, the angular distance/similarity $\alpha$, the ratio $\alpha_{t-1}/\alpha_t$ |

*Morphosyntax and Syntactic Dependencies*

The function of the word in the sentence was determined with two pieces of (morpho)syntactic information: the word's part of speech and its syntactic dependency relation. The word's part of speech was obtained from the output of the part-of-speech taggers mentioned in Chapter 5. The morphosyntactic classes of TreeTagger (Schmid, 1994) and LGTagger (Constant & Sigogne, 2011) were then transformed to a common set of morphosyntactic categories, including adjectives, adverbs, nouns, verbs, grammatical words, named entities, and foreign words. A further indicator of complexity was the extent to which the target word was functionally ambiguous, which was computed from the total number of different parts of speech attributed to the word in Lexique3. A notation of these features is given in Definition 6.12.

**Definition 6.12: morphosyntactic features**

| | |
|---|---|
| `msy.class` | morphosyntactic class, which is the output of the part-of-speech tagger |
| `msy.categ` | morphosyntactic category where *a* is adjective, *e* is named entity, *g* is grammatical word, *n* is noun, *v* is verb, *r* is adverb, and *f* is foreign word |
| `msy.nbcla.Lexique381` | number of different parts of speech for the target word, as found in Lexique3 |

The word's syntactic dependency relation was computed with spaCy's neural dependency parser (Honnibal & Montani, 2017). The dependency parser outputs a dependency tree for each sentence in the data set and labels each dependency with the universal dependencies (UD) relations (De Marneffe et al., 2014). Besides the word's UD relation, the following indicators of syntactic complexity were computed: the length of the sentence, the word's depth in the dependency tree, and the positional distance in the sentence from either the dependency head or the dependency root. A notation of these features is given in Definition 6.13.

**Definition 6.13: syntactic features**

| | |
|---|---|
| `len.sen` | length of sentence in number of words |
| `syn.dep.rel` | syntactic dependency relation |
| `syn.dep.rdis` | distance from root (position $i$ in the sentence, $i_{\text{word}} - i_{\text{root}}$) |
| `syn.dep.hdis` | distance from head (position $i$ in the sentence, $i_{\text{word}} - i_{\text{head}}$) |
| `syn.dep.dept` | depth in the dependency tree |

### 6.1.4 *Exposure*

The common ground of the previously described features is that their value is the same for all learners as they do not depend on the reading order. At the same time, there are two reasons why the order in which the texts are read may be interesting to include in the analyses. The first reason is that reading order determines the number of repeated encounters with a given lexical unit. Word repetition is an essential predictor of incidental vocabulary acquisition through reading (cf. Definition 6.14). The second reason is that difficulty measures may be influenced by the sequential position of the word in the text (cf. Definition 6.14). The following features were examined: the number of previous exposures to a lexical unit, the spacing between these

exposures, and the sequential order in the text. A notation of these feature is given in Definition 6.14.

**Definition 6.14: word occurrence in reading task**

| | | |
|---|---|---|
| `occ.expo.`*span*`.`*unit* | | number of previous exposures to the target *unit* in *span* |
| `occ.ordr.`*span* | | order of occurrence (i.e., sequential position) of the word in *span* |
| `occ.spac.`*span*`.`*unit* | | spacing between the previous occurrence of the target *unit* in *span* |
| | where | *span* is: `task` (entire reading task), `docu` (text document), or `sent` (sentence) |
| | | *unit* is the tallying unit: `f` (form), `l` (lemma), `p` (part of speech), `fl` (form + lemma), `lp` (lemma + part of speech), `flp` (form + lemma + part of speech) |

### 6.1.5 *Etymology*

The link between perceived difficulty and the word's etymology has not yet been thoroughly investigated. Bingel et al. (2016), for instance, examined whether the etymological origin (i.e., Latin root) of the word could predict whether the word would be difficult or not. More recently, Palmero Aprosio et al. (2020) proposed a method for identifying cognates and false friends in CWI, but they did not validate this feature on any user or learner data. To further investigate the potential effect of word etymology on perceived difficulty, several features were extracted from Etymological WordNet (de Melo, 2014). A notation of these features is given in Definition 6.15.

**Definition 6.15: Etymological WordNet features**

| | |
|---|---|
| `ety.ance` | number of ancestors, the length of the path from the word to the root of the ancestry tree |
| `ety.desc` | number of descendants, the length of the path from the word to the root of the descent tree |
| `ety.borr` | number of languages having borrowed the word |
| `ety.cogn` | path distance in the etymological tree between the word and the closest cognate in the L1 |
| `ety.lat` | has a Latin root |

## 6.2 GENERALIZED LINEAR MIXED-EFFECTS MODELS

Based on the features set introduced in the previous section, a GLMM was developed with the *lme4* library in R (Bates et al., 2015). The model was constructed in three steps, starting with a mathematical definition of the most comprehensive model (Section 6.2.1). Next, a stepwise forward-backward selection procedure was used to select the best model configuration (Section 6.2.2). Finally, the most explanatory factors and their effect sizes were identified and analyzed (Section 6.2.3).

### 6.2.1 *Definition*

The definition of the GLMM model started with a simpler statistical model: a binomial logistic regression model. This model was equivalent to a generalized linear model (GLM) with a logit link function and a binomial random variable $Y$ as the dependent variable. Here, $Y$ was the number of difficult words observed in the two trials described in Chapter 5. The data followed a binomial

distribution defined either at the general level (6.1a) or at the level of the individual learner $j$ (6.1b):

$$Y \sim B(N, p) \tag{6.1a}$$

$$Y_j \sim B(n_j, p_j) \qquad \forall j \in \{1, ..., M\} \tag{6.1b}$$

where

$Y$ is the number of words perceived as difficult

$N$ is the number of words read

$p$ is the probability that a word is perceived as difficult

$M$ is the number of learners

$Y_j$ is the number of words perceived as difficult by learner $j$

$n_j$ is the number of words read by learner $j$ such that $\sum_{j=1}^{M} n_j = N$

$p_j$ is the probability that a word is perceived as difficult by learner $j$

Notably, the data from both trials were combined such that the regression model was estimated on a more considerable amount of observations. As a result, the data comprised 261,942 lexical units read by a sample of 56 learners and followed $Y \sim B(261942, 0.049)$.

A drawback of a simple logistic GLM was that it could only model the response based on a single varying factor. The intercept for learner $j$ ($\beta_{0j}$) and the slope of factor $p$ for learner $j$ ($\beta_{pj}$) were the same as, respectively, the average intercept for all learners (6.2b) and the average slope of factor $p$ for all learners (6.2c). As a result, the model could only make predictions for the group of learners on average; it could not account for the between-learner variance observed in Chapter 5. The model defined in Equation (6.2) was extended so that the intercept $\gamma_{00}$ and slope $\gamma_{p0}$ could vary between learners. The most comprehensive mixed-effects model included a random intercept and a random slope for each fixed effect. Additionally, the random intercept and random slopes also included an interaction with the learner's characteristics (i.e., native language and proficiency level). This 'full' GLMM model is defined in Definition 6.17.

**Definition 6.16: generalized linear model**

Let GLM be a multiple logistic regression model given by a set of $N$ observations drawn from a sample of $M$ learners. The model's logistic function is a linear combination of a set of $p$ factors

$$\ln\left[\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right] = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{pj}X_{pij} \tag{6.2a}$$

$$\beta_{0j} = \gamma_{00} \tag{6.2b}$$

$$\beta_{pj} = \gamma_{p0} \tag{6.2c}$$

such that the model predicts whether learner $j \in \{1, \ldots, M\}$ will perceive word $i \in \{1, \ldots, n_j\}$ as difficult or not. Within this model,

$P$ is the probability that word $i$ is perceived as difficult $P(y_{ij}=1)$ or not $P(y_{ij}=0)$ by learner $j$

$i$ is the word as it occurs in the text read by the learner

$j$ is the learner

$X_p$ is the $p$-th lexical feature

$\beta_{0j}$ is the intercept for learner $j$

$\gamma_{00}$ is the average intercept for all learners

$\beta_{pj}$ is the slope of factor $p$ for learner $j$

$\gamma_{p0}$ is the average slope of factor $p$ for all learners

**Definition 6.17: generalized linear mixed model**

Let GLMM be an extension of GLM given by a set of $N$ observations drawn from a sample of $M$ learners. The model's logistic function is a linear combination of a set of $p$ factors of lexical complexity, a set of $q$ learner characteristics, and a learner-specific part $u$

$$\ln\left[\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right] = \beta_{0j} + \beta_{1j}X_{1ij} + \ldots + \beta_{pj}X_{pij} \tag{6.3a}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \ldots + \gamma_{0q}Z_{qj} + u_{0j} \tag{6.3b}$$

$$\beta_{pj} = \gamma_{p0} + \gamma_{p1}Z_{1j} + \ldots + \gamma_{pq}Z_{qj} + u_{pj} \tag{6.3c}$$

such that the intercept $\beta_{0j}$ and the slope $\beta_{0p}$ vary between learners. Within this model,

$Z_{qj}$ is the $q$-th characteristic of learner $j$

$u_{0j}$ is the random part of the intercept attributed to learner $j$

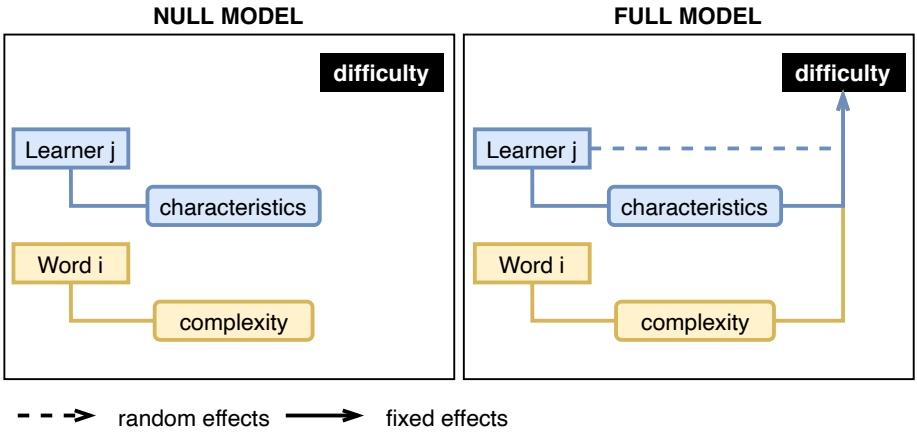
$\sigma^2_{u_0}$ is the between-intercepts variance

$u_{pj}$ is the random part of the slope of factor $p$ attributed to learner $j$

$\sigma^2_{u_p}$ is the between-slopes variance for factor $p$

$\rho_{0p}$ is the correlation between the random intercept and the random slope for factor $p$

**Figure 6.2**
*Null and Full Generalized Linear Mixed-Effects Models*




### 6.2.2 *Selection*

If all possible fixed effects, random effects, and interactions were kept, the model defined in Definition 6.17 would need a considerable amount of data to converge. It was key to select only those fixed and random effects which contributed significantly to the prediction of difficulty.

A stepwise regression procedure was used to search for the optimal configuration of fixed and random effects. Starting from a null model (i.e., which did not include any parameters besides the intercept), various parameters were iteratively added to the null model until (possibly) the full model defined in Definition 6.17 was attained (Figure 6.2). The search ended when no parameter could be included or removed so as to improve the model's fit with the data. However, before starting the stepwise procedure, three issues had to be addressed first: (a) standardizing the model's parameters, (b) detecting and removing multicollinearity from the model, and (c) determining the model's goodness of fit with the data.

STANDARDIZATION The first issue was the range of possible values of each variable in the regression model. Because the regression model required all factors to be numeric, all categorical variables were represented as a one-

hot encoding. For numeric variables, the range of possible values varied considerably. Because the regression model required all factors to have a comparable scale, all numeric variables were standardized. Each (numeric) factor $p$ was therefore centered and scaled to have zero mean and unit variance such that $X_p \sim \mathcal{N}(0, 1)$.

MULTICOLLINEARITY The second issue was a multicollinear regression model, or a model where two or more explanatory variables have an almost perfect linear relation. This phenomenon could originate from the inclusion of two or more factors that were minor transformations of the same variable (e.g., different $n$-gram orders). Although multicollinearity would not affect the model's prediction, it would produce a bias estimating the model's coefficients. The existence of a collinear relation between two or more variables would result in an inflation of the standard errors of their respective coefficients, leading to misleading interpretations.

Multicollinearity was detected with the variance inflation factor (VIF), which was computed for each fixed effect by regressing all other fixed effects onto that particular factor. The $R^2$ statistic of the regression model for each fixed effect then yielded the VIF as follows:

$$VIF = \frac{1}{1 - R^2} \Leftrightarrow R^2 = \frac{VIF - 1}{VIF} \tag{6.4}$$

Generally, a variable has been considered highly collinear when $VIF \geq 10$, corresponding to an $R^2 \geq .90$ (i.e., the other fixed effects explain at least 90% of the variance in a fixed effect). However, the inclusion a categorical variable could result in high VIF values for each category. For instance, a variable 'morphosyntactic category' with several dummy variables for each category (e.g., adverb, adjective, noun, and verb) would yield a high VIF for each of these highly interdependent levels. What is more, because these dummy variables were binary factors, regressing all other factors onto each dummy variable would yield a logistic regression model with a pseudo $R^2$ statistic. Conversely, regressing all other factors onto a numeric variable would yield a linear regression model with an actual $R^2$ statistic with which a VIF could be computed. For these reasons, the existence of multicollinearity was checked for numeric variables only.

GOODNESS-OF-FIT The last issue was the selection of an adequate criterion for selecting the optimal model. To statistics could be used to assess the model's goodness-of-fit, namely the Akaike information criterion (AIC) and the Bayesian information criterion (BIC):

$$AIC = -2\ln(\mathcal{L}) + 2d \tag{6.5a}$$

$$BIC = -2\ln(\mathcal{L}) + \ln(N)d \tag{6.5b}$$

where

$\mathcal{L}$ is the model's likelihood function

$d$ is the number of model dimensions or estimated parameters

$N$ is the number of observations

For both AIC and BIC, a lower value indicated that the model fitted the data better. Searching for an optimal model configuration entailed selecting the model with the minimal value for either AIC or BIC. The difference between AIC and BIC was that the latter additionally uses the number of observations as a penalty on the number of dimensions. Because all models were fitted on the same number of observations during the stepwise procedure, this additional penalty did not seem necessary. Consequently, the AIC criterion was used as the model selection criterion.

*Stepwise Forward-Backward Regression Procedure*

The stepwise forward-backward regression procedure was carried out in three stages. In the first stage, the data were preprocessed, and all problematic features were removed. All categorical variables were transformed to a one-hot encoding, and all numeric variables were transformed to a standardized scale. Next, all variables with a considerable amount of missing values were removed. In addition, various sets of highly collinear features were identified (e.g., different $n$-gram orders). For each set of collinear features, only the feature whose single-factor GLM achieved the best fit (i.e., the lowest AIC) was kept. Eighty-one features were remaining at the end of this first stage.

In the second stage, the remaining features were used to construct a GLM model with a stepwise forward selection procedure. At each step, the remain-

**Table 6.17**

*Feature and Model Selection*

| | | GLM | | GLMM | | |
|---|---|---|---|---|---|---|
| | Factor | $X$ | $Z$ | RI | RS | $X \times Z$ |
| $X_1$ | ngr.word.1.ngr.frcow16ax.$S$ | ✔ | ✔ | ✔ | ✔ +RS | ✔ |
| $X_2$ | msy.categ [e] | ✔ | ✔ | ✔ | ✔ +RS | ✔ |
| $X_3$ | msy.categ [g] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_4$ | res.FLELex-TT.A1.SFI | ✔ | ✔ | ✔ | ✔ | ✔ |
| $Z_1$ | proficiency level [A2] | | ✔ | ✔ | ✔ | ✔ |
| $X_5$ | nlm.CamemBERT.MaskedLM.$S$ | ✔ | ✔ | ✔ | ✔ | ✔ |
| $Z_2$ | proficiency level [B1] | | ✔ | ✔ | ✔ | ✔ |
| $X_6$ | msy.categ [r] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_7$ | ety.borr | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_8$ | ngr.char.2.seq.frcow16ax.$S_{\text{norm}}$ | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_9$ | occ.expo.docu.l | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{10}$ | occ.ordr.docu | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{11}$ | rea.list.stopwords | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{12}$ | nrm.OLD20.lexique381 | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{13}$ | ngr.word.1.seq.frcow16ax.$S_{\text{norm}}$ | ✔ | ✔ | ✔ | ✘ | ✘ |
| $X_{14}$ | occ.expo.sent.f | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{15}$ | len.let | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{16}$ | msy.categ [n] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{17}$ | syn.dep.rel [nsubj] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{18}$ | syn.dep.rel [acl] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{19}$ | syn.dep.rel [case] | ✔ | ✔ | ✔ | ✘ | ✘ |
| $X_{20}$ | syn.dep.rel [expl] | ✔ | ✔ | ✔ | ✘ | ✘ |
| $X_{21}$ | ngr.char.3.seq.frcow16ax.$S_{\text{norm}}$ | ✔ | ✔ | ✔ | ✘ | ✘ |
| $X_{22}$ | occ.ordr.sent | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{23}$ | syn.dep.rel [conj] | ✔ | ✔ | ✔ | ✔ | ✔ |
| $X_{24}$ | syn.dep.rel [appos] | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{25}$ | syn.dep.rel [mark] | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{26}$ | nlm.CamemBERT.MaskedLM.cloze1 | ✔ | ✘ | ✘ | ✘ | ✘ |
| $X_{27}$ | res.Manulex.G1.SFI | ✔ | ✘ | ✘ | ✘ | ✘ |

RI = random intercept model RS = random slopes model
$X \times Z$ = interaction between word complexity and learner characteristics

ing features were entered one by one into the model, starting with a null model (i.e., an empty GLM) at step one. At each step, features were removed from the remaining features when there was collinearity, and their AIC was higher than the best AIC at that step. The feature with the best AIC was added to the model and removed from the remaining features. The stepwise procedure continued until there were no remaining features. The selected features are listed in the third column of Table 6.17. Next, all learner characteristics were added to the GLM model with another stepwise forward-backward procedure. At each step, the characteristic that achieved the lowest AIC was added to the model.The previously selected features were removed from the model at each step if their removal decreased the model's AIC. The remaining features are listed in the fourth column of Table 6.17.

In the final stage, the remaining features were used to construct a GLMM model with a stepwise forward-backward selection procedure. This procedure was carried out in three passes. First, a random intercept was added to the previously fitted GLM model. The model's parameters were removed if their removal decreased the AIC of the GLMM model or if the model no longer converged. Second, a random slopes model was tested for each fixed effect in the model. Again, the model's parameters were removed if their removal decreased the AIC of the GLMM model or if the model no longer converged. Third, several interactions were examined between the word-level factors and the learner-level factors, but no significant interactions were detected. The selected parameters are listed in the fifth, sixth, and seventh columns of Table 6.17.

*Optimization*

The model was optimized with a hybrid algorithm combining the Expectation-Maximization and Quasi-Newton methods. Both the Laplace and Adaptive Gauss-Hermite Quadrature (AGHQ) methods (Pinheiro & Chao, 2006) were explored as an approximation of the log-likelihood function. The Laplace approximation is the default method used in the *lme4* package. The *GLMMadaptive* package[70] was used to compute the AGHQ approximation. Although the model's coefficients did not significantly differ between either method, the

70 https://drizopoulos.github.io/GLMMadaptive/

AGHQ approximation (with $Q = 7$ points) was used because it produced a slightly better log-likelihood and because previous studies have shown the advantage of using the adaptive quadrature over the non-adaptive quadrature in logistic mixed-effects models (see Lesaffre & Spiessens, 2001; Rabe-Hesketh et al., 2002). Lastly, Monte Carlo simulations ($n = 300$) were performed to compute prediction confidence intervals.

### 6.2.3 *Analysis*

The final GLMM model is given in Equation (6.6). The model predicted the difficulty of a word $i$ for a learner $j$ with 18 fixed effects – 16 factors at the level of the word and two factors at the level of the learner – as well as three learner-specific coefficients $u$. The model was fitted on the data from Chapter 5, which counted 4.9% ($p = .049$) of difficult items. Based on this prior probability of difficulty, the effect sizes for each regression coefficient (odds ratio) were determined following Chen et al. (2010), who proposed the following effect sizes when $p = .05$:

- $e^{\beta} \geq 1.52$ for a small effect (similar to Cohen's $d \geq 0.2$),
- $e^{\beta} \geq 2.74$ for a medium effect (similar to Cohen's $d \geq 0.5$), and
- $e^{\beta} \geq 4.72$ for a large effect (similar to Cohen's $d \geq 0.8$).

Table 6.18 describes the model coefficients. The coefficients of determination for mixed-effects models (Nakagawa & Schielzeth, 2013) showed that the model achieved a substantial explanatory power. The marginal coefficient of determination $R^2_{\text{GLMM}(m)} = .68$ indicated that a considerable proportion (68%) of the variability in difficulty was explained by the model's fixed effects alone. This result corroborated the explanatory power of various well-established factors included in the model (e.g., surprisal, frequency, and exposure). The conditional coefficient of determination $R^2_{\text{GLMM}(c)} = .76$ showed that the model achieved an increase of 8% in explanatory power when taking into account individual differences between learners on top of these fixed factors. In sum, the results show that a substantial proportion (76%) of the variability in difficulty was explained both by fixed traits of the word and the learner and by random factors attributed to each learner.

**Definition 6.18: the final GLMM model**

$$
\begin{aligned}
\ln\left[\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right] = & -5.01 + u_{0j} \\
& + 1.09\ \texttt{ngr.word.1.ngr.FRCOW16AX.}S_i + u_{1j} \\
& - 3.25\ \texttt{msy.categ [e]}_i + u_{2j} \\
& - 1.66\ \texttt{msy.categ [g]}_i \\
& - 0.51\ \texttt{res.FLELex-TT.A1.SFI}_i \\
& + 1.17\ \texttt{proficiency level [A2]}_j \\
& + 0.22\ \texttt{nlm.CamemBERT.MaskedLM.}S_i \\
& + 0.47\ \texttt{proficiency level [B1]}_j \\
& - 0.70\ \texttt{msy.categ [r]}_i \\
& - 0.25\ \texttt{ety.borr}_i \\
& - 1.67\ \texttt{occ.expo.docu.l}_{ij} \\
& - 0.24\ \texttt{ngr.char.2.seq.FRCOW16AX.}S_{\texttt{norm}i} \\
& - 0.14\ \texttt{nrm.OLD20.lexique381}_i \\
& - 0.53\ \texttt{rea.list.stopwords}_i \\
& - 0.30\ \texttt{occ.expo.sent.f}_{ij} \\
& - 0.25\ \texttt{syn.dep.rel [nsubj]}_i \\
& + 0.15\ \texttt{msy.categ [n]}_i \\
& + 0.31\ \texttt{syn.dep.rel [acl]}_i \\
& + 0.15\ \texttt{syn.dep.rel [conj]}_i
\end{aligned}
\tag{6.6}
$$

**Table 6.18**

*Generalized Linear Mixed-Effects Model of Perceived Lexical Difficulty*

| **Fixed Effects** | $\beta$ | *SE* | 95% CI | $e^{\beta}$ | VIF |
|---|---|---|---|---|---|
| Intercept | −5.01 | .21 | [−5.42, −4.61] | 0.01 | |
| [U] ngr.word.1.ngr.FRCOW16AX.*S* | 1.09 | .05 | [ 0.99, 1.19] | 2.97 | 1.1 |
| [U] msy.categ [e] | −3.25 | .21 | [−3.67, −2.83] | 0.04 | 1.0 |
| [U] msy.categ [g] | −1.66 | .08 | [−1.82, −1.51] | 0.19 | 1.5 |
| [U] res.FLELex-TT.A1.SFI | −0.51 | .01 | [−0.54, −0.48] | 0.60 | 1.1 |
| [L] proficiency level [A2] | 1.17 | .31 | [ 0.56, 1.79] | 3.23 | 1.6 |
| [U] nlm.CamemBERT.MaskedLM.*S* | 0.22 | .01 | [ 0.20, 0.24] | 1.24 | 1.1 |
| [L] proficiency level [B1] | 0.47 | .25 | [−0.02, 0.96] | 1.60 | 1.6 |
| [U] msy.categ [r] | −0.70 | .06 | [−0.82, −0.58] | 0.50 | 1.2 |
| [E] ety.borr | −0.25 | .01 | [−0.27, −0.22] | 0.78 | 1.1 |
| [X] occ.expo.docu.l | −1.67 | .16 | [−1.99, −1.36] | 0.19 | 1.1 |
| [F] ngr.char.2.seq.FRCOW16AX.$S_{norm}$ | −0.24 | .02 | [−0.28, −0.20] | 0.79 | 1.1 |
| [F] nrm.OLD20.lexique381 | −0.14 | .01 | [−0.16, −0.11] | 0.87 | 1.1 |
| [U] rea.list.stopwords | −0.53 | .06 | [−0.64, −0.41] | 0.59 | 1.4 |
| [X] occ.expo.sent.f | −0.30 | .04 | [−0.37, −0.23] | 0.74 | 1.1 |
| [U] syn.dep.rel [nsubj] | −0.25 | .04 | [−0.33, −0.16] | 0.78 | 1.0 |
| [U] msy.categ [n] | 0.15 | .02 | [ 0.10, 0.19] | 1.16 | 1.2 |
| [U] syn.dep.rel [acl] | 0.31 | .05 | [ 0.20, 0.41] | 1.36 | 1.1 |
| [U] syn.dep.rel [conj] | 0.15 | .04 | [ 0.08, 0.22] | 1.16 | 1.0 |

| **Random Effects** | | $\sigma^2$ | 95% CI | $\rho$ | |
|---|---|---|---|---|---|
| sbj_id | Intercept | 0.86 | [0.54, 1.38] | — | — |
| | msy.categ [e] | 1.462 | [0.86, 2.86] | 0.16 | — |
| | ngr.word.1.ngr.frcow16ax.*S* | 0.098 | [0.14, 0.11] | −0.69 | −0.17 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *N* | 261,942 | $R^2_{GLMM(m)}$ | 0.68 | *ICC* | 0.24 | $\ell$ | −32,995.74 |
| *M* | 56 | $R^2_{GLMM(c)}$ | 0.76 | | | | |

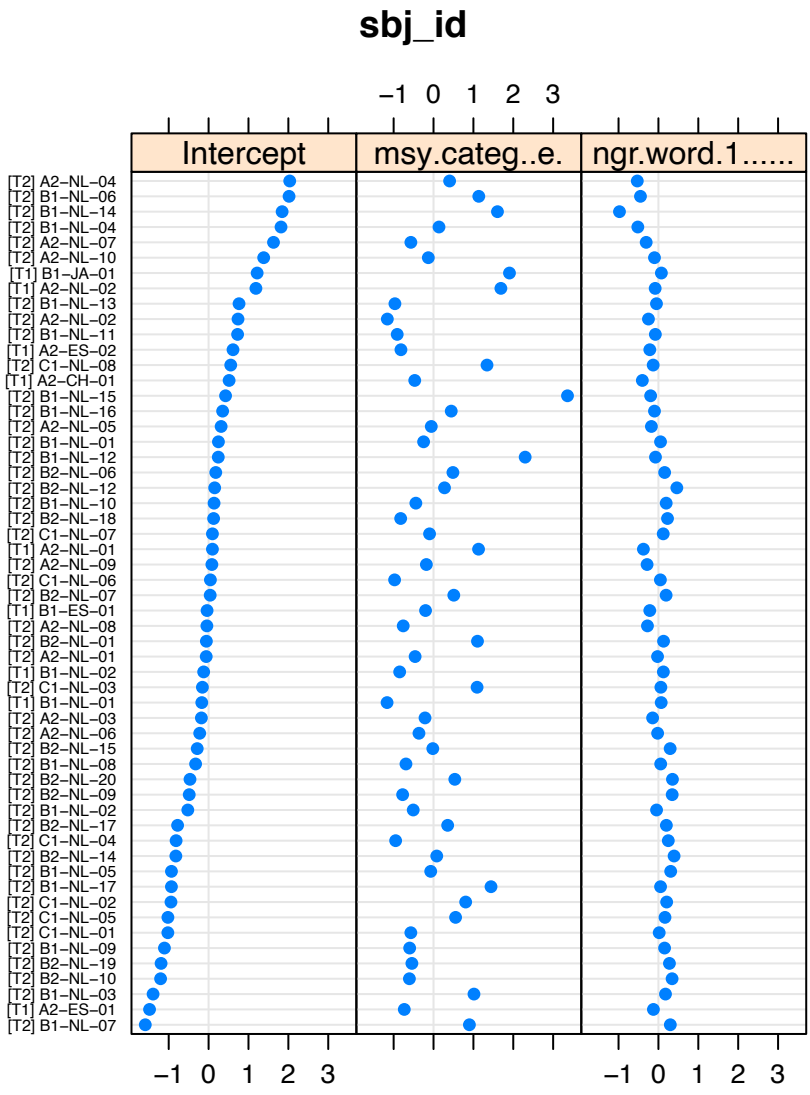
[E] etymology
[F] form
[M] meaning
[U] use
[X] exposure
[L] learner characteristic

At the individual learner level, the model included the following effects: a learner-specific intercept, two learner-specific slopes (one for morphosyntactic category and one for word surprisal; cf. *infra*), and two learner characteristics. The learner-specific coefficients for the intercept are given in Figure 6.3. The coefficients displayed a considerable degree of between-learner variance. Again, this result indicated that learners varied considerably in the extent to which they perceived difficulty. However, there was no clear-cut interaction between the learners' perceptions of difficulty and their proficiency level. Considering the various learner-specific intercepts in Figure 6.3, some intercepts were higher (i.e., $\beta_{0j} > 0$) or lower (i.e., $\beta_{0j} < 0$) than the average intercept, both for the basic (A2), intermediate (B1), and advanced (C1) proficiency levels. Similarly, the two learner characteristics selected for inclusion in the model (viz., the A2 and B1 proficiency levels) did not strongly contribute to the prediction, as can be observed from the high standard errors as well as the inclusion of zero in the 95% confidence intervals (Table 6.18).

At the individual word level, the model included mainly ($n = 11$) factors related to word use. In particular, the model predicted perceived difficulty from the word's surprisal, frequency, and grammatical function. Besides these usage-based indices, the model also included a smaller number ($n = 5$) of factors related to the degree of exposure to the word, its etymology, and its form. Two comments should be made before analyzing the effect of these various word-level features in further detail. First, it should be noted that the error term at the level of the individual word $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ was not reported since its variance is fixed in a standard logistic regression, $\sigma^2 = \pi^2/3 \approx 3.29$. Because the error term is a constant in any mixed logistic model, further analysis of this within-learners variance would not be informative. Second, the model did not suffer from multicollinearity. Table 6.18 shows that the VIF for each fixed effect was below two, which indicates that there was no fixed effect of which more than half (50%) of the variance could be explained by other factors included in the model. Because there was no inflation of the regression coefficients, the effect of the various explanatory factors could be further interpreted without bias.

**Figure 6.3**

*Random Effects for Learners in the Prediction of Perceived Lexical Difficulty*



*General Effects: Word Surprisal and Frequency*

The first and most pivotal fixed effect in the mixed-effects model was a standard information-theoretic complexity metric: unigram surprisal. Word surprisal is equivalent to the word's information content, or the logarithm of the reciprocal probability of the word occurring in the French language:

$$S_{\text{isolated}} = -\log_{10} P(w_i)$$

The more surprising the use of the word, the more likely it was perceived as difficult. A similar effect was observed for contextual surprisal, which is the logarithm of the reciprocal probability of the word occurring given the words that precede and follow the word in the sentence:

$$S_{\text{contextual}} = -\log_{10} P(w_i \mid w_1^{i-1}, w_{i+1}^{N})$$

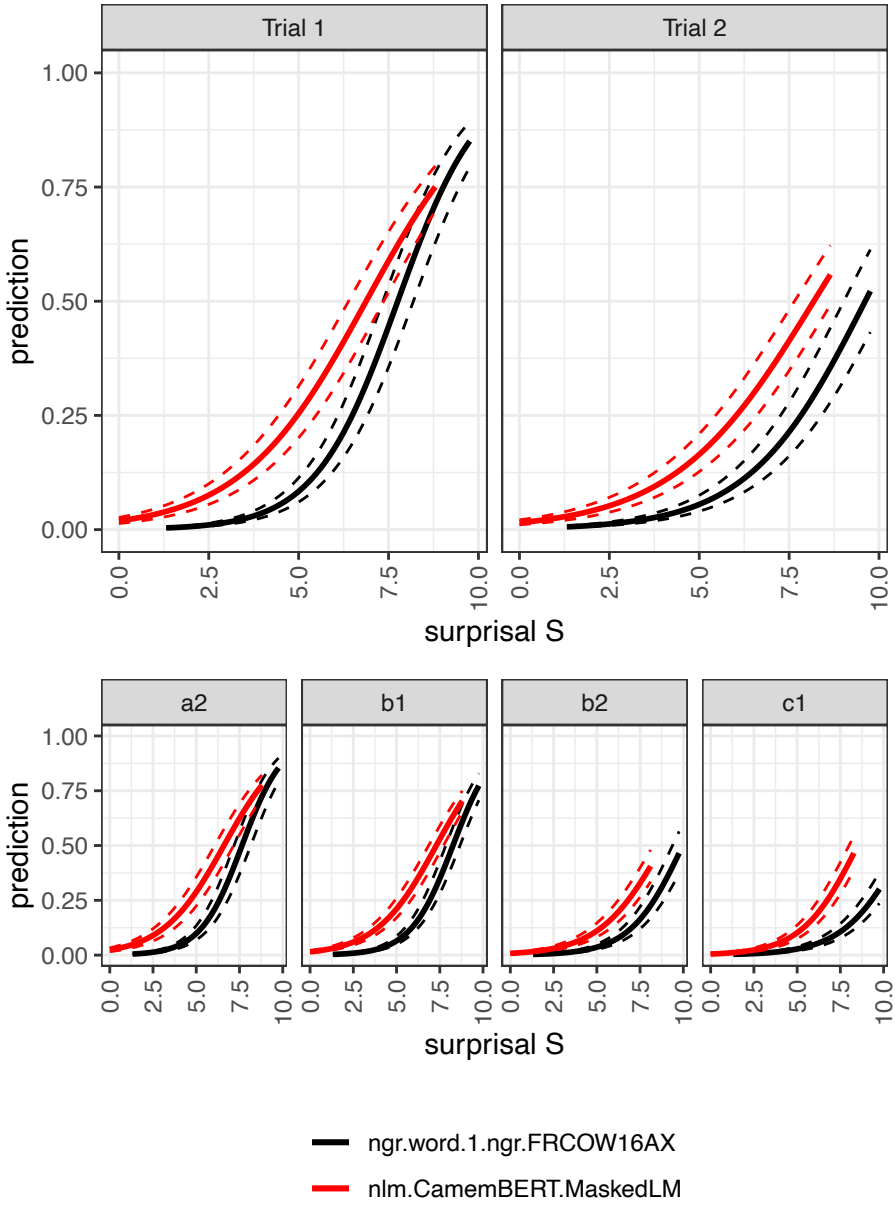
The more surprising the use of the word given its surrounding context, the more likely the word was perceived as difficult. However, Table 6.18 shows that these two measures did not have the same effect size. Whereas unigram surprisal (i.e., ngr.word.1.ngr.FRCOW16AX.*S*) had a medium effect size ($e^{\beta} > 2.74$), contextual surprisal (i.e., nlm.CamemBERT.MaskedLM.*S*) only made a minimal contribution ($e^{\beta} < 1.52$).

Figure 6.4 further illustrates the effects of isolated and contextual surprisal per each trial and per each proficiency level. The figure shows that the effects of isolated and contextual surprisal were both stronger for the first trial than for the second trial. Of course, this was very likely because the data for the first trial was more sizable as the participants read more extensively. Nevertheless, the effect of word surprisal remained noticeable even when the predictions were performed on a shorter reading task and, consequently, on less sizable data. Furthermore, the figure shows that the effect of word surprisal decreased as the proficiency level increased. Learners with a lower proficiency level tended to perceive difficulty on words that occurred less rarely, whereas those with a higher proficiency level perceived difficulty on words that occurred more rarely.

The random slope for unigram surprisal further accounted for these individual differences between learners. There was, however, only a modest

**Figure 6.4**

*The Effects of Isolated and Contextual Word Surprisal for the Prediction of Perceived Lexical Difficulty*



deviation in the effect of isolated word surprisal, as evidenced by the low between-learners variance for ngr.word.1.ngr.FRCOW16AX.*S* in Table 6.18 and by the closeness of the learner-specific coefficients in Figure 6.3. At the same time, there was a strong negative correlation between the learner-specific intercepts and the learner-specific slopes, $\rho = -.69$. The higher the extent of difficulty perceived by the learner, the smaller the slope of word surprisal, and vice versa. Put differently, when a learner perceived more words as difficult, the effect of word surprisal was more gradual. Conversely, when a learner perceived fewer words as difficult, the effect of word surprisal was steeper (i.e., a word had to be more surprising for it to be perceived as difficult).

The inverse effect was observed for CEFR-graded word frequencies from FLELex (Francois et al., 2014), as measured by the standard frequency index (SFI). The more frequently the word occurred in textbooks and readers for basic-level (A1) language learners, the less likely it was perceived as difficult. The inclusion of this frequency measure in the model gave further substance to an initial finding by Tack et al. (2016b), namely that these L2 word frequencies were better predictors than non-L2 frequencies attested in Manulex (i.e., frequencies extracted from graded L1 textbooks) or in Lexique3 (i.e., general frequencies extracted from films and books). However, contrary to the medium effect size observed for isolated word surprisal ($e^{\beta} = 2.97$; cf. *supra*), this factor only had a small negative effect, $e^{-\beta} = 1.67$. Smoothed word surprisal estimates from large-scale general language corpora were therefore better predictors than absolute frequencies from an L2-specific lexical resource, while the latter were better predictors than absolute frequencies extracted from non-L2 lexical resources.[71]

In sum, the effects observed for word surprisal underscored a logical finding: learners perceived difficulty on words that rarely occurred in the target language. This finding was consistent with the general surprisal and frequency effects observed in psycholinguistic and L2 research (cf. Chapters 2 and 3). Similarly, Paetzold and Specia (2016c) showed a significant difference in unigram log-likelihood between words perceived as difficult or not by non-native

71 As a point of comparison, it would have been preferable to have had the same surprisal estimates computed with a probabilistic language model on the FLELex corpus. However, these estimates were not available at the time of writing since the FLELex resource only included absolute normalized frequencies.

readers of English. However, on this type of data, not many studies have provided concrete and coherent regression coefficients or effect sizes. For instance, Davoodi and Kosseim (2016) showed that word frequency extracted from the Google N-gram corpus only achieved a small information gain ($IG$ = .007), whereas Ronzano et al. (2016) observed a more significant information gain ($IG$ = .38) for frequency estimates extracted from Wikipedia. Moreover, because the data was aggregated, the individual differences between learners could not be considered. The results of the current study shed more light on the effect size and degree of variability in sword surprisal on a similar difficulty measure for French.

*Bidirectional Effects: Morphosyntactic and Dependency Function*

Another set of fixed effects were related to the word's grammatical and syntactic function in the sentence. Learners perceived difficulty on nouns, clausal noun modifiers (i.e., `acl` in UD), and conjuncts (i.e., `conj` in UD), but these effects were all weak, $e^{\beta} < 1.52$. Conversely, most learners skipped named entities, grammatical words, adverbs, stop words, and nominal subjects (i.e., `nsubj` in UD). The category of grammatical words achieved a large negative effect on difficulty ($e^{-\beta} = 5.26$), whereas adverbs and stop words had a small negative effect, respectively $e^{-\beta} = 2.01$ and $e^{-\beta} = 1.70$. The nominal subject dependency relation had a weak negative effect $e^{-\beta} = 1.28$.

The largest negative effect on perceived difficulty was observed for named entities, $e^{-\beta} = 25.8$. However, the perception of difficulty on named entities varied greatly between learners, as evidenced by the random slope for msy.categ [e]. Two observations can be made from the random slope coefficients in Table 6.18 and Figure 6.3. First, the degree of between-learner variance in the effect of named entities was considerable. Second, contrary to the random slope for word surprisal, there was no strong positive correlation between the random intercept and slope. There was a slight tendency for learners who experienced more difficulty overall to also perceive more difficulty on named entities, but this correlation was weak, $\rho = .16$.

*Individual Effects: Word Exposure*

Besides the main factors of word surprisal and function, the model also included two factors: the number of repeated exposures to the lemma in the reading text and the form in the sentence. A large negative effect size was observed for the former, $e^{-\beta} = 5.31$. The more previous exposures to the lemma in the document, the less likely the word was difficult. Similarly, the number of previous exposures to the form in the sentence affected difficulty, but the size of this effect was negligible, $e^{-\beta} = 1.35$.

The frequency of exposure bears apparent similarities with word surprisal. Like word surprisal, frequency of exposure provides information on how the word is used and how frequently the word occurs. In L2 research, this variable has therefore been referred to as the frequency of occurrence (Zahar et al., 2001), rate of occurrence (Elgort & Warren, 2014), order of occurrence (Elgort et al., 2018), etc. In this study, the term frequency of exposure was used to distinguish this variable from the general frequency of occurrence of the word in the target language. At the same time, there were notable differences between the effects of word surprisal and frequency of exposure. Unlike word surprisal, the effect of frequency of exposure differed greatly between learners, as was observed from the fairly large standard error, $SE = 0.16$. This finding was consistent with previous studies and meta-analyses in L2 research showing notable differences between learners and proficiency levels in the effect of word exposure and word repetition when learning vocabulary through reading (Elgort et al., 2018; Elgort & Warren, 2014; Uchihara et al., 2019; Zahar et al., 2001).

*Minor and Inconsistent Effects: Word Etymology and Form*

Lastly, the model also predicted difficulty based on the word's etymology and form. However, these factors achieved either minor effect sizes or effects inconsistent with expectations. On the one hand, the etymological factor pertained to the number of languages having borrowed the word form. The more languages borrowed the word, the less likely the word was perceived as difficult. This factor could therefore be seen as another way of measuring the prevalence of the word. The more languages having borrowed the word,

the more prevalent the existence of the word in other languages and the more widespread the use of the word across languages. As such, this factor bears similarities with the factors of word surprisal and frequency of exposure, which also measure the prevalence of the word in the target language and the text. However, contrary to the medium effect size observed for word surprisal and the large negative effect size observed for frequency of exposure, the effect size of this etymological 'prevalence' was insignificant, $e^{-\beta} = 1.28$.

On the other hand, the two form-related factors (viz., OLD20 and character bigrams) achieved effects – albeit of little substance – that were inconsistent with the expected outcome observed in visual form recognition. Words with a higher orthographic sparsity should induce a higher latency in visual word recognition; in other words, the cognitive difficulty in recognizing these word forms would be higher. In this case, the direction of the effect was reversed. The higher the orthographic sparsity, the less likely the word was difficult. Similarly, it was expected that words with more surprising combinations of character bigrams would be more difficult. Again, the direction of the effect was reversed: the more surprising the combinations of character bigrams, the less likely the word was difficult. Nevertheless, the sizes of these two negative effects were both insubstantial, respectively $e^{-\beta} = 1.15$ and $e^{-\beta} = 1.27$. The apparent incongruity in these effects could be explained by the fact that, for instance, the normalized surprisal of character bigrams was especially high for short and frequent grammatical words such as *À* ('at', $S = 3.71$), *y* ('there', $S = 3.15$), and *eu* ('had', $S = 1.74$). As such, the effect of form-related factors on the dependent measure of perceived difficulty was unsubstantiated.

*A Final Note on Predictive Power*

Although the results showed that well-established factors could explain a substantial proportion of difficulty, it should be noted that there were limitations to the predictive power of the GLMM model. When the model was used to estimate the probability of difficulty on the entire data, the model predicted difficult words with 28% certainty on average, whereas non-difficult words were predicted as non-difficult with 96.2% certainty on average (see Table 6.19). Furthermore, the coefficient of discrimination (i.e., a coefficient that computes the absolute difference between the average estimated probability

**Table 6.19**

*Probability of Difficulty and Non-Difficulty Estimated by the GLMM Model*

| | $\hat{P}(y = 1)$ | | | | | | $1 - \hat{P}(y = 1)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | on difficult words | | | | | | on non-difficult words | | | |
| | min | $Q_1$ | *Mdn* | *M* | $Q_3$ | max | $Q_1$ | *Mdn* | *M* | $Q_3$ |
| Data | .00 | .10 | .23 | .28 | .42 | .98 | .97 | .99 | .96 | .99 |
| Trial 1 | .00 | .10 | .23 | .29 | .43 | .98 | .97 | .99 | .96 | .99 |
| Trial 2 | .00 | .10 | .20 | .26 | .37 | .94 | .98 | .99 | .97 | .99 |
| AvgPr | .00 | .07 | .15 | .20 | .30 | .94 | .70 | .85 | .80 | .93 |

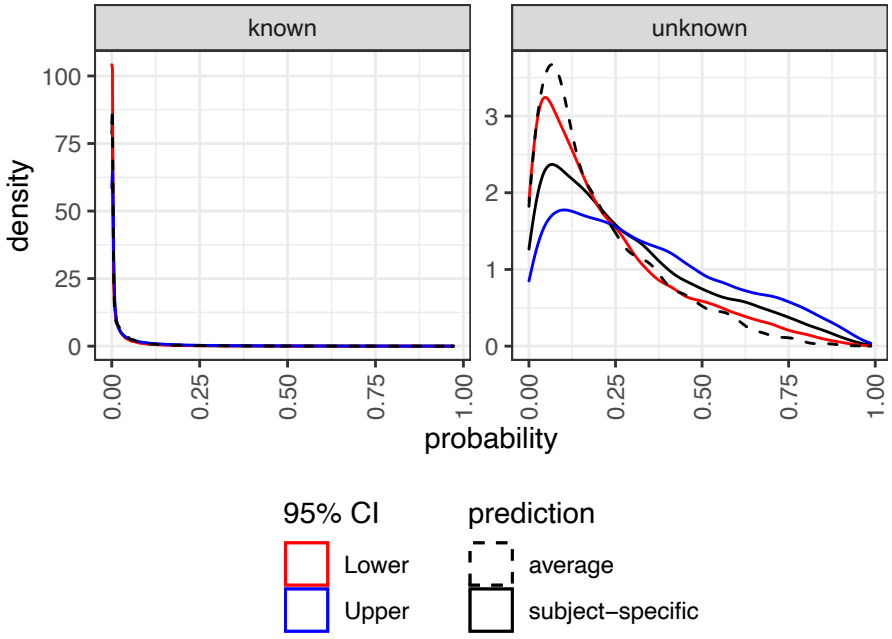
AvgPr = probability when making averaged predictions

on the positive class and the average estimated probability on the negative class; Tjur, 2009) showed that the model achieved a discriminatory power of 24.5% between difficult and non-difficult words, $D_{\text{Tjur}} = .25$. This is illustrated more clearly in Figure 6.5. In order to achieve a full discriminatory power ($D_{\text{Tjur}} \approx 1$), we would expect a peak density at $\hat{P}(y_{ii} = 1) \approx 0$ for non-difficult words – which was the case – and a peak density at $\hat{P}(y_{ii} = 1) \approx 1$ for difficult words – which was not the case.

There were two probable causes for the low discriminatory power of the GLMM model. On the one hand, the model was not robust against class imbalance, as was shown by the sensitivity towards the majority class (i.e., non-difficult words). Table 6.19 and Figure 6.5 show that the model was more certain about non-difficulty than difficulty. On the other hand, there may also have been a drawback in assuming a linear relationship between the response and explanatory variables.

Despite these drawbacks, the results showed the advantage of making learner-specific predictions over averaged predictions of difficulty. Compared to the GLMM model, the discriminatory power of making averaged predictions (i.e., by setting the random coefficients to zero) dropped 8%, $D_{\text{Tjur}} = .17$. This drop in certainty on both difficult and non-difficult words is shown more clearly in Table 6.19 and in Figure 6.5. These results were taken as support for adopting a mixed-effects modeling approach to predicting perceived difficulty.

**Figure 6.5**
*Probability of Difficulty on Non-Difficult and Difficult Words Estimated by the GLMM Model*




## 6.3 CONCLUSION

This chapter described and analyzed the construction of a generalized linear mixed model (GLMM) on subjective judgments of difficulty. The model included a total of 18 fixed factors and three learner-specific random effects. The most explanatory fixed factors were (a) the degree of isolated and contextual word surprisal, which achieved a medium and weak positive effect on perceived difficulty; (b) the number of previous exposures to the word, which achieved a large negative effect on perceived difficulty; and (c) the word's morphosyntactic function (grammatical word and named entity), which achieved a large negative effect on perceived difficulty. Notable similarities were found between, on the one hand, these subjective judgments of difficulty and, on the other hand, various objective measures of difficulty. On the one hand, the results were similar to the effects of word surprisal observed on implicit

measures of word processing difficulty (e.g., the N400 ERP in EEG measures; Frank, 2013; Frank et al., 2015). On the other hand, the results showed the effect of word repetition – or the frequency of previous exposures to the word – observed on both implicit measures (e.g., eye movements) and explicit measures (e.g., vocabulary knowledge tests) in L2 research (inter alii, Elgort et al., 2018; Elgort and Warren, 2014; Rott, 1999; Webb, 2007). In sum, the findings of the first research question (RQ6.1) showed that the most explanatory factors of the subjective judgments of difficulty introduced in Chapter 5 were similar to various well-established effects.

The findings of the second research question (RQ6.2) showed that the integration of subject-specific random effects on top of the fixed factors of difficulty resulted in an increase in explanatory and predictive power. These random effects accounted for both the varying degree of difficulty perceived by individual learners (i.e., random intercept) and the varying effect of various explanatory factors on individual learners (i.e., random slopes). On the one hand, the results showed robust effects for word surprisal. There was a low but significant between-learner variance in the effect of isolated word surprisal, which was negatively correlated with the random intercept (i.e., the effect of surprisal was stronger on learners who perceived less difficulty overall). On the other hand, the results showed a high between-learner variance in the perception of difficult named entitites.

It should be noted that the model bore similarities with but also notable differences from the statistical methods reviewed in Chapters 2 and 3.

1. For all running words in a reading text, the model estimated the probability that the word was perceived as difficult by a learner.

   This implied two differences with respect to previous studies. First, contrary to the studies reviewed in Chapter 2 and which, on average, examined various predictors of incidental vocabulary learning on a rather small selection of target words, all words in the text were considered. Second, contrary to the two CWI shared tasks reviewed in Chapter 3 and which compared various systems that predicted difficulty in either sentences or paragraphs, the full text was considered. As such, the model was trained not only on a much larger number of observations, but also on a more usual reading task. At the same time, this also required

the model to work well on a task with highly imbalanced observations, where only a small percentage of the running words were difficult.

2. The probability of difficulty was predicted from a generalized linear combination of explanatory factors, which corresponded to a set of common indicators of lexical complexity.

   Various predictors were taken into consideration based on the literature reviewed in Chapters 2 and 3. Because of practical limitations, however, not all previously studied explanatory variables could be accounted for. The most obvious drawback was that factors pertaining to, for instance, eye movements could not be computed because no eye-tracking measures were available.

3. The predictions were made by taking into account the individual differences between learners through the integration of subject-specific random effects on top of the fixed explanatory factors.

   In L2 research, a number of studies have used such mixed-effects models to investigate individual differences between learners when reading words in context (e.g., see Elgort et al., 2018; Elgort & Warren, 2014). In the two CWI shared tasks, however, such a mixed-effects approach could not be adopted because the data consisted of aggregated measures of difficulty where individual differences were no longer accounted for. As such, the study reported in this chapter was the first that investigated the use of mixed-effects models on data similar to the two CWI benchmarks.

Lastly, the findings showed advantages and disadvantages of a mixed-effects modeling approach to the prediction of perceived difficulty in reading. There was a decrease in predictive power when personal difficulties were predicted with an average model. When the learner-specific random effects were set to zero, the model predicted personal difficulty with estimates for the group of learners on average. These average-subject predictions achieved 17% of discriminative power between difficult and non-difficult words, whereas subject-specific predictions attained 25% of discriminative power. Having learners-specific training data may therefore provide an answer to the limitations observed by Finnimore et al. (2019) and Zampieri et al. (2017), showing a ceiling in system performance when making predictions of personal difficulty

based on a model trained on aggregated data. At the same time, the 25% of discriminative power attained by the mixed-effects model was far from perfect: the model achieved a much higher certainty on difficulty than on non-difficulty (e.g., see Figure 6.1 on page 251). For this reason, another study was conducted using deep learning models to account for class imbalance and possible non-linearities. This study will be described in the next chapter (Chapter 7).

## 6.A APPENDIX

This appendix provides some details on how the mixed-effects analysis was conducted. The following information is provided: the computing infrastructure (Section 6.A.1), the dependencies (Section 6.A.2), and the R commands (Section 6.A.3).

### 6.A.1 *Computing Infrastructure*

```
MacBook Pro (Retina, 15-inch, Mid 2015)

macOS Catalina 10.15.4 (19E287)
Processor: 2,2 GHz Intel Core i7 quad-core
Memory: 16 GB 1600 MHz DDR3
Graphics: Intel Iris Pro 1536 MB
```

### 6.A.2 *Dependencies*

The following R libraries were used: *car* (Fox & Weisberg, 2019), *lme4* (Bates et al., 2015), *MASS* (Venables & Ripley, 2002), *report* (Makowski et al., 2019), *sjPlot* (Lüdecke, 2020), *tydr* (Wickham & Henry, 2020).

```
R version 4.0.3

car==3.0.10
dplyr==1.0.0
ggplot2==3.3.2
GLMMadaptive==0.7.15
lattice==0.20.41
lme4==1.1.25
MASS==7.3.53
report==0.1.0
```

```
sjPlot==2.8.4
tidyr==1.1.0
```

### 6.A.3 *Commands*

GLMM model fitted with lme4

```
glmer(
  difficulty ~
    ngr.word.1.ngr.frcow16ax.surprisal + msy.categ_e + msy.categ_g +
    res.FLELex.TT.A1_SFI + pro_level_a2 + surprisal + pro_level_b1 +
    msy.categ_r + ety.borr + occ.expo.docu.l +
    ngr.char.2.seq.frcow16ax.H + nrm.lexique381.OLD20 +
    rea.list.stopwords +  occ.expo.sent.f + syn.dep.rel_nsubj +
    msy.categ_n + syn.dep.rel_acl +  syn.dep.rel_conj +
    (msy.categ_e + ngr.word.1.ngr.frcow16ax.surprisal | sbj_id)
  data=data,
  family=binomial(link="logit"))
```

GLMM model fitted with GLMMadaptive

```
mixed_model(
  fixed = "difficulty ~
    ngr.word.1.ngr.frcow16ax.surprisal + msy.categ_e + msy.categ_g +
    res.FLELex.TT.A1_SFI + pro_level_a2 + surprisal + pro_level_b1 +
    msy.categ_r + ety.borr + occ.expo.docu.l +
    ngr.char.2.seq.frcow16ax.H + nrm.lexique381.OLD20 +
    rea.list.stopwords +  occ.expo.sent.f + syn.dep.rel_nsubj +
    msy.categ_n + syn.dep.rel_acl +  syn.dep.rel_conj",
  random = ~ (msy.categ_e + ngr.word.1.ngr.frcow16ax.surprisal | sbj_id),
  data=data,
  family=binomial(link="logit"))
```

CHAPTER 7

# *Deep Learning Of Difficulty*
# A COMPARISON OF NEURAL NETWORKS WITH NON-SENSITIVE PERFORMANCE METRICS

**Abstract** This chapter presents a deep learning approach to the prediction of perceived difficulty. The learner data presented in Chapter 5 is used for training several artificial neural networks. The basic components of these neural networks are FastText word embeddings and character-based convolutional neural networks. The study compares the predictive power of personalized and contextualized networks with a repeated-measures tenfold cross-validation procedure. The study highlights that, when assessing model performance on different learners, it is necessary to use metrics that are not sensitive towards changes in test data between learners. Because previously used performance metrics are sensitive towards differences between learners, the $D$ and $\phi$ coefficients are proposed.

Artificial neural networks are founded on the idea that neural activity can be described with a basic mathematical formulation (McCulloch & Pitts, 1943). An artificial neuron receives a series of inputs, each associated with a weight, and transmits this weighted information with a simple activation function (e.g., Heaviside step function, sigmoid, hyperbolic tangent, or ReLU). As single neurons are combined and stacked into a more extensive network, the information transmitted through this computational graph allows for complex but powerful pattern recognition (Bishop, 1995). Consequently, artificial neural networks have not only been of benefit for elaborating connectionist models of human cognition (Rumelhart & McClelland, 1986) but have also shown to

be a powerful method for several NLP tasks over the past decade(s) (Bengio et al., 2003; Collobert et al., 2011; Devlin et al., 2019; Mikolov et al., 2013). Although multi-layer artificial neural networks have previously been studied in the fields of cybernetics, connectionism, and PDP, the field is currently referred to as deep learning (Goodfellow et al., 2016).

In recent years, several studies examined the use of deep learning for the prediction of lexical difficulty in reading. During the two CWI shared tasks, several participating teams evaluated the predictive power of neural network models to predict the difficulty of words in a text, as judged by a sample of L2 and L1 readers. Although the neural network model submitted to the first shared task (Bingel et al., 2016) did not outperform the top-performing systems, the neural network models that were submitted to the second shared task achieved a good performance. De Hertog and Tack (2018), for instance, developed a deep learning architecture for the prediction of lexical difficulty in English and Spanish texts. The study showed that a good performance could be achieved with the inclusion of word and character embeddings alone; the addition of other, standard features of lexical complexity as well as topic information did not further enhance the model's performance. Similarly, Bingel and Bjerva (2018) developed a neural network that achieved top performance on cross-lingual CWI. The study showed that a neural network optimized on data for English, German, and Spanish could effectively generalize these predictions to new languages such as French.

While previous studies have shown the potential of neural networks for predicting lexical difficulty in L2 reading, they have not yet fully addressed two key limitations. First of all, due to the aggregated nature of the training data used in the two CWI shared tasks, the neural networks developed by Bingel and Bjerva (2018), Bingel et al. (2016), and De Hertog and Tack (2018) were optimized to predict lexical difficulty for the target reader population in general. By contrast, other studies used neural networks to predict personal lexical difficulty from reader-specific data. Bingel, Barrett, et al. (2018), for instance, developed a neural network that predicted misreadings from the eye movements of dyslexic children while they were reading. The neural network made personalized predictions by means of multi-task learning: each individual task was set to predict whether a specific dyslexic reader would

misread a word while reading. Similarly, the preliminary results of a pilot study by Tack et al. (2016b) showed that a learner-specific neural network achieved better predictions of self-reported vocabulary knowledge in French L2 reading than a neural network trained on aggregated data for the group of learners on average. However, because the participant sample size was small ($N < 5$), the study lacked a statistically valid between-subjects analysis.

Other recent studies by Ehara (2019, 2020) have similarly adopted a personalized approach to the prediction of L2 lexical knowledge and difficulty, but these studies presented another important limitation. While the aim of Ehara's studies was to predict lexical difficulty in L2 reading with a personalized neural network, the data used to train and evaluate this network differed considerably from the data used in the CWI shared tasks. The personalized neural network proposed by Ehara was developed on data obtained from a vocabulary size test administered to a sample of EFL learners. As such, the neural network made personalized but decontextualized predictions of lexical difficulty.

On the other hand, the previously mentioned neural networks were all developed on data where the difficulty of words was measured while the participants were reading words in context (Bingel, Barrett, et al., 2018; Bingel & Bjerva, 2018; Bingel et al., 2016; De Hertog & Tack, 2018; Tack et al., 2016b). Still, even though this lexical difficulty was measured in context, these networks did not take advantage of the contextualized nature of the data to predict the difficulty of the word. In fact, because they were all standard feedforward neural networks, these models learned to predict the difficulty of a word in isolation, without integrating the surrounding context into this prediction.

In sum, it seems that previous studies have not yet fully investigated the use of neural networks to make personalized and contextualized predictions of lexical difficulty in reading. This study therefore addresses the following hypothesis: if an empirical measure of lexical difficulty in reading is both contextualized and personalized (cf. Chapter 5), this measure of difficulty is better predicted with a (neural network) model that is also both contextualized and personalized. In particular, the following research questions are examined:

RQ7.1 Does a contextualized neural network discriminate better between difficulty and non-difficulty and make predictions that are better correlated with learner perceptions?

RQ7.2 Does personalizing a contextualized neural network lead to a better discrimination between difficulty and non-difficulty and predictions that are better correlated with learner perceptions?

To this end, the study compares the predictive power of a contextualized and personalized neural network with respect to a non-contextualized and non-personalized neural network.

The neural networks were based on the deep learning architecture proposed by De Hertog and Tack (2018). However, to ensure that the predictions were contextualized and personalized, the original architecture had to be modified in two ways. On the one hand, the fully-connected feedforward layer was replaced with a recurrent neural network (RNN) layer, more specifically a bidirectional long-short term memory (BiLSTM) layer, which predicted the probability that a word was difficult based on the surrounding words in the sentence. On the other hand, the weights of the network were personalized through the integration of learner-specific information into the network.

The chapter is structured as follows. Section 7.1 describes the neural network architectures. First, the section introduces the architecture's basic components: word and character embeddings (Section 7.1.1). Next, the section illustrates the contextualized and personalized neural network architectures (Section 7.1.2). Section 7.2 presents an evaluation of the networks' predictive power. First, the section presents a sensitivity analysis of previous performance metrics (Section 7.2.1). Next, the section describes an analysis of model performance with tenfold cross-validation and multiple pairwise repeated measures analyses. Section 7.3 concludes the chapter.

## 7.1 DEEP LEARNING ARCHITECTURES

In this study, four neural network models were implemented with TensorFlow (v2.1, Keras API). These models are listed in Table 7.1. The first model was a reimplementation of the deep learning architecture for CWI proposed by De Hertog and Tack (2018) and served as a baseline in the comparative analyses. The other three models, in contrast, were adapted from this original architecture and were specifically designed to compare the effects of making contextualized and/or personalized predictions. Importantly, in order for these

**Table 7.1**
*Deep Learning Architectures*

| | Contextualized | Personalized |
|---|---|---|
| MODELS | | |
| BiLSTM-learner | Yes | Yes |
| BiLSTM-average | Yes | No |
| FFNN-average | No | No |
| BASELINE | | |
| De Hertog 2018 | No | No |

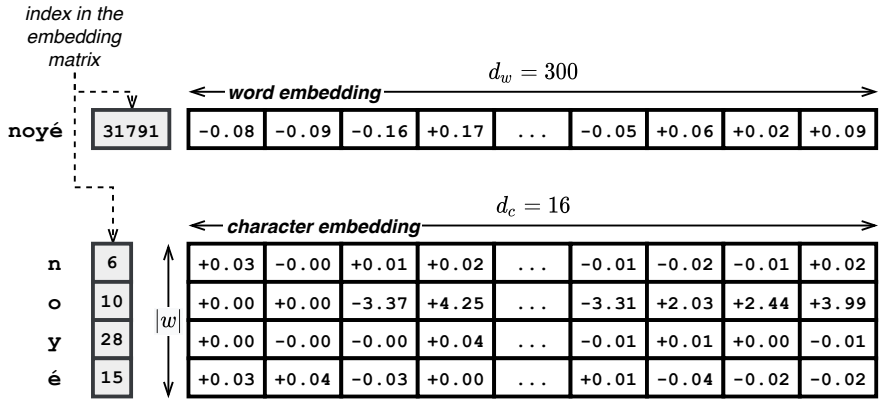
models to be comparable, it was ensured that they had the same hyperparameters (i.e., number of neurons, number of layers, etc.) and that they differed only in terms of whether they were contextualized and/or personalized. Furthermore, all of these models had a similar architecture that was composed of three parts: (a) an input layer, including word and character embeddings; (b) one or more hidden layers; and (c) an output layer, estimating the probability that the word would be perceived as difficult. This architecture will be further detailed below. First, Section 7.1.1 will describe the word and character embeddings that were used to represent the input to each neural network. Next, Section 7.1.2 will describe the global architecture of the four neural networks.

### 7.1.1 *Word and Character Embeddings*

The input to the neural network architectures was a vectorized representation of a particular word. A word was represented in two ways (Figure 7.1): as a vector of word-level features (i.e., a word embedding) and as a sequence of character-level features vectors (i.e., a stack of character embeddings). Each word-level vector, on the one hand, yielded a usage-based, distributional representation of the word. These distributional features characterized the word in terms of morphosyntax, frequency, collocations, etc. Each sequence of character-level vectors, on the other hand, was used to extract various infralexical features (e.g., complex patterns of characters or morphemes).

**Figure 7.1**

*An Illustration of Word and Character Embeddings*




*Word Embeddings*

For each word, a features vector was extracted from a matrix of pre-trained embeddings $\mathbf{W}_{\text{FastText}} \in \mathbb{R}^{V \times d_w}$, where $V$ was equal to the size of the vocabulary and $d_w$ was equal to the 300 dimensions of the pre-trained vectors. In order to pre-train a word embedding, various models have been proposed so far. De Hertog and Tack (2018), for instance, used vectors that were pre-trained with word2vec (Mikolov et al., 2013) on the English and Spanish COW corpora (Schäfer, 2015; Schäfer & Bildhauer, 2012). A disadvantage was, however, that word2vec could not provide a vector representation for out-of-vocabulary (OOV) words (i.e., words in the data that were not seen during training). As a result, the vectors were set to zero for all OOV words. A more applicable solution to the representation of OOV words was to use the FastText model (Bojanowski et al., 2017). With this model, a vector could be computed for each OOV word based on their orthographic similarity with words that were seen during training. This entailed that the model could compute a vector representation for an unseen word form (e.g., *noyèrent*, the *passé simple* of the verb *noyer*; 'to drown') because it resembled other words that were seen during training (e.g., *noy[é]* and *[pass]èrent*). For this reason, the FastText

**Table 7.2**

*$R^2$ and Pseudo-$R^2$ Values for Features of Complexity Explained by FastText Word Embeddings*

| | $R^2$ | VIF |
|---|---|---|
| ngr.word.1.ngr.FRCOW16AX.*S* | .87 | 7.98 |
| msy.categ [e] | .85 [T] | 6.59 |
| msy.categ [g] | .94 [T] | 17.26 |
| res.FLELex-TT.A1.SFI | .67 | 3.00 |
| nlm.CamemBERT.MaskedLM.*S* | .47 | 1.89 |
| msy.categ [r] | .82 [T] | 5.70 |
| ety.borr | .42 | 1.71 |
| occ.expo.docu.l | .21 | 1.27 |
| ngr.char.2.seq.FRCOW16AX.$S_{norm}$ | .84 | 6.35 |
| nrm.OLD20.lexique381 | .74 | 3.86 |
| rea.list.stopwords | .89 [T] | 9.43 |
| occ.expo.sent.f | .17 | 1.20 |
| syn.dep.rel [nsubj] | .50 [T] | 2.00 |
| msy.categ [n] | .83 [T] | 5.76 |
| syn.dep.rel [acl] | .17 [T] | 1.21 |
| syn.dep.rel [conj] | .09 [T] | 1.10 |

[T] Tjur's $R^2$ on logistic regression with liblinear solver

vectors pre-trained on the French Common Crawl by Grave et al. (2018) were used in this study.[72]

A simple regression analysis showed that these pre-trained embeddings accounted for most of the factors that explained the data and which have been covered in Chapter 6. For each word, the 300-dimensional vector was regressed on each one of the 16 explanatory factors. For each numeric factor, an OLS linear regression model was fitted with the explanatory variable as the dependent variable and with the 300 dimensions as the independent variables. For binomial factors, a logistic regression model was fitted with a liblinear solver. Table 7.2 gives the coefficients of determination for each regression model. The FastText vectors explained most of the

72 The pre-trained FastText word vectors were downloaded from the following web page: https://fasttext.cc/docs/en/crawl-vectors.html. The FastText model implemented in the *gensim* library (Řehůřek & Sojka, 2010) was used to compute a new vector for each OOV word ($V_{OOV}$ = 257).

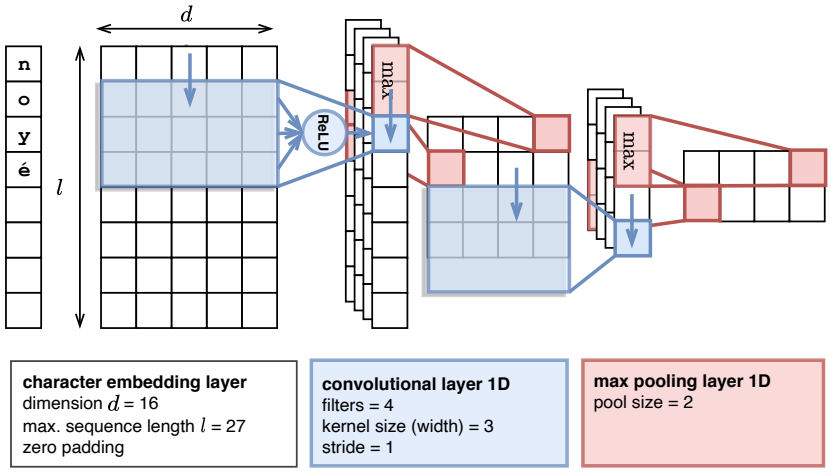
variance in the explanatory variables that achieved the highest effect sizes. The vectors explained about 90% of the variance in isolated word surprisal (i.e., ngr.word.1.ngr.FRCOW16AX.*S*, which achieved a medium positive effect on difficulty) and in morphosyntactic function (i.e., msy.categ [e] and msy.categ [g], which had a large negative effect on difficulty). Similarly, the vectors explained more than 50% of the variance in factors that achieved small effects on difficulty, such as other grammatical functions (i.e., msy.categ [r], msy.categ [n], rea.list.stopwords), word frequency (i.e., res.FLELex-TT.A1.SFI), and form-based indices of complexity (i.e., ngr.char.2.seq.FRCOW16AX.$S_{\text{norm}}$ and nrm.OLD20.lexique381). This corroborated the finding of De Hertog and Tack (2018) that a deep learning architecture captures the same amount of information as features of lexical complexity. Nevertheless and unsurprisingly, the vectors did not account for information beyond the level of the word. All $R^2$ values were below 50% for factors pertaining to contextual word surprisal (i.e., nlm.CamemBERT.MaskedLM.*S*), to the word's syntactic dependency function (i.e., syn.dep.rel [nsubj], syn.dep.rel [acl], syn.dep.rel [conj]), to the word's etymology (i.e., ety.borr), and to the learner's frequency of exposure to the word (i.e., occ.expo.docu.l and occ.expo.sent.f).

*Character Embeddings*

Besides the pre-trained word embedding, each word was also represented with a sequence of character embeddings. These character embeddings were, however, not pre-trained. Instead, they were randomly initialized and trained with a convolutional neural network (CNN), which is illustrated in Figure 7.2. The idea of using CNNs with character embeddings was first proposed by Zhang et al. (2015). De Hertog and Tack (2018) applied this idea to the task of word difficulty prediction. The first layer of the CNN included the sequence of character embeddings, which were taken from an embedding matrix $\mathbf{W}^{A \times d_c}$ where $d_c$ was equal to the number of dimensions in each character vector and $A$ was equal to the size of the alphabet. The alphabet corresponded to all distinct characters that were found in the entire data ($A = 107$). The number of vector dimensions ($d_c = 16$) were the same as those used by De Hertog and Tack (2018). Next, a convolutional layer moved a sliding window (size = 3) over the sequence of character embeddings, one by one (stride = 1), in order to extract a single

**Figure 7.2**

*Convolutional Neural Network with Character Embeddings*




piece of information from a trigram character window. The convolutional layer was composed of four filters, each of which filtered out different pieces of information. A maximum pooling layer then reduced the length of the convolutions by taking the maximum value for each two consecutive scalars in each filter. The purpose of these two layers was to identify the most salient areas of interest inside a word (e.g., character combinations or morphemes). This process of temporal convolution and maximum pooling was repeated a second time to further narrow down the areas of interest inside a word. Lastly, the output of the character CNN network was connected to an auxiliary output layer with a sigmoid activation function. The purpose of the auxiliary output layer was to optimize the weights of the character embeddings and CNN layers directly from the judgments of difficulty. As such, the aim of the auxiliary training step was to learn an association between difficulty and the most salient parts of the word.

### 7.1.2 *Contextualized and Personalized Neural Networks*

The word and character embeddings described in the previous section were used as inputs to four different neural networks, which differed in whether they were contextualized and/or personalized. The networks that were developed in this study included two feedforward neural networks (FFNNs), which made non-contextualized predictions, and two bidirectional long-short term memory (BiLSTM) networks, which made contextualized predictions. Moreover, three of these networks were non-personalized (i.e., made predictions for the group of learners on average), whereas one neural network included learner-specific encodings (i.e., made personalized predictions).

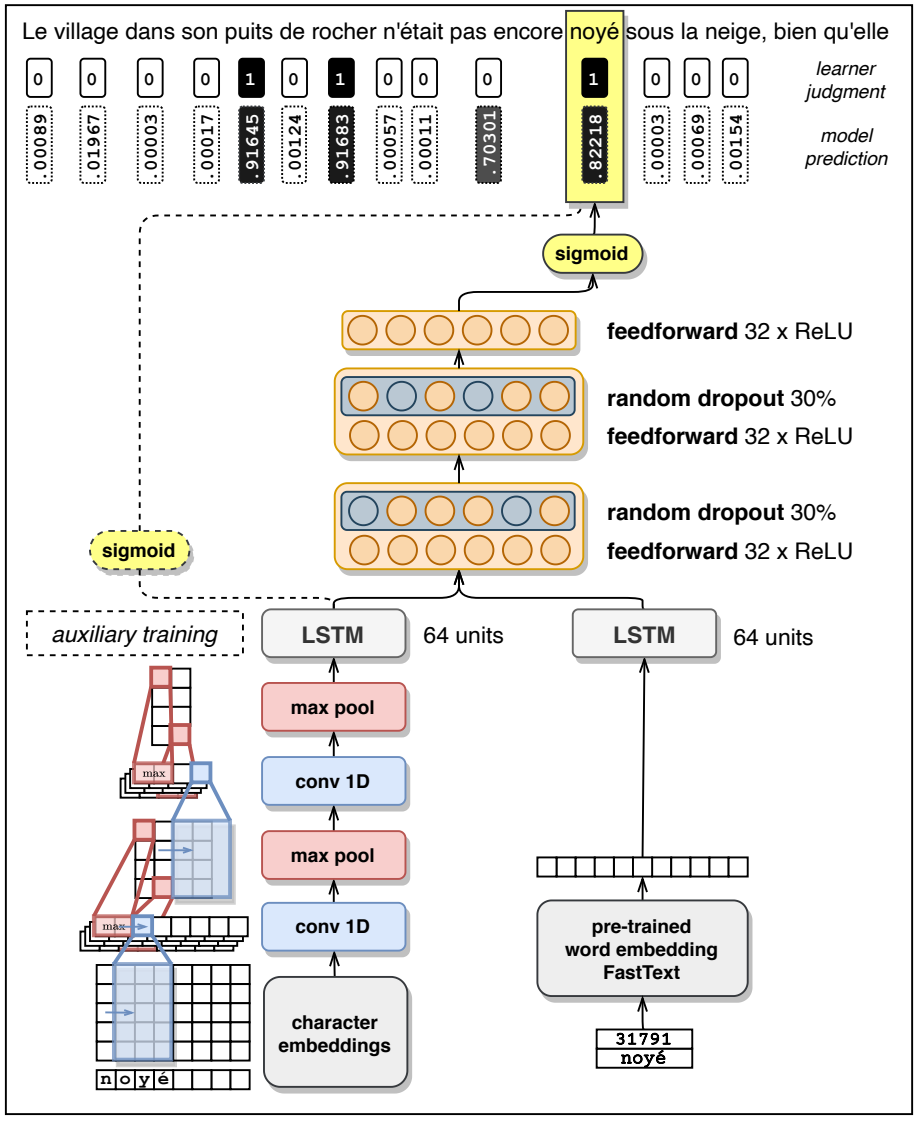
#### *Feedforward Networks*

The two feedforward neural networks (FFNNs) were non-contextualized and non-personalized models of difficulty. These networks were non-contextualized because they connected the input (i.e., a word) to the output (i.e., a prediction of difficulty) through one or more hidden layers of non-recurrent neurons. Because these networks were non-recurrent, they could only predict the difficulty of a word by taking into account the representation of that particular word. In other words, they predicted the difficulty of a word in isolation. Furthermore, these networks were non-personalized because they did not include any learner-specific encodings.

The first FFNN architecture (Figure 7.3) was a reimplementation of the model proposed by De Hertog and Tack (2018). In the original paper, the model included topic embeddings and other psycholinguistic features on top of the word and character embeddings. However, because the results showed that these additional features did not lead to an increase in performance, only the word and character embeddings were kept in this study. The model was adapted to French by replacing the English word embeddings with pre-trained vectors for French and by adapting the character alphabet from English to French. As can be observed from Figure 7.3, the model included a multilayered architecture, with three hidden layers of 32 neurons with a rectified linear unit (ReLU) activation function. In order to avoid the model from overfitting the data, random dropout was applied to the first and second hidden layers.

**Figure 7.3**

*Deep Learning Architecture for CWI (De Hertog & Tack, 2018)*



The second FFNN architecture (Figure 7.4) was a simplification of the previous architecture. The reason why the previous architecture was simplified was because its structure was too different from the final BiLSTM network that was developed in this study (cf. *infra*). In order to make reliable comparisons, it had to be ensured that both the non-contextualized FFNNs and the contextualized BiLSTMs had the same number of hidden layers, neurons, and activation functions. In consequence, a second FFNN was developed with the same number of hidden layers, neurons, and activation functions as the BiLSTM networks described below. The FFNN architecture also included three additional input layers that corresponded to the learner's identifier, proficiency level, and native language. However, these inputs were set to zero such that the FFNN model would remain non-personalized and thus only make predictions for the group of learners on average.

*Bidirectional Long-Short Term Memory Networks*

The two bidirectional long-short term memory (BiLSTM) networks were contextualized models of difficulty. They were contextualized because they connected the input (i.e., a word) to the output (i.e., a prediction of difficulty) through one or more hidden layers of recurrent neurons. Because these networks were recurrent, they predicted the difficulty of a word not only by taking into account the representation of the word, but also by looking at the surrounding words in the sentence. In other words, they predicted the difficulty of a word in context.

The BiLSTM architecture (Figure 7.5) was both a contextualized and a personalized model of difficulty. On the one hand, the model made contextualized predictions with a bidirectional layer of long short-term memory (LSTM) units. The LSTM units included three gates: an input gate, an output gate, and a forget gate (see Hochreiter & Schmidhuber, 1997). These gates regulated the information coming from the previously seen words in the sequence and determined to what extent the learned representations were retained and passed on to the next words in the sequence. With the bidirectional layer, this stream of information was coming from the words that either preceded or followed the current word in the sentence.

**Figure 7.4**

*Feedforward Neural Network Architecture*

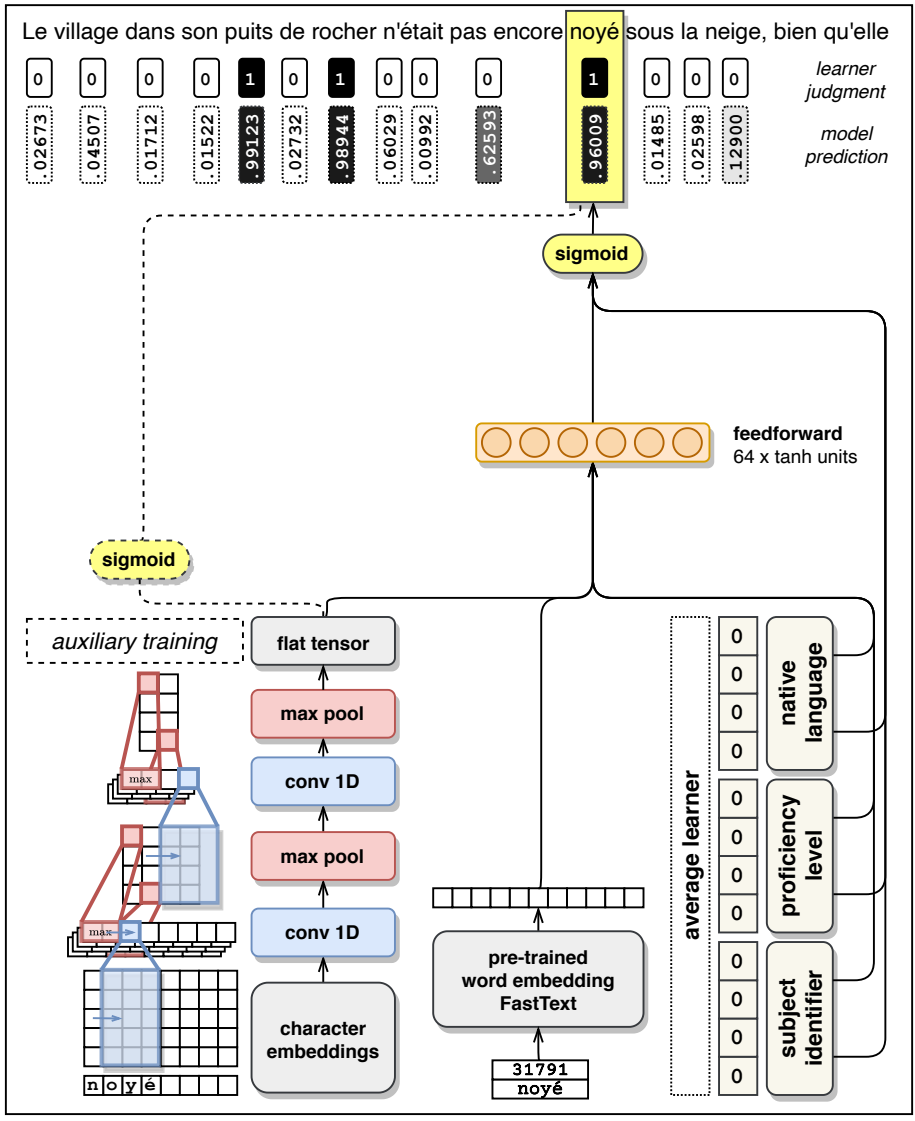
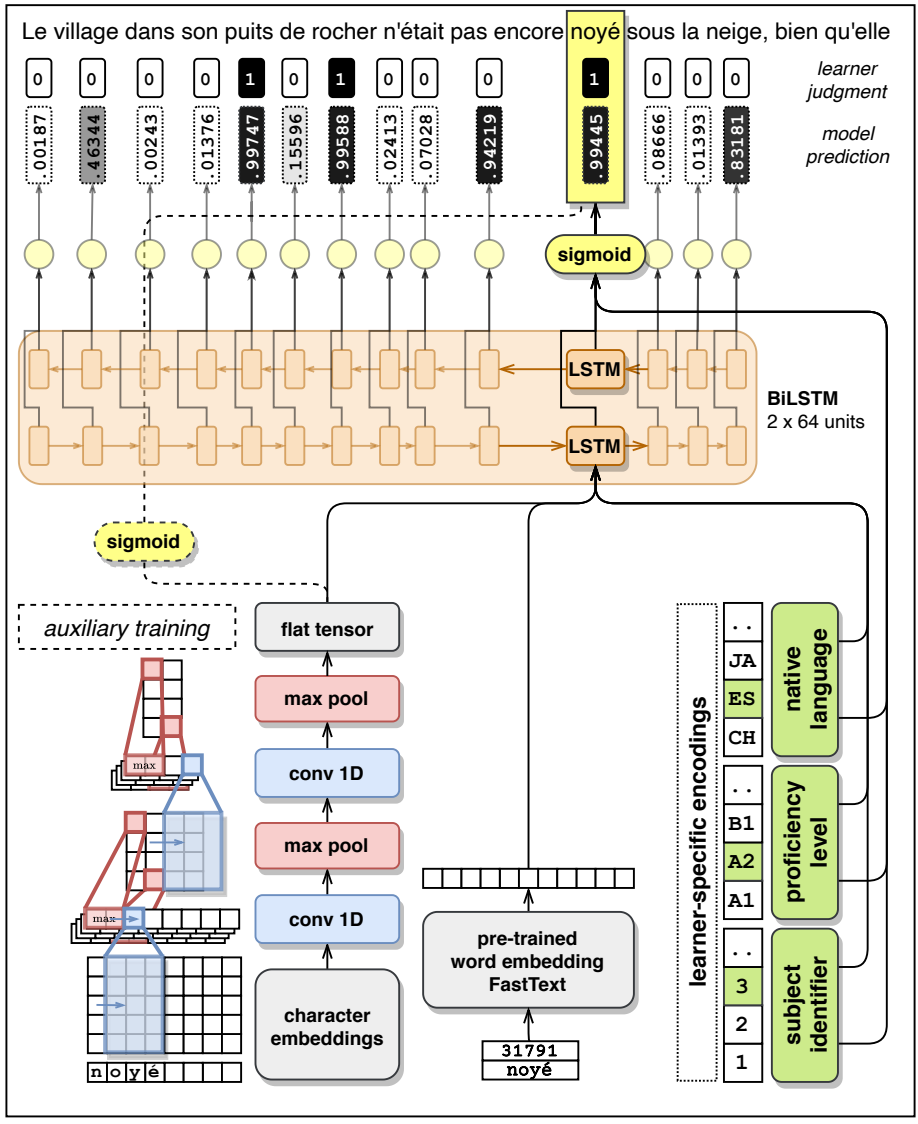


**Figure 7.5**

*Bidirectional Long-Short Term Memory Neural Network Architecture*



On the other hand, the network also integrated subject-specific information into the predictions, namely the subject's unique identifier, proficiency level, and native language, all of which were transformed to one-hot encodings. These encodings were used to add personalized weights to both the BiLSTM layer and the output layer. A second BiLSTM architecture was developed with the same architecture as in Figure 7.5, but with the learner-specific encodings set to zero as in Figure 7.4. Contrary to the previous BiLSTM, this model made non-personalized predictions of difficulty.

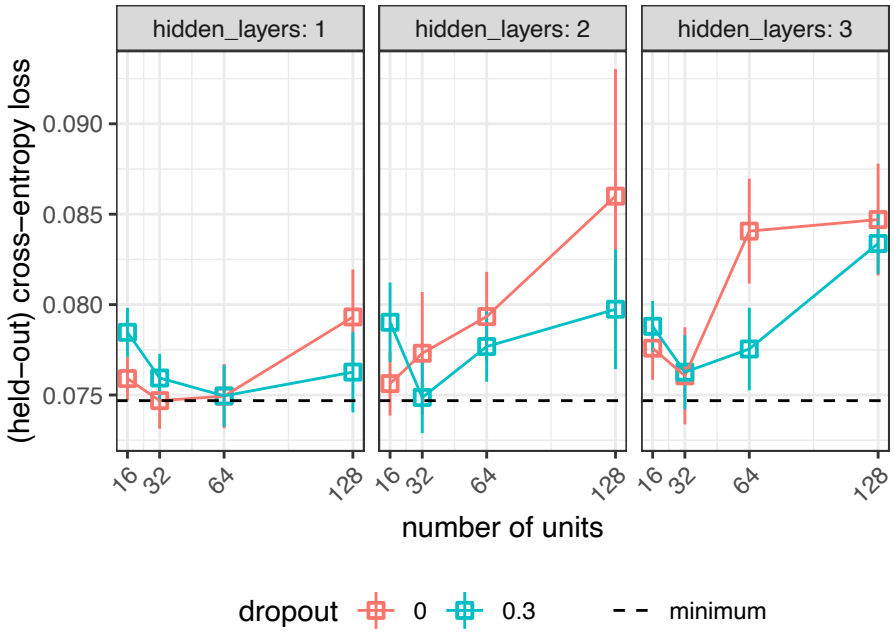
*Model Optimization and Selection*

In order to make sure the models were comparable, the same procedure was used to optimize the neural networks. Prior to optimization, all sequences were processed such that a zero mask was set to all non-alphanumeric tokens. All sentences were padded to the maximum sentence length (97 words) and all labels were binarized (1 = *difficult* and 0 = *non-difficult*). The weights for each layer were initialized with the same fixed seed set to zero. The networks were optimized on binary cross-entropy loss with the Adam algorithm. The loss function was optimized with balanced class weights to avoid overfitting the majority class (viz., non-difficult words). Early stopping was used to stop the optimization procedure when the loss increased on 10% validation data that was held out from the training data. Overall, the optimization procedure did not take more than 10 iterations.

Finally, a grid search was performed to optimize the model's hyperparameters. The personalized BiLSTM model was trained with a varying number of hidden layers (1, 2, and 3) and hidden units (16, 32, 64, and 128) and with or without the 30% random dropout used in the original architecture. The selection procedure was performed on the same 10% of validation data that was held out from the training data. The procedure was repeated ten times during cross-validation (see Section 7.2). Figure 7.6 shows the validation cross-entropy loss of each model configuration. The search showed that a model with a single hidden layer achieved the lowest loss. When a second or third hidden layer was added, the model showed signs of overfitting. Furthermore, the search also showed that the lowest loss was achieved with 32 or 64 hidden units. Because a model with 64 hidden units did not seem to overfit the data,

**Figure 7.6**

*Model Selection on Held-out Tenfold Cross-Validation*



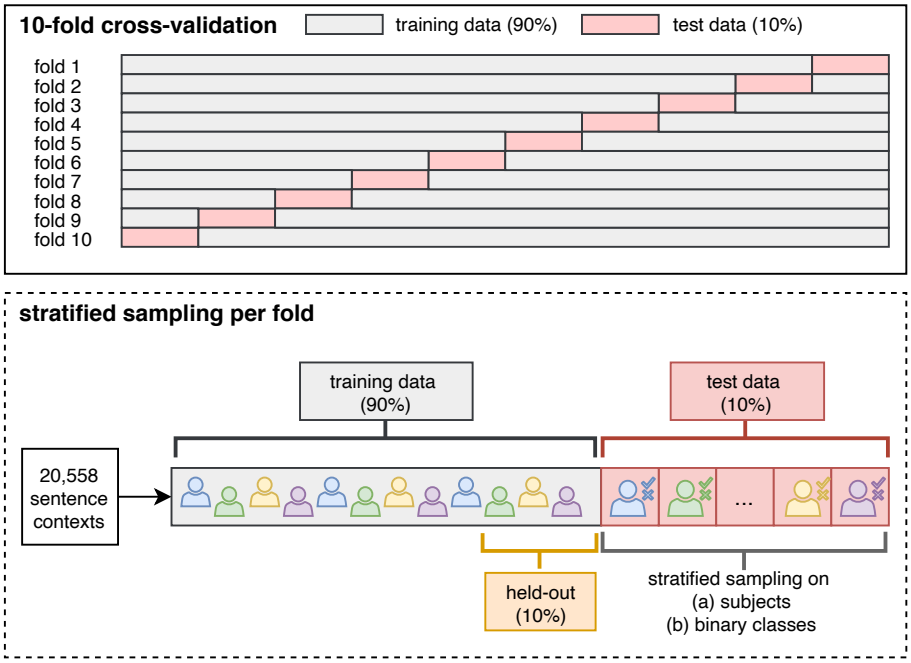

a model with one hidden layer of 64 units was chosen. Finally, because the random dropout did not lead to an decrease in validation loss, no dropout was applied. In consequence, the model configuration that was selected for training the BiLSTM and FFNN models was composed of a single hidden layer of 64 units without random dropout.

## 7.2 MODEL EVALUATION

The performance of the models described in the previous section was evaluated with a standard tenfold cross-validation procedure. Like in Chapter 6, this evaluation was performed on the data described in Chapter 5. There were, however, two important differences with respect to the previous study. First, because the contextualized networks required a sequence learning task, the 261,942 observations were grouped into their respective sentence contexts. This grouping resulted in a total of 20,558 of sentences with judgments of

**Figure 7.7**

*Tenfold Cross-Validation with Stratified Sampling*




difficulty. Second, instead of fitting each model on all observations, the sentence contexts were sampled into a training set and a test set (Figure 7.7). For each cross-validation fold, the test set was defined with stratified sampling. More specifically, it was ensured that each test set contained data for each subject and that there was at least one judgment of difficulty for each subject. This was needed to evaluate how well each model performed on unseen data for each individual subject. Finally, the ten training and test splits were fixed beforehand such that the same cross-validation procedure was used for each repeated modeling experiment. For more detailed information, see the Appendix (Section 7.A).

### 7.2.1 *Sensitivity Analysis of Performance Metrics*

Before comparing model performance on learner-specific test data, there was one concern that had to be addressed. As described in Chapter 5, the data

showed a considerable degree of variance between learners. In particular, there were two important sources of variance that emerged from the learner-specific distributions of difficulty $Y_j \sim B(n_j, p_j)$. Compared to others, some learners read more or less extensively (i.e., parameter $n_j$, the number of words read by a subject) and/or experienced more or less difficulty (i.e., parameter $p_j$, the probability that the subject finds a word difficult). Consequently, the main concern was that, when evaluating the predictive power of a model on data for different learners, the performance metric would be influenced by this between-learner variance and would favor cases where more words were read or a larger proportion of words were found difficult. To address this concern, a sensitivity analysis was conducted with the aim to find at least one performance metric that was not sensitive towards these two sources of variance between learners. In the analysis, the potential sensitivity of the previously used benchmark metrics was compared against other performance metrics used in binary classification and logistic regression.

*Performance Metrics*

Previous studies used different metrics to evaluate the accuracy of the predictions of difficulty with respect to the true class (i.e., manual lexical simplification, difficulty judgment, etc.). Shardlow (2013a) compared the performance of computational methods for identifying complex words based on four common statistical measures, namely accuracy, precision, recall, and $F_1$-score. In order to focus on a better recall of difficult words for automatic lexical simplification, Paetzold and Specia (2016a) proposed the use of the G-score (i.e., the harmonic mean between accuracy and recall). In order to balance performance on both the positive and negative classes, other shared tasks in simplification and difficulty prediction resorted to weighted and macro-averaged $F_1$-scores (Štajner et al., 2016; Yimam et al., 2018). Although these metrics are standard performance metrics, they have, however, two drawbacks.

A first drawback of the previously used performance metrics is that they do not handle class imbalance, which refers to the case where a particular class is overrepresented in the data. In lexical difficulty prediction, the majority class is the negative or non-difficult class, corresponding to, for instance, 95% of observations (i.e., words in a reading text; cf. Chapter 5). The existence of

such imbalanced prior distributions not only poses a problem for machine and deep learning methods (He & Garcia, 2009; Johnson & Khoshgoftaar, 2019), it also affects system performance. For instance, a constant decision function that always yields the negative class will achieve 95% of accuracy, provided that the negative class represents 95% of observations. In other words, the accuracy score is sensitive towards class imbalance in that it favors decision functions that are optimized to predict the majority class. Furthermore, a constant decision function that always yields the negative class will achieve a higher accuracy on data where a learner has less overall difficulty (e.g., 99% of non-difficult words, 99% accuracy) than in cases where a learner has more overall difficulty (e.g., 90% of non-difficult words, 90% accuracy), even though the predictions remain uninformative in both cases. The accuracy score is, therefore, not only sensitive towards imbalance but also sensitive towards changes in the prior class distribution.

By contrast, a measure that appears robust towards class imbalance and which is frequently used in bioinformatics is Matthews (1975) correlation coefficient or MCC (e.g., see Boughorbel et al., 2017). The MCC metric is equivalent to Pearson's (1904) $\phi$ (Phi) correlation coefficient, which measures the strength of association between two binomial variables (in this case, the true and predicted classes) from a 2x2 contingency table:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{7.1}$$

An important corollary is that, if the Phi correlation coefficient is insensitive to an imbalance in the prior class distribution, the coefficient will also be insensitive to changes in this prior class distribution. Put differently, if the Phi coefficient is insensitive towards an imbalance in the percentage of words that are found difficult or not, the coefficient will also be insensitive to a variance in the percentage of words that are found difficult between learners. For this reason, the $\phi$ coefficient was examined as a non-sensitive performance metric.

A second drawback of the previously mentioned benchmark metrics is that they require binary predictions. If a system predicts a binary class, this binary value can be compared directly to the true class. However, if a system predicts the probability of the observation belonging to the positive class $P(y = 1) \in [0, 1]$, this continuous outcome needs to be discretized to a binary

value. A common way of discretizing the probability is to set a cut-off point at $P(y = 1) > 0.5$, which classifies an observation into the positive class $y = 1$ if the predicted probability is at least 50%. A shortcoming of this discretization method is that the outcome no longer retains information about the degree of certainty in the predictions. For instance, if a given system predicts the difficulty of a truly difficult word with 55% probability and another system predicts the same word as difficult with 95% probability, both predictions will correctly classify the word as difficult based on a 50% cut-off value. If both systems consistently produce the same classifications, they will obtain the same performance score, even though the latter makes more certain predictions than the former. Consequently, the previously mentioned performance metrics could attribute a high classification score to a system that predicts the positive class with a low degree of certainty.

A metric that retains information about the degree of certainty in the predictions is Tjur's (2009) coefficient of discrimination, $D \in [0, 1]$. The $D$ coefficient – also referred to as Tjur's pseudo-$R^2$ – is a goodness-of-fit metric used in logistic regression models. The coefficient measures the absolute difference between the mean probability (of a positive outcome) on the $n_1$ observations belonging to the positive class and the mean probability (of a positive outcome) on the $n_0$ observations belonging to the negative class.

$$D = \left| \frac{\sum_{i=1}^{n_1} P(y_i = 1)}{n_1} - \frac{\sum_{j=1}^{n_0} P(y_j = 1)}{n_0} \right| \tag{7.2}$$

The coefficient has a maximum value ($D = 1$, i.e., full discriminative power) if the mean probability on the positive class is 100% and the mean probability on the negative class is 0%. The coefficient has a mimimum value ($D = 0$, i.e., no discriminative power) if the mean probability on the positive and negative classes are equal. The advantage of Tjur's $D$ is therefore that it determines how well a system or model can discriminate between two classes without discretizing the probabilistic outcome to a binary value. For this reason, the $D$ coefficient was examined as a potentially non-sensitive performance metric.

*Local Sensitivity Analysis*

A one-factor-at-a-time (OFAT) sensitivity[73] analysis was used to verify whether the F-score, G-score, $\phi$, and $D$ could be used as non-sensitive performance metrics. The OFAT analysis determined the local sensitivity of a function $f$ towards changes in one parameter while keeping all other parameters at a constant value. In this study, $f$ corresponded to the outcome of a given performance metric function, which had the following underlying parameters: those of the true binomial data distribution $Y \sim B(N, p)$ and those of the predictive model $m$. The main assumption was that, if a given metric $f$ was insensitive towards changes in both $N$ and $p$, it was certain that a change in $f$ would only reflect a change in the model $m$.

For each performance metric and for each parameter $N$ and $p$, the sensitivity of $f$ was computed as a percent change from a base value. The base value for $N$ was set to the maximal number of words read by a learner ($N_{\text{base}}$ = 21,048). The base value for $p$ was set to the average percentage of words perceived as difficult ($p_{\text{base}}$ = 0.05), which coincidentally corresponded to the threshold of known words (95%) required for adequate reading comprehension (Laufer & Ravenhorst-Kalovski, 2010). The model $m$ was set to a constant decision function that predicted the positive class with $P(y = 1) = 1$. Finally, a local sensitivity coefficient (7.3; see Hamby, 1994) was computed as an approximation of the partial derivative of $f$ with respect to factor $x$.

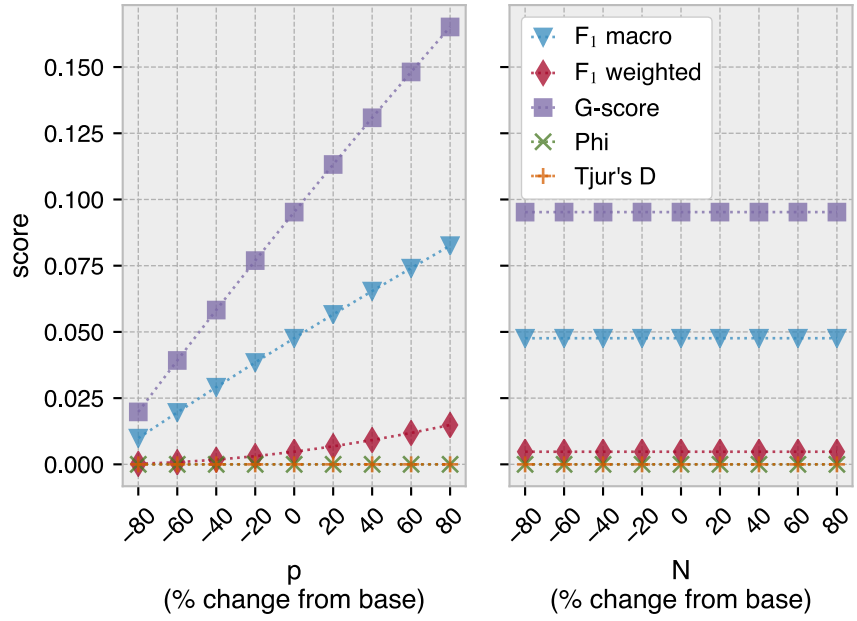
$$S_x^{\text{local}} = \frac{\%\Delta f}{\%\Delta x} = \frac{f - f_{\text{base}}}{f_{\text{base}}} \div \frac{x - x_{\text{base}}}{x_{\text{base}}} \tag{7.3}$$

The results showed that all previous benchmark metrics were biased towards a larger $p$ and, to a trivial degree, a larger $N$. The local sensitivity coefficients showed comparable degrees of sensitivity for the G-score and the macro-averaged $F_1$-score ($S_p^{\text{local}}$ = 0.98; $S_N^{\text{local}}$ = 0.00005), the accuracy score ($S_p^{\text{local}}$ =

73 The goal of a sensitivity analysis is to identify parameters of a mathematical model or function that cause a variability in the outcome of that model or function. Put differently, a sensitivity analysis determines whether the outcome of a mathematical model or function is sensitive towards changes in, for instance, the underlying data distribution. The literature is abundant in methods for analyzing sensitivity, ranging from local analyses, recording the sensitivity of one factor of uncertainty (Eschenbach, 1992; Hamby, 1994), to global analyses for complex mathematical models (Saltelli & Annoni, 2011; Sobol', 2001). Because a complex global analysis was beyond the needs of this study, a local sensitivity analysis was used instead.

**Figure 7.8**

*Spider Plot of Changes in Constant Baseline Performance*




1.0; $S_N^{\text{local}}$ = 0.00005), and the weighted $F_1$-score ($S_p^{\text{local}}$ = 1.39; $S_N^{\text{local}}$ = 0.0001). More specifically, the local sensitivity coefficients indicated the rate of change in $f$ when only the $p$ factor was allowed to vary: if the percentage of words found difficult increased by 1%, the performance would result in a 1% increase for the macro-averaged $F_1$ and accuracy scores and a 1.39% increase for the weighted $F_1$-score. In other words, all performance metrics increased when the percentage of difficult words increased. As a result, it could not be ascertained that an increase in performance would only reflect an increase in the predictive power of a model $m$. By contrast, the spider plot[74] for $p$ and $N$ (Figure 7.8) showed that the $\phi$ and $D$ coefficients were both insensitive to changes in $p$. If the percentage of words found difficult increased, both coefficients remained constant. Because both coefficients remained unchanged when the percentage

74 A spider plot is a commonly used technique for visualizing the sensitivity of an outcome (y-axis) with respect to percent changes in the value of one or more factors (x-axis) (see Eschenbach, 1992).

of difficult words increased, it was ascertained that a change in performance would only reflect a change in the predictive power of a model *m*.

In conclusion, the results of the local sensitivity analysis showed that the previous benchmark metrics were sensitive to between-learner variance. When more words were found difficult, the $F_1$, G, and accuracy scores for a constant baseline increased. When comparing these performance metrics on learner-specific data, we therefore cannot say whether a higher score reflects a growth in performance or simply an increase in personal difficulty. Conversely, the $\phi$ and $D$ coefficients appeared insensitive towards differences between learners in terms of the percentage of words that were found difficult. For this reason, the $\phi$ and $D$ coefficients were used to compare model performance between learners.

### 7.2.2 *Repeated Measures Performance Comparisons*

A series of repeated measures analyses were run to compare the models described in Section 7.1. These models included (a) the contextualized and personalized model (BiLSTM-learner), (b) the contextualized and non-personalized model (BiLSTM-average), (c) the non-contextualized and non-personalized model (FFNN-average), and (d) the baseline model (De Hertog 2018). Multiple Friedman tests were carried out to compare model performance on both discriminatory power ($D$) and correlation ($\phi$), with a view to providing answers to the following hypotheses:

1. A contextualized model achieved a better performance than a non-contextualized model.

   $H_1$ : BiLSTM-average > FFNN-average

2. A personalized model achieved a better performance than a non-personalized model.

   $H_1$ : BiLSTM-learner > BiLSTM-average

   $H_1$ : BiLSTM-learner > FFNN-average

3. A contextualized and personalized model achieved a better performance than the baseline.

$H_1$ : BiLSTM-learner > De Hertog 2018

Post hoc one-tailed Wilcoxon signed-rank tests were run to identify a significant difference in performance between the four models. The p-values were adjusted for multiple comparisons using the Holm method. The rank-biserial correlation coefficient was used to measure the size of the effect.

Three series of analyses were conducted. The first analysis compared the models in terms of the average probability of difficulty predicted on difficult and non-difficult words. The second analysis compared the models' overall performance on cross-validation. The third and final analysis compared the predictive power of the models per subject. The results of these three analyses will be discussed in the next three sections.

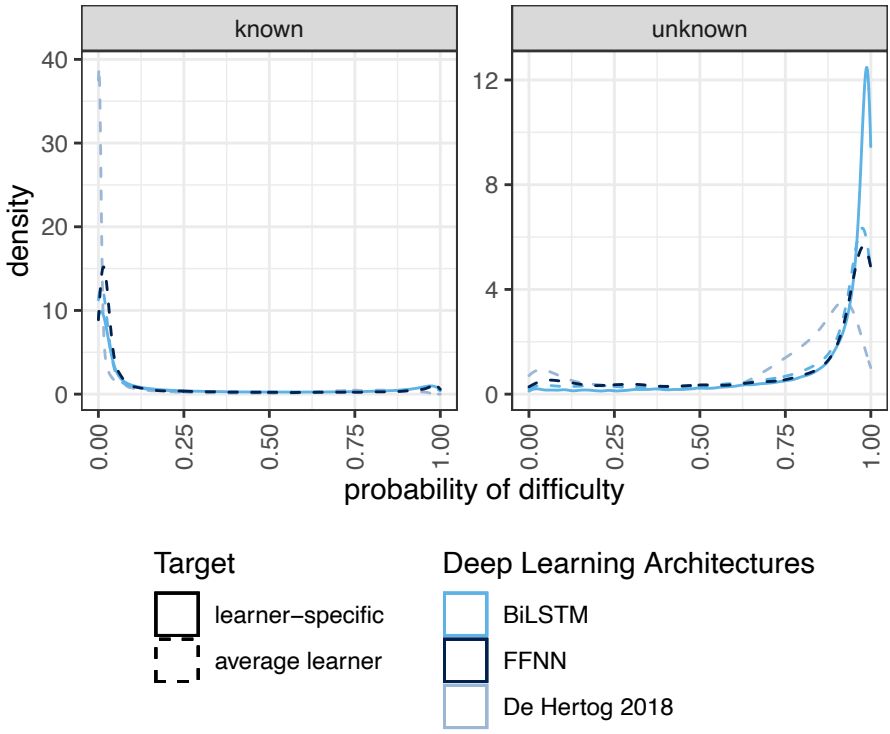
*Probability of Difficulty on Difficult and Non-Difficult Words*

In order for the model to accurately discriminate between difficult and non-difficult words, the ultimate goal was to achieve a maximal probability of difficulty on difficult words and a minimal probability of difficulty on non-difficult words. The results showed that all four neural network models were generally good at distinguishing between difficult and non-difficult words. Figure 7.9 shows that all models predicted the majority of difficult words with a probability higher than 50% and the majority of non-difficult words with a probability below 50%. However, the figure also shows that the contextualized and personalized model (viz., BiLSTM-learner) made more certain predictions than the non-personalized model, the non-contextualized model, and the baseline model. The performance of the baseline model, on the other hand, appeared to be critically skewed towards achieving a higher certainty of non-difficulty rather than difficulty.

A Friedman test confirmed this initial observation and showed that all models differed significantly in terms of the average probability of difficulty predicted on difficult words, $\chi^2_F(3) = 28.9$, $p < .001$. Table 7.3 gives the average predicted probabilities per each model and per each class of words. Post hoc one-tailed Wilcoxon signed rank tests showed that the BiLSTM-average model predicted difficult words with a significantly higher certainty than the FFNN-average model ($W = 54$, $p = .002$, $r = .85$, large effect size) and the baseline model ($W = 55$, $p = .002$, $r = .87$, large effect size). Similarly, the

**Figure 7.9**

*Probability of Difficulty on Difficult and Non-Difficult Words*




BiLSTM-learner model predicted difficult words with a significantly higher probability than the BiLSTM-average model ($W = 55$, $p = .003$, $r = .87$, large effect size), the FFNN-average model ($W = 55$, $p = .003$, $r = .87$, large effect size), and the baseline model ($W = 55$, $p = .003$, $r = .87$, large effect size). Because all pairwise comparisons achieved large effect sizes, it was concluded that a contextualization and personalization of predictions lead to a large and significant increase in certainty on difficult words.

A second Friedman test showed that the models also differed significantly in terms of the average probability of difficulty predicted on non-difficult words, $\chi^2_F(3) = 25.7$, $p < .001$. Post hoc one-tailed Wilcoxon signed rank tests showed that the BiLSTM-average model predicted non-difficult words

**Table 7.3**

*Average Probability of Difficulty on Difficult and Non-Difficult Words*

| | C | P | *N* | $P(y = 1)$ difficult words | | $P(y = 1)$ non-difficult words | |
|---|---|---|---|---|---|---|---|
| | | | | $M \pm SD$ | *Mdn* | $M \pm SD$ | *Mdn* |
| MODELS | | | | | | | |
| BiLSTM-learner | ✔ | ✔ | 10 | $0.87 \pm 0.02$ | 0.87 | $0.23 \pm 0.01$ | 0.23 |
| BiLSTM-average | ✔ | ✘ | 10 | $0.80 \pm 0.01$ | 0.80 | $0.21 \pm 0.01$ | 0.21 |
| FFNN-average | ✘ | ✘ | 10 | $0.78 \pm 0.01$ | 0.78 | $0.19 \pm 0.01$ | 0.19 |
| BASELINES | | | | | | | |
| De Hertog 2018 | ✘ | ✘ | 10 | $0.69 \pm 0.04$ | 0.71 | $0.15 \pm 0.02$ | 0.15 |

[C] contextualized model [P] personalized model

with a significantly lower certainty[75] than the FFNN-average model ($W = 51$, $p = .007$, $r = .76$, large effect size) and the baseline model ($W = 55$, $p = .002$, $r = .87$, large effect size). Similarly, the BiLSTM-learner model predicted difficult words with a significantly lower certainty than the BiLSTM-average model ($W = 54$, $p = .004$, $r = .85$, large effect size), the FFNN-average model ($W = 54$, $p = .004$, $r = .85$, large effect size), and the baseline model ($W = 55$, $p = .003$, $r = .87$, large effect size). However, it should be observed that the decrease in certainty on non-difficult words was considerably lower than the increase in certainty on difficult words. Compared to the baseline model, the BiLSTM-learner model displayed a larger (18%) increase in certainty on difficult words, but a much smaller (8%) decrease in certainty on non-difficult words. It was therefore concluded that a contextualized and personalized model achieved a significantly higher certainty of difficulty, whilst retaining a low probability of difficulty on non-difficult words.

*Discriminatory Power and Correlation with Learner Judgments*

A second series of analyses were performed to examine the degree of discriminatory power and correlation on tenfold cross-validation. Table 7.4 shows the

75 Because the model predicted non-difficult words with a significantly higher probability of difficulty on average, this entailed that the model predicted non-difficult words with a significantly lower certainty of non-difficulty.

**Table 7.4**

*Mean and Median Performance on Tenfold Cross-Validation*

| | C | P | N | $D$ | | $\phi$ | |
|---|---|---|---|---|---|---|---|
| | | | | $M \pm SD$ | $Mdn$ | $M \pm SD$ | $Mdn$ |
| MODELS | | | | | | | |
| BiLSTM-learner | ✔ | ✔ | 10 | $0.64 \pm 0.02$ | 0.64 | $0.36 \pm 0.01$ | 0.36 |
| BiLSTM-average | ✔ | ✘ | 10 | $0.59 \pm 0.01$ | 0.60 | $0.34 \pm 0.01$ | 0.34 |
| FFNN-average | ✘ | ✘ | 10 | $0.59 \pm 0.02$ | 0.59 | $0.35 \pm 0.01$ | 0.35 |
| BASELINES | | | | | | | |
| De Hertog 2018 | ✘ | ✘ | 10 | $0.54 \pm 0.03$ | 0.55 | $0.35 \pm 0.01$ | 0.35 |

[C] contextualized model [P] personalized model

mean and median performance on discriminatory power ($D$) and correlation ($\phi$). A Friedman test showed that all neural network models differed significantly in terms of discriminatory power, $\chi^2_F(3) = 28.08$, $p < .001$. Post hoc one-tailed Wilcoxon signed rank tests showed that the BiLSTM-average model achieved a significantly higher $D$ than the FFNN-average model ($W = 50$, $p = .01$, $r = .73$, large effect size) and the baseline model ($W = 55$, $p = .002$, $r = .89$, large effect size). Similarly, the BiLSTM-learner model achieved a significantly higher $D$ than the BiLSTM-average model ($W = 55$, $p = .003$, $r = .89$, large effect size), the FFNN-average model ($W = 55$, $p = .003$, $r = .89$, large effect size), and the baseline model ($W = 55$, $p = .003$, $r = .89$, large effect size). The results showed that a contextualized and personalized model achieved a significant large increase in discriminatory power over non-personalized and non-contextualized models.

A second Friedman test showed that all neural network models differed significantly in their correlation with learner judgments, $\chi^2_F(3) = 14.88$, $p = .002$. Post hoc one-tailed Wilcoxon signed rank tests showed that the BiLSTM-average model did not achieve a significantly higher $\phi$ than the FFNN-average model ($W = 9$, $p = 1$) and the baseline model ($W = 1$, $p = 1$). By contrast, the BiLSTM-learner model achieved a significantly higher $\phi$ than the BiLSTM-average model ($W = 55$, $p = .003$, $r = .89$, large effect size), but did not significantly outperform the FFNN-average model ($W = 41$, $p = .19$) nor

**Table 7.5**

*Model Ablation on Tenfold Cross-Validation*

| | | $D$ | | $\phi$ | |
|---|---|---|---|---|---|
| | $N$ | $M \pm SD$ | $Mdn$ | $M \pm SD$ | $Mdn$ |
| BiLSTM-learner | 10 | 0.64 ± 0.02 | 0.64 | 0.36 ± 0.01 | 0.36 |
| — native language | 10 | 0.65 ± 0.01 | 0.65 | 0.36 ± 0.01 | 0.36 |
| — subject identifier | 10 | 0.62 ± 0.01 | 0.62 | 0.34 ± 0.01 | 0.34 |
| — proficiency level | 10 | 0.59 ± 0.01 | 0.59 | 0.35 ± 0.01 | 0.35 |
| — character CNN | 10 | 0.64 ± 0.01 | 0.63 | 0.35 ± 0.01 | 0.35 |
| — word embedding | 10 | 0.43 ± 0.02 | 0.42 | 0.23 ± 0.01 | 0.23 |

the baseline model ($W = 41$, $p = .19$). The results therefore showed that personalizing a contextualized model achieved a significantly larger correlation with learner judgments, but its performance was not significantly different from non-contextualized models.

An ablation study was performed to investigate whether there was a significant decrease in performance when parts of the contextualized and personalized network (viz., BiLSTM-learner) were removed. Table 7.5 shows the average cross-validated performance scores after removing one specific part of the model. The ablation of word embeddings had a significant large effect ($r = .89$) on both correlation and discriminatory power ($W = 55$, $p = .005$). When no pre-trained word vectors were included in the model, the power of discrimination dropped below 50%. The ablation of character embeddings, on the other hand, had a significant large effect ($r = .89$) on correlation ($W = 55$, $p = .005$), but an insignificant effect on discriminatory power ($W = 42$, $p = .16$). The ablation of the subject's unique identifier had a significant large effect on both discriminatory power ($W = 54$, $p = .006$, $r = .89$) and correlation ($W = 55$, $p = .005$, $r = .85$). The ablation of the subject's proficiency level had a large ($r = .89$) effect on discriminatory power ($W = 55$, $p = .005$), but an insignificant effect on correlation ($W = 37$, $p = .38$). Finally, the ablation of the subject's native language had an insignificant effect on both discriminatory power ($W = 14$, $p = .92$) and correlation ($W = 8$, $p = .98$).

In sum, the results showed that a contextualization and personalization of predictions lead to a significantly better discrimination between difficult and non-difficult words. However, only a personalization of predictions lead to a significantly better correlation with learner judgments. Furthermore, the results showed that all components except the learner's L1 made a significant contribution to the predictive power of the model. The word embeddings as well as the subject's identifier and proficiency level contributed to a better discrimination between difficult and non-difficult words. The word and character embeddings as well as the subject's identifier contributed to a better correlation with learner judgments.

*Predictive Power per Subject*

A final series of analyses was conducted to examine whether the contextualized and personalized model also achieved a better performance for each individual learner. Instead of computing a general score per each cross-validation fold (cf. *supra*), a performance score was computed for each learner per each cross-validation fold. Three linear mixed-effects models were fitted with *lme4* (Bates et al., 2015) on (a) the average probability of difficulty predicted on difficult words, (b) the discriminatory power $D$ between difficult and non-difficult words, and (c) the correlation $\phi$ with learner judgments. Each mixed-effects model included the models as fixed factors (with the baseline model as the intercept) and the subjects and the cross-validation folds as random factors. Because the $D$ values and the average predicted probabilities followed a heavy-tailed distribution, the Gamma link function was used instead. Post hoc pairwise Tukey comparisons were performed on the estimated marginal means with the *emmeans* package (Lenth, 2020). Cohen's $d$ was used to measure the size of the effect. All three models were fitted on a total of 2,240 observations, corresponding to four repeated measures (i.e., four models) with 560 scores each (i.e., 10 folds $\times$ 56 learners).

Table 7.6 shows the results of a Gamma generalized linear mixed-effects analysis on the average predicted probability of difficulty on difficult words. The personalized and contextualized model (BiLSTM-learner) achieved a significant but small increase in certainty (+4.3%) over a non-personalized and contextualized model (BiLSTM-average) as well as a significant but small in-

**Table 7.6**

*Mixed-Effects Analysis of the Average Predicted Probability on Difficult Words*

| **Fixed Effects** | Context | Personal | Beta | *SE* | $\chi^2_{LRT}(1)$ | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | ✘ | ✘ | 0.791 | 0.017 | 2,184 | <.001 |
| BiLSTM-learner | ✔ | ✔ | 0.143 | 0.004 | 1,101 | <.001 |
| BiLSTM-average | ✔ | ✘ | 0.100 | 0.004 | 568 | <.001 |
| FFNN-average | ✘ | ✘ | 0.085 | 0.004 | 418 | <.001 |

| **Random Effects** | Subject | Fold:Subject | Residual |
|---|---|---|---|
| Variance | 0.002 | 0.004 | 0.012 |

| **Tukey Comparisons** | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 1 | BiLSTM-learner | 0.934 | 0.344 | 0.464 | 1.139 |
| 2 | BiLSTM-average | +0.043 *** | 0.891 | 0.120 | 0.795 |
| 3 | FFNN-average | +0.058 *** | +0.015 .004 | 0.876 | 0.675 |
| 4 | De Hertog 2018 | +0.143 *** | +0.100 *** | +0.085 *** | 0.791 |

diagonal = estimated average probability of difficulty on difficult words lower triangle = estimated marginal means $p$-value upper triangle = Cohen's *d* effect size *** $p < .001$

crease in certainty (+5.8%) over a non-personalized and non-contextualized model (FFNN-average). The contextualized model (BiLSTM-average) achieved a significant but weak increase in certainty (+1.5%) over a non-contextualized model (FFNN-average). Furthermore, all three models achieved a medium to large increase in certainty over the baseline model. The personalized and contextualized model (BiLSTM-learner) achieved a statistically significant, large increase in certainty (+14.3%) over the baseline model, whereas the contextualized model (BiLSTM-average) and non-contextualized model (FFNN-average) achieved a significant, medium increase in certainty (+10% and +8.5%, respectively) over the baseline model. Consequently, the pairwise mixed-effects comparisons showed results that were similar to those obtained on tenfold cross-validation, although the effect size estimates were more conservative.

Table 7.7 shows the results of a Gamma generalized linear mixed-effects analysis on discriminatory power (*D*) between difficult and non-difficult words. The personalized and contextualized model (BiLSTM-learner) achieved a significant but small increase in *D* (+3.3%) over a non-personalized and contextual-

**Table 7.7**

*Mixed-Effects Analysis of Model Performance on Discriminatory Power*

| **Fixed Effects** | Context | Personal | Beta | *SE* | $\chi^2_{LRT}(1)$ | *p* |
|---|---|---|---|---|---|---|
| (Intercept) | ✘ | ✘ | 0.661 | 0.015 | 1,845 | <.001 |
| BiLSTM-learner | ✔ | ✔ | 0.075 | 0.004 | 417 | <.001 |
| BiLSTM-average | ✔ | ✘ | 0.042 | 0.004 | 140 | <.001 |
| FFNN-average | ✘ | ✘ | 0.038 | 0.004 | 113 | <.001 |

| **Random Effects** | Subject | Fold:Subject | Residual |
|---|---|---|---|
| Variance | 0.001 | 0.004 | 0.016 |

| **Tukey Comparisons** | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 1 | BiLSTM-learner | 0.736 | 0.257 | 0.289 | 0.588 |
| 2 | BiLSTM-average | +0.033 *** | 0.704 | 0.032 | 0.331 |
| 3 | FFNN-average | +0.037 *** | +0.004 .68 | 0.699 | 0.299 |
| 4 | De Hertog 2018 | +0.075 *** | +0.042 *** | +0.038 *** | 0.661 |

diagonal = estimates of *D* lower triangle = estimated marginal means *p*-value
upper triangle = Cohen's *d* effect size *** $p < .001$

ized model (BiLSTM-average) as well as a significant but small increase (+3.7%) over a non-personalized and non-contextualized model (FFNN-average). The contextualized model (BiLSTM-average), however, did not achieve a significant increase in *D* (+0.4%) over a non-contextualized model (FFNN-average). As compared to the baseline model, all three models achieved a small to medium increase in discriminatory power. The personalized and contextualized model (BiLSTM-learner) achieved a statistically significant, medium increase in *D* (+7.5%) over the baseline model, whereas the contextualized model (BiLSTM-average) and non-contextualized model (FFNN-average) achieved a significant, small increase in *D* (+4.2% and +3.8%, respectively) over the baseline model. Consequently, the pairwise mixed-effects comparisons showed results that were similar to the results of the pairwise comparisons performed on tenfold cross-validation, but only regarding personalization. The difference in discriminatory power between the contextualized (BiLSTM-average) and non-contextualized (FFNN-average) models was no longer significant when comparing performance between subjects.

**Table 7.8**

*Mixed-Effects Analysis of Model Performance on Correlation*

| **Fixed Effects** | Context | Personal | Beta | *SE* | $\chi^2_{LRT}(1)$ | $p$ |
|---|---|---|---|---|---|---|
| (Intercept) | ✘ | ✘ | 0.347 | 0.011 | 1,011 | <.001 |
| BiLSTM-learner | ✔ | ✔ | 0.007 | 0.003 | 5 | .026 |
| BiLSTM-average | ✔ | ✘ | −0.025 | 0.003 | 66 | <.001 |
| FFNN-average | ✘ | ✘ | −0.015 | 0.003 | 23 | <.001 |

| **Random Effects** | Subject | Fold:Subject | Residual |
|---|---|---|---|
| Variance | 0.006 | 0.004 | 0.003 |

| **Tukey Comparisons** | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 BiLSTM-learner | 0.354 | 0.617 | 0.418 | 0.133 |
| 2 BiLSTM-average | +0.032 *** | 0.321 | −0.199 | −0.484 |
| 3 FFNN-average | +0.022 *** | −0.010 .005 | 0.332 | −0.285 |
| 4 De Hertog 2018 | +0.007 .12 | −0.025 *** | −0.015 *** | 0.347 |

diagonal = estimates of $\phi$ lower triangle = estimated marginal means $p$-value
upper triangle = Cohen's $d$ effect size *** $p < .001$

Table 7.8 shows the results of a Gaussian linear mixed-effects analysis on correlation ($\phi$) with learner judgments. The personalized and contextualized model (BiLSTM-learner) achieved a significant, medium increase in $\phi$ (+.032) over a non-personalized and contextualized model (BiLSTM-average) as well as a significant, small increase (+.022) over a non-personalized and non-contextualized model (FFNN-average). The contextualized model (BiLSTM-average), however, achieved a significant, small decrease in $\phi$ (−.010) over a non-contextualized model (FFNN-average). Moreover, all three models achieved either a non-significant increase or a significant decrease in correlation over the baseline model. The personalized and contextualized model (BiLSTM-learner) achieved an insignificant increase in $\phi$ (+.007) over the baseline model, whereas the contextualized model (BiLSTM-average) and non-contextualized model (FFNN-average) achieved a significant, small decrease in $\phi$ (−.025 and .015, respectively) over the baseline model. Consequently, the results of the pairwise mixed-effects comparisons showed that personalizing a contextualized model achieved a significantly larger correlation with

learner judgments, whereas a contextualized model did not outperform a non-contextualized model on correlation.

As a final note, it should be observed that there was no sizable variance between learners and between cross-validation folds. As can be seen in Tables 7.6 to 7.8, the random effects variances remained small on all three metrics, albeit with some minor differences. More specifically, there appeared to be a larger variance in discriminatory power between folds than between learners. Conversely, there appeared to be a larger variance in correlation between learners. More notable differences were, however, observed when comparing model performance per trial and per proficiency level. Figures 7.10 and 7.11 show, on the one hand, that a comparable performance was achieved on both trials. In other words, there were no considerable differences in performance between a trial with more data but fewer learners (viz., Trial 1) and a trial with more learners but less data (viz., Trial 2). On the other hand, Figures 7.10 and 7.11 also show that model performance differed more considerably between proficiency levels and on discriminatory power in particular. Figure 7.10 shows that all models achieved a considerably lower degree of discriminatory power on the lowest proficiency level (viz., A2). By contrast, Figure 7.11 shows that all models achieved a similar degree of correlation on all proficiency levels, except for the highest proficiency level (viz., C1). In conclusion, although all models achieved a similar performance between subjects and between trials, model performance differed more considerably between proficiency levels.

*Fine-Tuning a Pre-Trained BERT Transformer for French*

Because no significant performance increase was observed for the contextualized network, a follow-up analysis was performed to determine whether a better performance could be achieved by fine-tuning a pre-trained transformer model. The analysis focused on the CamemBERT model (Martin et al., 2019), which included state-of-the-art pre-trained contextualized word representations for French. This pre-trained model (viz., CamembertForTokenClassification) was further fine-tuned on the data and evaluated with the same cross-validation setup and loss weights. Table 7.9 shows that fine-tuning a pre-trained transformer resulted in a considerable discriminatory power ($D$ = .55). However, this model could not outperform the other three neural

**Figure 7.10**

*Discriminatory Power per Trial and per Proficiency Level*

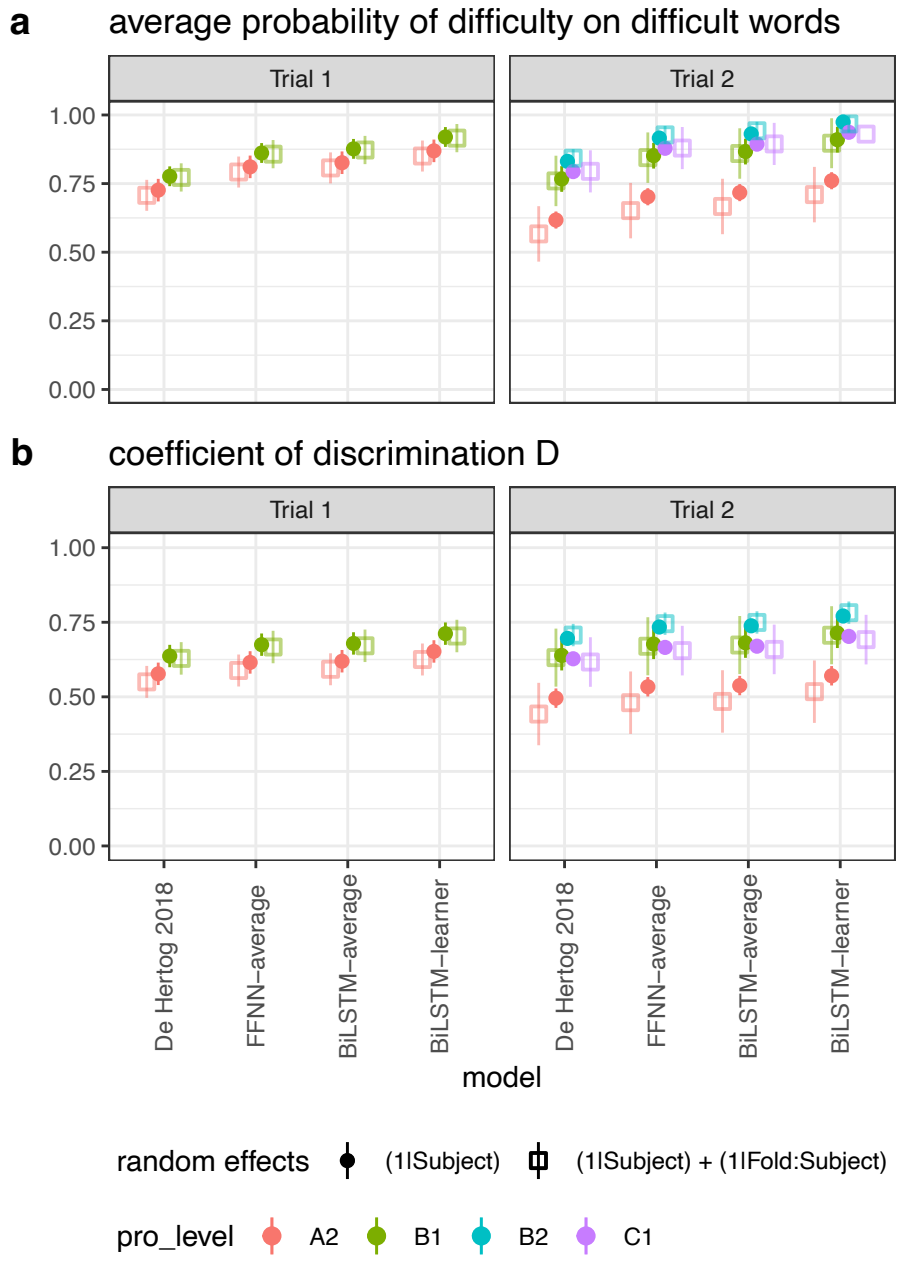
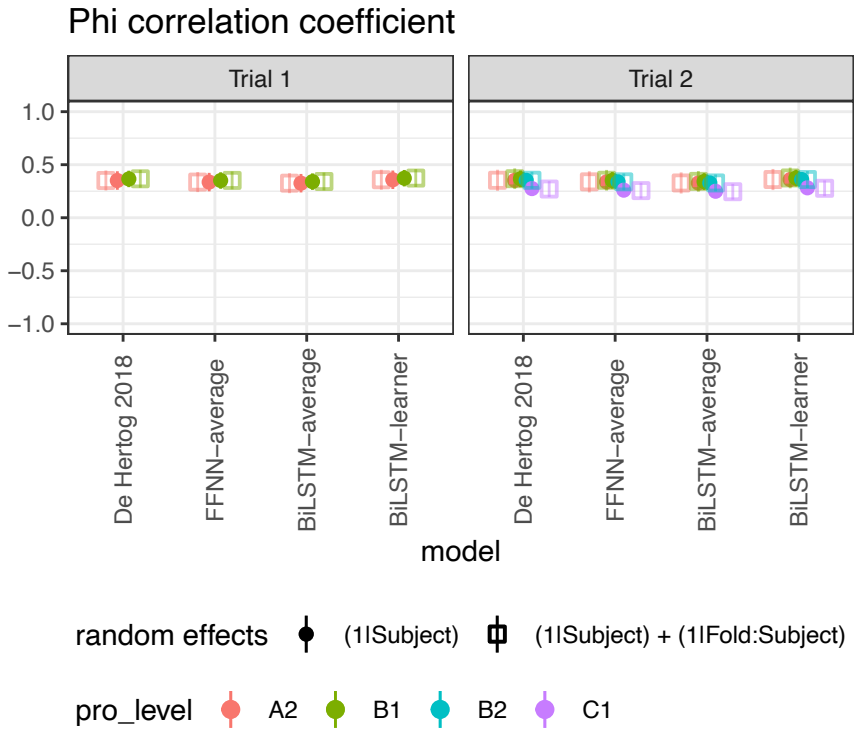


**Figure 7.11**
*Correlation per Trial and per Proficiency Level*




network models. This finding suggested that the best predictive power could not be obtained from contextualized representations alone.

*Summary of Results*

Table 7.10 gives a summary of the results obtained with respect to the three hypotheses of this study. The first hypothesis was that a contextualized model would achieve a significantly better performance than a non-contextualized model. The results did not, however, support this hypothesis. Although a contextualized model could discriminate better between difficult and non-difficult words on tenfold cross-validation, a contextualized model did not significantly outperform a non-contextualized model on correlation. Furthermore, the contextualized model no longer outperformed the non-contextualized model when performance was compared between learners.

**Table 7.9**

*Performance on Tenfold Cross-Validation of a Fine-Tuned BERT Transformer*

| | C | P | *N* | *D* | | *ϕ* | |
|---|---|---|---|---|---|---|---|
| | | | | *M* ± *SD* | *Mdn* | *M* ± *SD* | *Mdn* |
| MODELS | | | | | | | |
| BiLSTM-learner | ✔ | ✔ | 10 | 0.64 ± 0.02 | 0.64 | 0.36 ± 0.01 | 0.36 |
| BiLSTM-average | ✔ | ✘ | 10 | 0.59 ± 0.01 | 0.60 | 0.34 ± 0.01 | 0.34 |
| FFNN-average | ✘ | ✘ | 10 | 0.59 ± 0.02 | 0.59 | 0.35 ± 0.01 | 0.35 |
| TRANSFORMERS | | | | | | | |
| CamemBERT | ✔ | ✘ | 10 | 0.55 ± 0.01 | 0.55 | 0.29 ± 0.01 | 0.30 |

[C] contextualized model [P] personalized model

**Table 7.10**

*Summary of Results*

| | Cross-Validation | | Between Learners | |
|---|---|---|---|---|
| Hypothesis | *D* | *ϕ* | *D* | *ϕ* |
| Contextualization > None | Yes | No | No | No |
| Personalization > None | Yes | Yes | Yes | Yes |
| BiLSTM-learner > Baseline | Yes | No | Yes | No |

The second hypothesis was that a personalized model would achieve a significantly better performance than a non-personalized model. The results provided unequivocal support for this hypothesis. A personalized model outperformed a non-personalized model, achieving a better correlation with learner judgments and a higher discrimination between difficult and non-difficult words. The individual characteristics that contributed significantly to this performance were the learners' proficiency level and their unique identifier. The inclusion of the learner's proficiency level lead to a better discrimination between difficult and non-difficult words, while integrating the individual learner in the model lead to a medium increase in correlation with learner judgments.

The third and final hypothesis was that a personalized and contextualized model would achieve a better performance than the baseline model. The results showed that the BiLSTM-learner model achieved a medium increase in discriminatory power over the baseline model. In fact, all three neural network models significantly outperformed the baseline model on discriminatory power, even a less complex, single-layer feedforward network (viz., FFNN-average). At the same time, none of the models significantly outperformed the baseline model on correlation with learner judgments. There were two possible reasons for this: either the low degree of discriminatory power was caused by the more complex structure of the baseline model (i.e., a three-layer neural network with random dropout) or it was caused by the fact that De Hertog and Tack's (2018)'s implementation attributed more weight to the auxiliary training of the character CNN than to the optimization of the fully-connected layers. The results of the ablation study might give support for the latter. While ablating the character CNN did not have a significant impact on discriminatory power, it did, however, lead to a significant decrease in correlation with learner judgments. It was therefore concluded that both the BiLSTM-learner model and the baseline model had a comparable correlation with learner judgments, although the former model was optimized to classify difficult words with a much higher probability of difficulty, whereas the latter model was optimized to classify difficult words within closer range of the default decision threshold of 50%.

## 7.3 CONCLUSION

The study gives support for the use of deep learning to predict when a learner will perceive lexical difficulty in reading French L2. Even with relatively few training data, an artificial neural network can learn to discriminate fairly well between difficult and non-difficult words, achieving 64% to 74% of discriminatory power on average as well as a moderate positive correlation with learner judgments ($\phi$ = .36). The strongest contribution to this performance comes from the inclusion of pre-trained distributional word vectors. These word embeddings capture most of the features of lexical complexity that explain why a learner perceives a word as difficult, such as word surprisal. The inclusion

of a character CNN, in contrast, does not significantly contribute to a better distinction between difficult and non-difficult words. The results therefore corroborate the conclusions of the previous chapter (Chapter 6), showing that perceived lexical difficulty in L2 reading mainly relates to word use rather than word form.

The study also provides answers to the hypothesis that better predictions can be made by taking into account the individual learner as well as the context surrounding the word. On the one hand, the study corroborates the need for adopting a personalized approach to the prediction of difficulty. By estimating learner-specific weights, an artificial neural network makes predictions that correlate significantly better with how a learner perceives difficulty. Moreover, by including the learner's proficiency level, the model distinguishes difficult from non-difficult words with a higher certainty. On the other hand, the results remain inconclusive as regards the need for adopting a contextualized approach. Although a contextualized model can discriminate significantly better between difficult and non-difficult words on the whole, the benefit of making contextualized predictions is no longer significant when model performance is compared between learners.

A plausible explanation for this inconclusiveness can be found in the specific nature of the data. In the measure used in this study, the difficulty of a word is mostly explained by isolated word surprisal, whereas contextual word surprisal only achieves a weak explanatory effect (see Chapter 6). In other words, learners have a tendency to notice difficulty on words that occur rarely in the target language, regardless of the context in which the word occurs. An important implication of the findings therefore relates to the types of measures that are used to develop a predictive model of difficulty (e.g., see Paetzold & Specia, 2016a; Yimam et al., 2018). If a predictive model is trained on a measure where subjects are asked to identify difficult words in a text and if this measure is mostly explained by isolated word surprisal, this predictive model will not be able to account for possible difficulties pertaining to word-to-context integration.

Nevertheless, because a contextualized BiLSTM learns more complex patterns than a non-contextualized FFNN, it could also be that the former generalizes its predictions to words that may be difficult in reality but to which the

learner is oblivious (i.e., perceived as non-difficult). For illustration purposes, the reader is referred to Figures 7.4 and 7.5 on page 305 and on page 306. As opposed to the FFNN network, the BiLSTM network appears to identify *puits de rocher* as a contiguous unit seeing that the token *de* is given a higher probability of difficulty. Similarly, the difficulty predicted for *pas encore* seems linked to the word it modifies (i.e., *noyé*), even though the learner did not perceive this adverb as difficult. Likewise, the network attributes a higher probability of difficulty to two immediate syntactic dependents of the predicate *noyé*, namely its passive nominal subject *village* and its oblique argument *neige*. Although the classifications of difficulty on *pas encore* and *neige* are incorrect with respect to the learner's perceptions, they could, however, be a correct projection of how the difficulty of *noyé* affects its syntactic modifiers and dependents. Consequently, it is logical that these seemingly incorrect generalizations lead to a lower correlation with the learner's perceptions. However, this potential advantage of a contextualized model cannot be assessed by a mere binary classification metric. Therefore, a more detailed experiment is required to further substantiate this tentative explanation.

Another important implication of the study relates to the metrics that are used to assess model performance. When determining how well a model can predict difficulty, it is important to use metrics that not only demonstrate the predictive power of the model, but also remain robust towards a variability in the extent of difficulty experienced across learners. More specifically, a performance metric should not attribute a higher score simply because the gold standard includes a larger proportion of difficult words. If this were the case and if such a metric were used to compare model performance for varying degrees of difficulty, there would be no way of knowing whether the performance score truly reflects an improvement in predictive power or simply an increase in perceived difficulty. For this reason, it appears that the use of non-sensitive performance metrics, such as Pearson's (1904) $\phi$ correlation coefficient and Tjur's (2009) coefficient of discrimination, should be preferred over previously used metrics, such as the $F_1$-score and the G-score, which are sensitive to varying degrees of difficulty.

There are, however, three elements that still remain unresolved and which require further study. Firstly, considering that all correlation coefficients

remain relatively unvarying between learners and even between proficiency levels, it is possible that the $\phi$ coefficient is too strict a measure of performance. Moreover, the outcome of the binomial correlation coefficient largely depends on the decision threshold that is used to transform the predicted probability of difficulty into a binomial class. Because this threshold is set to 50% by default, it may be interesting to further examine (a) whether it is more applicable to find an optimal decision threshold (e.g., see Bingel et al., 2016) and (b) how this optimal threshold impacts the results obtained on binomial correlation. Secondly, it may also be required to further investigate how the character CNN contributes to predicting other measures of difficulty. In particular, it may be interesting to examine the contribution of the character CNN to the prediction of lexical difficulty for other target readers for whom the complexity of the word form is known to be more critical, such as dyslexic readers. Finally, and more crucially, it is necessary to further investigate the use of recurrent neural networks to learn word-to-context integration difficulties from other – either direct or indirect – measures of lexical difficulty in L2 reading.

## 7.A APPENDIX

This appendix provides some details on how the deep learning analysis was conducted. The appendix is based on the *Machine Learning Reproducibility Checklist*[76] and provides the following information: the computing infrastructure (Section 7.A.1), the dependencies (Section 7.A.2), the parameters and hyperparameters (Section 7.A.3), the cross-validation setup (Section 7.A.4), and the evaluation metrics (Section 7.A.5).

### 7.A.1 *Computing Infrastructure*

The same computing infrastructure was used for all analyses. In addition, an NVIDIA GeForce GPU was used for fine-tuning the pre-trained CamemBERT model (Martin et al., 2019), which was available from the *transformers* library (Wolf et al., 2020).

```
MacBook Pro (Retina, 15-inch, Mid 2015)

macOS Catalina 10.15.4 (19E287)
Processor: 2,2 GHz Intel Core i7 quad-core
Memory: 16 GB 1600 MHz DDR3
Graphics: Intel Iris Pro 1536 MB
```

```
NVIDIA Driver

Ubuntu 18.04.5 LTS
Driver Version: 460.32.03
CUDA Version: 11.2
GPU: GeForce GT 740M
```

76 See also: https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (Pineau et al., 2020) and https://2020.emnlp.org/call-for-papers#new-reproducibility-criteria.

### 7.A.2 *Dependencies*

The following Python libraries were used: *gensim* (Řehůřek & Sojka, 2010), *numpy* (Harris et al., 2020), *pandas* (McKinney, 2010), *scikit-learn* (Pedregosa et al., 2011), *tensorflow* (Abadi et al., 2015), and *transformers* (Wolf et al., 2020).

```
Python 3.6.2

gensim==3.8.1
numpy==1.17.3
pandas==0.25.2
scikit-learn==0.21.3
tensorflow==2.1.0
tensorflow-gpu==2.1.0
transformers==3.1.0
```

### 7.A.3 *Parameters and Hyperparameters*

All model parameters were initialized with the default initializer. A fixed seed (`seed=0`) was used for initializing the weights. A summary of the deep learning architectures and their respective parameters is given below.

```
Summary of the DeHertog2018 Architecture

Layer (type)                Output Shape   Param #     Connected to
==================================================================
char_inp (InputLayer)       [(?, 27)]      0
char_emb (Embedding)        (?, 27, 16)    102544      char_inp[0][0]
char_conv1 (Conv1D)         (?, 27, 4)     196         char_emb[0][0]
char_max1 (MaxPooling1D)    (?, 13, 4)     0           char_conv1[0][0]
char_conv2 (Conv1D)         (?, 13, 4)     52          char_max1[0][0]
char_max2 (MaxPooling1D)    (?, 6, 4)      0           char_conv2[0][0]
char_lstm (LSTM)            (?, 64)        17664       char_max2[0][0]
------------------------------------------------------------------
```

```
word_inp (InputLayer)      [(?, 1)]       0
fasttext (Embedding)       (?, 1, 300)    600077400   word_inp[0][0]
word_lstm (LSTM)           (?, 64)        93440       fasttext[0][0]
----------------------------------------------------------------------
subj_inp (InputLayer)      [(?, 56)]      0
cefr_inp (InputLayer)      [(?, 4)]       0
lang_inp (InputLayer)      [(?, 4)]       0
----------------------------------------------------------------------
concat_1 (Concatenate)     (?, 192)       0           char_lstm[0][0]
                                                      word_lstm[0][0]
                                                      subj_inp[0][0]
                                                      cefr_inp[0][0]
                                                      lang_inp[0][0]
----------------------------------------------------------------------
dense_1 (Dense)            (?, 32)        6176        concat_1[0][0]
dropout_1 (Dropout)        (?, 32)        0           dense_1[0][0]
dense_2 (Dense)            (?, 32)        1056        dropout_1[0][0]
dropout_2 (Dropout)        (?, 32)        0           dense_2[0][0]
dense_3 (Dense)            (?, 32)        1056        dropout_2[0][0]
----------------------------------------------------------------------
concat_2 (Concatenate)     (?, 96)        0           dense_3[0][0]
                                                      subj_inp[0][0]
                                                      cefr_inp[0][0]
                                                      lang_inp[0][0]
----------------------------------------------------------------------
main_out (Dense)           (?, 1)         97          concat_2[0][0]
aux_out (Dense)            (?, 1)         65          char_lstm[0][0]
======================================================================
Total params: 600,299,746
Trainable params: 222,346
Non-trainable params: 600,077,400
```

Summary of the FFNN Architecture

```
Layer (type)               Output Shape   Param #     Connected to
======================================================================
char_inp (InputLayer)      [(?, 27)]      0
char_emb (Embedding)       (?, 27, 16)    102544      char_inp[0][0]
char_conv1 (Conv1D)        (?, 27, 4)     196         char_emb[0][0]
```

```
char_max1 (MaxPooling1D)   (?, 13, 4)     0           char_conv1[0][0]
char_conv2 (Conv1D)        (?, 13, 4)     52          char_max1[0][0]
char_max2 (MaxPooling1D)   (?, 6, 4)      0           char_conv2[0][0]
char_flat (Flatten)        (?, 24)        0           char_max2[0][0]
__________________________________________________________________________
word_inp (InputLayer)      [(?, 1)]       0
word_emb (Embedding)       (?, 1, 300)    600077400   word_inp[0][0]
fasttext (Flatten)         (?, 300)       0           word_emb[0][0]
__________________________________________________________________________
subj_inp (InputLayer)      [(?, 56)]      0
cefr_inp (InputLayer)      [(?, 4)]       0
lang_inp (InputLayer)      [(?, 4)]       0
__________________________________________________________________________
concat_1 (Concatenate)     (?, 388)       0           char_flat[0][0]
                                                      fasttext[0][0]
                                                      subj_inp[0][0]
                                                      cefr_inp[0][0]
                                                      lang_inp[0][0]
__________________________________________________________________________
dense (Dense)              (?, 32)        12448       concat_1[0][0]
__________________________________________________________________________
concat_2 (Concatenate)     (?, 96)        0           dense[0][0]
                                                      subj_inp[0][0]
                                                      cefr_inp[0][0]
                                                      lang_inp[0][0]
__________________________________________________________________________
main_out (Dense)           (?, 1)         97          concat_2[0][0]
aux_out (Dense)            (?, 1)         25          char_flat[0][0]
==========================================================================
Total params: 600,192,762
Trainable params: 115,362
Non-trainable params: 600,077,400
```

Summary of the BiLSTM Architecture

```
Layer (type)             Output Shape     Param #    Connected to
==========================================================================
char_inp (InputLayer)    [(?, 94, 27)]    0
char_emb (TimeDist.)     (?, 94, 27, 16)  102544     char_inp[0][0]
```

```
char_conv1 (TimeDist.)   (?, 94, 27, 4)   196        char_emb[0][0]
char_max1 (TimeDist.)    (?, 94, 13, 4)   0          char_conv1[0][0]
char_conv2 (TimeDist.)   (?, 94, 13, 4)   52         char_max1[0][0]
char_max2 (TimeDist.)    (?, 94, 6, 4)    0          char_conv2[0][0]
char_flat (TimeDist.)    (?, 94, 24)      0          char_max2[0][0]
----------------------------------------------------------------------
word_inp (InputLayer)    [(?, 94)]        0
fasttext (Embedding)     (?, 94, 300)     600077400  word_inp[0][0]
----------------------------------------------------------------------
subj_inp (InputLayer)    [(?, 94, 56)]    0
cefr_inp (InputLayer)    [(?, 94, 4)]     0
lang_inp (InputLayer)    [(?, 94, 4)]     0
----------------------------------------------------------------------
concat_1 (Concatenate)   (?, 94, 388)     0          char_flat[0][0]
                                                    fasttext[0][0]
                                                    subj_inp[0][0]
                                                    cefr_inp[0][0]
                                                    lang_inp[0][0]
----------------------------------------------------------------------
bilstm (Bidirectional)   (?, 94, 128)     231936     concat_1[0][0]
----------------------------------------------------------------------
concat_2 (Concatenate)   (?, 94, 192)     0          bilstm[0][0]
                                                    subj_inp[0][0]
                                                    cefr_inp[0][0]
                                                    lang_inp[0][0]
----------------------------------------------------------------------
main_out (TimeDist.)     (?, 94, 1)       193        concat_2[0][0]
aux_out (TimeDist.)      (?, 94, 1)       25         char_flat[0][0]
======================================================================
Total params: 600,412,346
Trainable params: 334,946
Non-trainable params: 600,077,400
```

Unless explicitly noted otherwise (previously or below), all functions were called with the parameters set to the default value. For each cross-validation run, an array of sample weights was constructed from the balanced class weights (cf. Section 7.A.4) that were computed for the array of targets in the training data, which was a 2-dimensional array of size (sentences,

max_padding). It was ensured that the sample weights of all masked tokens were set to zero.

Hyperparameters

| Function | Parameter | DeHertog2018 | FFNN/BiLSTM |
|---|---|---|---|
| compile | loss | binary_crossentropy | binary_crossentropy |
| | loss_weights | [0.5, 1.0] | [1, 0.25] |
| | optimizer | Adam | Adam |
| EarlyStopping | monitor | val_loss | val_loss |
| fit | epochs | 50 | 50 |
| | sample_weights | np.ndarray | np.ndarray |
| | shuffle | False | False |
| | validation_split | 0.1 | 0.1 |

### 7.A.4 *Cross-Validation Setup*

The data set included the collection of texts that were read by each individual learner and counted a total of 278,756 tokens. A mask was set to all tokens that were non-alphanumeric (e.g., punctuation) and which did not receive a label during data collection (e.g., numbers). As a result, the data counted a total of 261,582 tokens that were not masked, which did not receive a zero sample weight, and which were therefore considered for training and testing. The masked tokens (e.g., punctuation, symbols, numbers, etc.) were not removed because they were needed for correctly representing the sentence context. In order to ignore these masked tokens during training, their input vectors were set to zero and they received a zero sample weight when computing the loss function. All tokens were grouped into their respective sentence contexts, which corresponded to a total of 20,303 sentences. All inputs and outputs were padded (with zero pre-padding) to the maximum sentence length.

Cross-Validation Runs

| Run | Tokens (train/val/test) | Sentences (idem) | Balanced class weights |
|---|---|---|---|
| 01 | 211159/23463/26960 | 16401/1823/2079 | {0: 0.50, 1: 74.3} |
| 02 | 211693/23522/26367 | 16408/1824/2071 | {0: 0.50, 1: 73.7} |
| 03 | 211774/23531/26277 | 16417/1825/2061 | {0: 0.50, 1: 73.7} |
| 04 | 212112/23568/25902 | 16429/1826/2048 | {0: 0.50, 1: 73.6} |
| 05 | 211404/23490/26688 | 16440/1827/2036 | {0: 0.50, 1: 73.7} |
| 06 | 212150/23573/25859 | 16449/1828/2026 | {0: 0.50, 1: 73.6} |
| 07 | 212528/23615/25439 | 16458/1829/2016 | {0: 0.50, 1: 73.8} |
| 08 | 212049/23561/25972 | 16472/1831/2000 | {0: 0.50, 1: 74.1} |
| 09 | 212319/23591/25672 | 16481/1832/1990 | {0: 0.50, 1: 73.9} |
| 10 | 211622/23514/26446 | 16494/1833/1976 | {0: 0.50, 1: 74.0} |

### 7.A.5 *Evaluation Metrics*

The $\phi$ coefficient was computed with the `matthews_corrcoef` function from *scikit-learn* (Pedregosa et al., 2011). The Python code used for computing Tjur's (2009) coefficient of discrimination $D$ is given below.

Python Code for Tjur's Coefficient of Discrimination

```
import numpy as np


def tjur_D(y_true, y_prob):
    """Compute Tjur's coefficient of discrimination.

    Tue Tjur. 2009. Coefficients of Determination in
    Logistic Regression Models-A New Proposal: The
    Coefficient of Discrimination. The American
    Statistician, 63(4):366-372.

    :param y_true: np.array of integers (binary classes)
    :param y_prob: np.array of floats (probabilities on true class)
    :return: float ranging from 0 to 1
    """
```

```
    # get mask of positive class
    y_pos = (y_true == 1).flatten()
    # split predicted probabilities for each class
    y_1 = y_prob[y_pos]
    y_0 = y_prob[np.invert(y_pos)]
    # compute mean predicted probabilities on y=1
    y_1_mean = np.mean(y_1)
    y_0_mean = np.mean(y_0)
    # return absolute difference between the two
    return abs(y_1_mean-y_0_mean)
```

Cross-Validation Scores for the D Coefficient

| Run | BiLSTM-learner | BiLSTM-average | FFNN-average | DeHertog2018 | CamemBERT |
|---|---|---|---|---|---|
| 01 | 0.6295856 | 0.5832078 | 0.5774344 | 0.5197178 | 0.5422198 |
| 02 | 0.6355799 | 0.599928 | 0.6026593 | 0.5809717 | 0.5522327 |
| 03 | 0.6745878 | 0.5994418 | 0.5992727 | 0.551846 | 0.5591149 |
| 04 | 0.6377962 | 0.6012275 | 0.5889308 | 0.5571004 | 0.5550294 |
| 05 | 0.6323688 | 0.5965163 | 0.5976764 | 0.5684803 | 0.5603604 |
| 06 | 0.6570263 | 0.5788949 | 0.5759474 | 0.5363099 | 0.5412584 |
| 07 | 0.6466628 | 0.6040243 | 0.596638 | 0.5633804 | 0.557315 |
| 08 | 0.6224575 | 0.5819291 | 0.5767351 | 0.5485028 | 0.5477578 |
| 09 | 0.6501746 | 0.5939686 | 0.5849083 | 0.5035037 | 0.5367005 |
| 10 | 0.63884 | 0.5739082 | 0.55195 | 0.5114896 | 0.5391348 |

Cross-Validation Scores for the Phi Coefficient

| Run | BiLSTM-learner | BiLSTM-average | FFNN-average | DeHertog2018 | CamemBERT |
|---|---|---|---|---|---|
| 01 | 0.350819 | 0.3406713 | 0.3341135 | 0.3608887 | 0.2884533 |
| 02 | 0.3531413 | 0.3523438 | 0.3787997 | 0.3716971 | 0.2985075 |
| 03 | 0.3619861 | 0.343781 | 0.3491474 | 0.3521966 | 0.2970902 |
| 04 | 0.3667619 | 0.3480832 | 0.3417608 | 0.3540494 | 0.2964936 |
| 05 | 0.349773 | 0.3394244 | 0.3555756 | 0.3372511 | 0.2976278 |
| 06 | 0.3639473 | 0.3231396 | 0.3309634 | 0.3382102 | 0.2776579 |
| 07 | 0.3605886 | 0.3461702 | 0.3645514 | 0.3552054 | 0.2971264 |

```
08   0.3473357      0.3306161      0.3440983    0.3370374    0.3021565
09   0.3656229      0.3481516      0.3512324    0.3622646    0.2907652
10   0.3434424      0.3287618      0.3347346    0.336666     0.2850509
```

# CONCLUSION

Nil tam difficile'st quin quærendo investigari possiet.

— Terentius, *Heautòn Timoroúmenos*, *IV.ii.675*

CHAPTER 8

# CONCLUSION

How can words in a reading text be automatically predicted as difficult for a foreign language (L2) learner? Since the start of the doctoral study, there has been a surge in papers published on this subject, especially in the years following the two CWI shared tasks in computational linguistics (Paetzold & Specia, 2016a; Yimam et al., 2018). Previous studies on the automated prediction of lexical difficulty in reading have pursued two common objectives: to establish a reference measure or gold standard and train a statistical learning model with the best predictive power. However, previous studies in NLP have also been limited in two respects. Some studies predicted difficulty in context for the group of learners in general (i.e., from aggregated judgments in a reading task). In contrast, other studies predicted personal difficulty based on decontextualized measurements (i.e., from learner-specific vocabulary test data).

The doctoral study's research aims were to investigate whether adopting a contextualized (RQ1.1) and personalized (RQ1.2) approach could improve the prediction of lexical difficulty in L2 reading. Although common insights into the L2 reading process have long underscored the importance of the surrounding context and the individual learner, these factors have not been comprehensively addressed in recent advances in NLP. In fact, a systematic review of the literature showed that the NLP field was fairly disconnected from this evidence base in SLA and CALL research (RQ2.1). Moreover, the systematic scoping review corroborated the need for more contextualized and personalized measurements and predictions (RQ2.2).

This chapter gives a concluding discussion of the several studies conducted throughout the doctoral research project. Section 8.1 focuses on the main conclusions and implications regarding the several research questions addressed in this dissertation. Section 8.2 identifies some limitations and perspectives for further research.

## 8.1 MAIN CONCLUSIONS AND IMPLICATIONS

The first general question (RQ1.1) was whether adopting a contextualized approach could improve the prediction of lexical difficulty in L2 reading. However, the results remained tentative. Although there was some support for using WSD in the computation of word occurrence and semantic complexity (RQ4.2), the use of WSD was discontinued because of coverage issues, and because recent developments in NLP provided more suitable methods. With the inclusion of a BiLSTM layer in a neural network, the overall discriminatory power between difficulty and non-difficulty increased significantly (RQ7.1). However, this contextualized model made classifications that were not significantly better correlated with individual learner perceptions.

There were two plausible reasons for these tentative results. Firstly, despite the observed effect for contextual constraints (RQ5.2), the learner data may have had more to do with word decoding and lexical access than with word-to-context integration (RQ6.1). This conclusion could be drawn based on the effects observed for word surprisal. Both isolated and contextual word surprisal contributed to predicting lexical difficulty. However, isolated surprisal achieved a medium effect size, while contextual surprisal was negligible. Consequently, it was assumed that the data primarily indicated an inability to recognize the word rather than an inability to infer the meaning of the word from its surrounding context. Secondly, the statistical effect of context may have been less clear-cut because the predictive analyses targeted all words in the text. Most other studies observed an effect of context after selecting a more rigorously chosen set of target words and contexts from eye-tracking and EEG data (e.g., see Chen et al., 2017; Godfroid et al., 2013). Similarly, other studies observed more clear-cut word-to-context integration difficulties in a

delimited sentence reading task than in a more natural full-text reading task (cf. Dolgunsöz & Sariçoban, 2016).

The second general question (RQ1.2) was whether adopting a personalized approach could improve the prediction of lexical difficulty in L2 reading. Here, the results were unequivocal. As expected, there was a considerable variance in how learners perceived lexical difficulty while reading (RQ5.1). Consequently, it would have been unreliable to aggregate this self-assessment data into a single measure for training a predictive model. Furthermore, the results corroborated the necessity and capability of predictive models to account for such relative lexical complexity. There was a significant increase in predictive power for personalized models, for both mixed-effects models (RQ6.2) and personalized neural networks (RQ7.2). As such, the doctoral study addressed previous issues regarding the aggregated nature of training data and the disregard of relative (personalized) complexity measures in CWI shared task data, similarly noted by Finnimore et al. (2019) and Zampieri et al. (2017).

Finally, it should be stressed that these predictive models were evaluated on an empirical measure of perceived lexical difficulty. Initially, another difficulty measure was examined by looking at word occurrence in graded L2 reading materials. However, even though novel entries occurring at increasing proficiency levels displayed a clear trend in lexical complexity (RQ4.1), there were some incongruities in the levels at which translation equivalents occurred between languages (RQ4.3). As a result of these inconsistencies, this measure was not used for training a predictive model. Of course, some may argue that these inconsistent results were due to erroneously labeling difficulty as the first CEFR level at which the word occurred. The argument may be that relabeling lexical entries with another level would have considerably changed the results. However, considering that more than 60% of lexical entries appeared in one level only, such a 'transformation rule' would not have altered the difficulty label for the vast majority of analyzed items.

Now, what are the most important implications for future interdisciplinary research on the automated prediction of lexical difficulty in L2 reading?

### 8.1.1 *Implications for Educational NLP and the CWI Shared Tasks*

The first implication pertains to the CWI shared tasks. In this study, a similar methodology has been adopted, with a reference measure of how non-natives perceive lexical difficulty while reading in French. However, there are notable differences regarding data collection and performance metrics. Firstly, some issues may arise from inadequately defining the data collection procedure as an 'annotation' task. In the CWI task description, participants are referred to as 'annotators' who are 'annotating' complex words in a text. Yet, in an annotation task, we expect to obtain one ground truth (e.g., named entity labels or part of speech tags) based on which either a quantitative linguistic analysis is performed, or a statistical language model is trained. Either expert or crowd-sourced annotators analyze a corpus based on a well-defined linguistic theory and strict guidelines. Inter-annotator reliability coefficients such as Krippendorff's (2011) $\alpha$ are computed to assess the quality of the annotations.

In contrast, the CWI data sets indicate what has been referred to as *relative* linguistic complexity or *difficulty* (see the introduction in Chapter 1). Because the 'complex word annotation' task depends on the individual's perception of complexity, the identification of 'complex words' is relative to this specific individual.[77] Therefore, one cannot expect to obtain a single, unequivocal ground truth from this task. What is more, because of these individual differences, it is only expected that the agreement between participants will be low. It would, therefore, be inaccurate to use an inter-annotator agreement coefficient as an assessment of data quality.

Secondly, if the measurement of complexity is relative to an individual and if the agreement between individuals is low, data-related issues may arise from aggregating this diversity of measurements into a single training set (again, see Finnimore et al., 2019; Zampieri et al., 2017). Even if the participants have similar profiles (e.g., L1 and proficiency level), the results show that the agreement between participants reading the same texts remains below the accepted thresholds (both $\alpha < .8$ and $\alpha < .67$). As a result, training a

77 This does not mean, however, that it is not possible to think of a corpus that is annotated in terms of lexical complexity. In this case, one needs to resort to strict guidelines based on (psycho)linguistic theories that result in a single ground truth of lexical complexity (e.g., expert lexical simplifications for dyslexic children; Gala et al., 2020). In this case, it would be logical to use inter-rater reliability coefficients as an assessment of data quality.

predictive model of difficulty on an aggregation of all these measurements would result in inaccurate conclusions. There are two solutions to the issue of data aggregation. On the one hand, one could set aside the objective of training a predictive model of relative complexity and, instead, establish a reliable measure of absolute lexical complexity with a well-defined annotation task. On the other hand, one could accept individual differences in a measure of relative complexity and, consequently, proceed to train personalized models.

The latter option has been adopted throughout the study. In the judgments of difficulty made by NNSs reading the same texts, there is a recurrent finding: there are few words that all readers perceive as difficult, while there are many words that no reader perceives as difficult. The former observations pertain to terms occurring in low-constraint contexts, whereas the latter pertains to grammatical words. In between these two extremes, there is either partial or no agreement between learners. As one would expect based on theories of attention in SLA (Schmidt, 2001), there are notable individual differences in how learners at a similar proficiency level attend to lexical difficulty when they are reading the same texts.

Lastly, metrics such as the F-score make it impossible to compare system performance between different data distributions. The study highlights that these performance metrics are sensitive to changes in the prior distribution of difficulty. More specifically, the F-score of a constant predictive model is higher on a data set with a larger proportion of difficult words than on a data set with a smaller proportion of difficult words. Consequently, it is impossible to know whether an increase in F-score reflects an increase in predictive power or an increase in the percentage of difficulty. This sensitivity poses a challenge for comparing the predictive power of personalized models on learners-specific difficulty. It also restricts us from comparing system performance on different distributions such as the CWI shared task data. For this reason, the Phi and Tjur's D coefficients have been proposed.

By contextualizing and personalizing the predictions of lexical difficulty, a deep learning model can discriminate significantly better between words perceived as difficult or not. Among the several neural network models developed in this study, the best discriminatory power on tenfold cross-validation ($M_D$ = .64) is achieved by a BiLSTM model based on a word embedding layer,

two character-based CNN layers, and three learner-specific feature layers (viz., learner ID, proficiency level, and L1). For the most part, this performance is attributable to the presence of pre-trained FastText word embeddings. These distributional vectors account for the lexical features that are the most indicative of difficulty, namely word surprisal and morphosyntactic category. Other negatives effects are observed for the word's etymology (i.e., the number of languages borrowing the word) and the number of previous exposures to the word in the reading task, although these factors are not accounted for by the distributional vectors. In consequence, we find that most factors that contribute to predicting difficulty pertain to word use rather than to the word's form or meaning.

In addition to better discriminatory power, a personalized model makes classifications better correlated with how a learner perceives difficult words while reading. A personalized model significantly outperforms a non-personalized model on binomial correlation, for the group of learners in general ($M_\phi$ = .36 on tenfold cross-validation) and for each individual learner ($M_\phi$ = .35 on individual test sets). Whereas the inclusion of the learner's proficiency level (CEFR scale) only leads to a significant increase in discriminatory power, the inclusion of the individual learner (unique identifier) has a significant large effect on discriminatory power and binomial correlation. The inclusion of the learner's native language does not lead to a significant increase in predictive power.

### 8.1.2 *Implications for SLA and CALL*

Regarding the most significant predictors, the effects observed for word surprisal (among others) are not unexpected nor do they shed light on unexplored factors. The results are therefore very much in line with established research evidence. Nevertheless, two methodological aspects may inform future research in SLA and CALL.

The first methodological difference is that the predictive analyses are based on many factors and consider all of the running words in a text. Some recent studies in SLA (e.g., see Elgort et al., 2018) perform similar mixed-effects regression analyses on a much more comprehensive range of factors. However,

the predictive models developed in these studies are trained on a small selection of words from the reading materials (e.g., several repeated exposures to 14 low-frequency words and nine control words in a 12,152-word text). Because of this small number of test items, a much larger sample of L2 learners is required to identify significant effects with sufficient statistical power (cf. Brysbaert, 2019; Brysbaert & Stevens, 2018). By contrast, the predictive models developed in this study are trained on 261,942 items from a sample of 56 learners. Even though the number of participants may seem small compared to what is currently expected in applied linguistics, the number of tested items is also much more extensive than what is presently the methodological standard in this area. Consequently, to perform an adequately powered study, it is also very much key to either considerably increase the selection of test items or develop a model that works for all of the words in the text.

The second methodological difference is concerned with automation. In SLA and CALL studies, target words are manually selected based on word frequency measures or on pre-knowledge and familiarity (i.e., either based on vocabulary pretests or on learners highlighting unknown words). Considering recent advances in NLP, these selection criteria can be easily automated thanks to the availability of various powerful pre-trained language models and distributional word vectors. Both can be used to select target words based on frequency: language models can be used to compute word surprisal values, whereas distributional word vectors capture information on word frequency. Moreover, the deep learning models developed in this study could automatically predict which words will be highlighted as unknown in a text by a given learner.

### 8.1.3 *Implications for Foreign Language Teaching*

From a theoretical point of view, lexical difficulty can be measured by looking at reading materials intended for different levels of L2 development. When considering the types of words introduced in texts at increasing levels on the CEFR scale, we find a notable increase in lexical complexity. Words that are introduced at elementary levels are more frequent and dispersed, refer to more generic concepts (i.e., hypernyms), point to more concrete referents, and are acquired earlier by native speakers. Words that are introduced at more

advanced levels are less frequent and dispersed, refer to more specific concepts (i.e., hyponyms), point to more abstract referents, and are acquired later by native speakers. Although these findings are solely based on an examination of a new graded lexical resource developed specifically for Dutch as a foreign language (viz., NT2Lex; Tack et al., 2018b), they are coherent with previous conclusions on the link between linguistic complexity and L2 proficiency (e.g., Crossley & Salsbury, 2010). In consequence, these findings support the potential use of this data as a measure of lexical difficulty in L2 reading.

Nevertheless, the results also highlight that a measure of lexical difficulty established from CEFR-graded reading materials is not easily comparable between languages. There is a weak correlation between the levels at which equivalently translatable cognates are introduced in Dutch and French reading materials. For example, the Dutch word *flexibel* is already used in reading materials intended for the A2 proficiency level, whereas its French equivalent *flexible* first occurs in reading materials intended for the C1 proficiency level. The reasons for this discrepancy are not clear. They may be related either to a diverging distribution in each language or to differences in the corpus of reading materials for each language. What is clear, however, is that other recent studies observe similar issues, either pertaining to the alignment of distributions between languages (Graën et al., 2020) or to the alignment between word distributions in CEFR-graded materials and the CEFR reference level descriptors (Pintard & François, 2020). Consequently, although the current study demonstrates a coherence between the level of difficulty at which the word first occurs in L2 reading materials and its degree of lexical complexity, more research is needed to determine whether a valid reference measure of lexical difficulty can be derived from this data.

Despite this disadvantage, there are advantages to the development and use of CEFR-graded lexical resources. Because these resources can be easily queried with available web-based tools, they are helpful tools for studying word occurrence in materials intended for reading instruction. As such, they could be used to reflect on current teaching practices and stimulate further improvements on the CEFR and the creation of new graded materials for L2 readers.

## 8.2 LIMITATIONS AND PERSPECTIVES

Although the study shows support for adopting a personalized approach to predicting lexical difficulty, several avenues remain unexplored. The first avenue for further research is concerned with the 'stability' criterion in individual differences research (Dörnyei, 2005). A characteristic that varies between learners has to exhibit some degree of stability over time to be qualified as an individual difference. In this study, three individual differences have been investigated: the learner's unique ID, the learner's proficiency level, and the learner's native language. However, there are inevitably many more individual differences left exploring. In Dörnyei's taxonomy and in the systematic literature review (Chapter 2), we find other learner characteristics that remain unexplored and which pertain to either the learner's personality, the learner's aptitude, the learner's motivation, and the learner's learning styles. Furthermore, it should be noted that this study only provides a snapshot of the learner at a given point in time. A longitudinal study is warranted to investigate stability in the effect of different personal characteristics and evaluate the performance of a personalized model over time.

A second avenue is concerned with reproducing the methodology on other samples of learners and different target languages. As the data is not representative of all possible individuals learning French as an L2, the study is limited in terms of generalization. One specific course of action is to reproduce the method on a more diverse sample in terms of native languages. This study would be necessary to investigate the inconclusive effect of L1 in the predictive model. One possibility would be to explore the data collected by Gala et al. (2019), who gathered similar data from a more varied sample of French L2 learners.

A final avenue is concerned with the question of what remains ultimately unpredictable. Although the results of this study indicate the most predictable indices of difficulty (e.g., word surprisal), there are still parts of the data that the model cannot explain. Seeing the low percentage of observations found difficult by all learners, it is not entirely sure whether it will be possible to explain all of the data. It may be that mere idiosyncrasies characterize parts of the data. Nevertheless, it is equally possible that some parts may be

explained by factors that have been reviewed in Chapter 2 but which have not been explored in this study. Besides considering a wider variety of learner characteristics, it is important to look further into task-related and other methodological factors.

In the automated identification of lexical difficulty in L2 reading, two things are essential to predict: (a) the words a learner will perceive as difficult and (b) the words with which a learner will have difficulty. As stated in the introduction to this dissertation (Chapter 1), this distinction is crucial because L2 learners do not necessarily notice words unfamiliar to them. When a system can predict both perceived and actual difficulty, this technology can be used to direct the learners' attention towards difficult words they may not perceive as difficult. All things considered, it seems more challenging to develop a system able to predict both perceived and actual difficulty. One way in which such a system could be developed is by administering a reading task in which several measures of lexical knowledge are tested (e.g., form and meaning recognition). On the basis of this data, a machine learning system is trained. Although this methodology is similar to the one adopted in this study, it poses several disadvantages. Firstly, this procedure will very likely result in a highly cognitively loaded task. Ideally, we would like to have an accurate measure of lexical knowledge for all of the running words in a text. This is practically impossible without negatively interfering with the natural flow of the reading task. Furthermore, because this cognitively loaded task will likely slow down the data collection procedure, there will be significantly less data on the basis of which an accurate machine learning system can be trained. An alternative way in which such a system could be developed is by tracking the learner in a language learning platform and by optimizing a model of lexical knowledge from different learning tasks. This model could then be used to contrast the predictions of perceived lexical difficulty with predictions of actual lexical difficulty.

# BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org/ (cit. on p. 334)

Abdollahi, M., & Farvardin, A. T. (2016). Demystifying the effect of narrow reading on EFL learners' vocabulary recall and retention. *Education Research International*, *2016*(5454031), 1–10. https://doi.org/10.1155/2016/5454031 (cit. on p. 62)

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, *21*(3), 199–226. https://doi.org/10.1080/09588220802090246 (cit. on pp. 4, 12, 15, 21, 32, 33)

AbuRa'ed, A., & Saggion, H. (2018). LaSTUS/TALN at Complex Word Identification (CWI) 2018 Shared Task. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 159–165. http://www.aclweb.org/anthology/W18-0517 (cit. on pp. 128–130, 145)

AbuSeileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. *Computers & Education*, *57*(1), 1281–1291. https://doi.org/10.1016/j.compedu.2011.01.011 (cit. on pp. 15, 64)

AbuSeileek, A. F. M. (2008). Hypermedia annotation presentation: Learners' preferences and effect on EFL reading comprehension and vocabulary acquisition. *CALICO Journal*, *25*(2), 260–275 (cit. on pp. 15, 64).

Adams, S. J. (1982). Scripts and the recognition of unfamiliar vocabulary: Enhancing second language reading skills. *Modern Language Journal*, *66*(2), 155–159 (cit. on p. 14).

Alfter, D., & Pilán, I. (2018). SB@GU at the Complex Word Identification 2018 Shared Task. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 315–321. http://www.aclweb.org/anthology/W18-0537 (cit. on pp. 128–130, 254)

Alfter, D., & Volodina, E. (2018). Towards single word lexical complexity prediction. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 79–88. https://doi.org/10/gfzmkx (cit. on pp. 18, 200)

Amancio, M., & Specia, L. (2014). An analysis of crowdsourced text simplifications. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 123–130. http://www.aclweb.org/anthology/W14-1214 (cit. on p. 105)

Aroyehun, S. T., Angel, J., Pérez Alvarez, D. A., & Gelbukh, A. (2018). Complex word identification: Convolutional neural network vs. feature engineering. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 322–327. http://www.aclweb.org/anthology/W18-0538 (cit. on pp. 128–130)

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2 (cit. on p. 222)

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies*, *14*(3), 281–299. https://doi.org/10/dhgtvh (cit. on p. 138)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10/gcrnkw (cit. on pp. 219, 223, 249, 266, 291, 321)

Beacco, J.-C. (2004). *Niveau B2 pour le français (utilisateur/apprenant indépendant): un référentiel*. Didier. (Cit. on p. 138).

Beacco, J.-C. (2008). *Niveau A1 et niveau A2 pour le français: textes et références : utilisateur/apprenant élémentaire*. Didier. (Cit. on p. 138).

Beacco, J.-C., Blin, B., Houles, E., & Riba, P. (2011). *Niveau B1 pour le français (apprenant/utilisateur indépendant). Niveau seuil: un référentiel*. Didier. (Cit. on p. 138).

Beacco, J.-C., de Ferrari, M., Lhote, G., & Tagliante, C. (2005). Niveau A1.1 pour le français: publics adultes peu francophones, scolarisés, peu ou non scolariés : référentiel et certification (DILF) pour les premiers acquis en français. (Cit. on p. 138).

Beacco, J.-C., Lepage, S., Porquier, R., Riba, P., & Pellieux, N. (2008). *Niveau A2 pour le français (utilisateur/apprenant élémentaire). Niveau intermédiaire: un référentiel*. Didier. (Cit. on p. 138).

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning*, *62*(3), 665–703. https://doi.org/10.1111/j.1467-9922.2011.00651.x (cit. on p. 6)

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Readability for foreign language learning: The importance of cognates. *ITL - International Journal of Applied Linguistics*, *165*(2), 136–162. https://doi.org/10.1075/itl.165.2.02bei (cit. on p. 148)

Bell, F. L., & LeBlanc, L. B. (2000). The language of glosses in L2 reading on computer: Learners' preferences. *Hispania*, *83*(2), 274–285 (cit. on p. 67).

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155 (cit. on p. 294).

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*(1), 15–31. https://doi.org/10.1111/j.1467-9817.1984.tb00252.x (cit. on p. 14)

Benzahra, M., & Yvon, F. (2019). Measuring text readability with machine comprehension: A pilot study. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 412–422. https://doi.org/10/gg9w8s (cit. on p. 207)

Billami, M., François, T., & Gala, N. (2018). ReSyf: a French lexicon with ranked synonyms. *Proceedings of COLING 2018* (cit. on pp. 141, 142).

Bingel, J., Barrett, M., & Klerke, S. (2018). Predicting misreadings from gaze in children with reading difficulties. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 24–34. http://www.aclweb.org/anthology/W18-0503 (cit. on pp. 120, 294, 295)

Bingel, J., & Bjerva, J. (2018). Cross-lingual complex word identification with multitask learning. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 166–174. http:

//www.aclweb.org/anthology/W18-0518 (cit. on pp. 128–131, 145, 294, 295)

Bingel, J., Paetzold, G., & Søgaard, A. (2018). Lexi: A tool for adaptive, personalized text simplification. *Proceedings of the 27th International Conference on Computational Linguistics*, 245–258. https://www.aclweb.org/anthology/C18-1021 (cit. on p. 12)

Bingel, J., Schluter, N., & Martínez Alonso, H. (2016). CoastalCPH at SemEval-2016 Task 11: The importance of designing your neural networks right. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1028–1033. http://www.aclweb.org/anthology/S16-1160 (cit. on pp. 128–131, 254, 265, 294, 295, 332)

Bird, S., Klein, E., & Loper, E. (2009, July 7). *Natural Language Processing with Python* (1st ed.). O'Reilly. (Cit. on p. 157).

Birjandi, P., Alavi, S. M., & Najafi Karimi, S. (2015). Effects of unenhanced, enhanced, and elaborated input on learning English phrasal verbs. *International Journal of Research Studies in Language Learning*, *4*(1), 43 (cit. on pp. 64, 231).

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press, Inc. (Cit. on p. 293).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008 (cit. on p. 46)

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, *66*, 113 (cit. on pp. 62, 64, 66).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10/gfw9cs (cit. on pp. 262, 298)

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric (Q. Zou, Ed.). *PLOS ONE*, *12*(6), e0177678. https://doi.org/10/gbgsqd (cit. on p. 311)

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, *87*(3), 541–552 (cit. on p. 119).

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Routledge. (Cit. on p. 119).

Brew, C., McKelvie, D., & Place, B. (1996). Word-pair extraction for lexicography. *Proceedings of the Second International Conference on New Methods in Language Processing*, 45–55 (cit. on p. 191).

Brooke, J., Uitdenbogerd, A., & Baldwin, T. (2016). Melbourne at SemEval 2016 Task 11: Classifying type-level word complexity using random forests with corpus and word list features. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 975–981. http://www.aclweb.org/anthology/S16-1150 (cit. on pp. 104, 128–130)

Brouwers, L., Bernhard, D., Ligozat, A.-L., & François, T. (2014). Syntactic sentence simplification for French. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 47–56. http://www.aclweb.org/anthology/W14-1206 (cit. on pp. 105, 208)

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10/gf5hfj (cit. on p. 351)

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 1–13. https://doi.org/10.3758/s13428-018-1077-9 (cit. on p. 259)

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977 (cit. on p. 167)

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10 (cit. on pp. 234, 351)

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010 (cit. on p. 171)

Butnaru, A., & Ionescu, R. T. (2018). UnibucKernel: A kernel-based learning method for complex word identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 175–183. http://www.aclweb.org/anthology/W18-0519 (cit. on pp. 128–130)

Campbell, L., & Mixco, M. J. (2007). *A glossary of historical linguistics*. Edinburgh Univ. Press. (Cit. on p. 146).

Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, *1*(01). https://doi.org/10.1017/S2041536210000048 (cit. on p. 138)

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, *3*. https://doi.org/10.1017/S2041536212000013 (cit. on pp. 138, 198)

Carger, C. L. (1993). Louie comes to life: Pretend reading with second language emergent readers. *Language Arts*, *70*(7), 542–547 (cit. on p. 6).

Carroll, J. B., Davies, P., & Richman, B. (1971, January 2). *The American heritage word frequency book*. Houghton Mifflin. (Cit. on pp. 140, 159).

Carver, R. P. (1982). Optimal rate of reading prose. *Reading Research Quarterly*, *18*(1), 56. https://doi.org/10/cjzx7b (cit. on p. 213)

Chen, B., Ma, T., Liang, L., & Liu, H. (2017). Rapid L2 word learning through high constraint sentence context: An event-related potential study. *Frontiers in Psychology*, *8*, 2285. https://doi.org/10.3389/fpsyg.2017.02285 (cit. on pp. 18, 117, 118, 125, 132, 134, 231, 346)

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693–713 (cit. on p. 61).

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864. https://doi.org/10/cxshr5 (cit. on pp. 227, 275)

Chen, I.-J. (2016). Hypertext glosses for foreign language reading comprehension and vocabulary acquisition: Effects of assessment methods. *Computer Assisted Language Learning*, *29*(2), 413–426. https://doi.org/10.1080/09588221.2014.983935 (cit. on pp. 15, 64, 65)

Chen, I.-J., & Yen, J.-C. (2013). Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages. *Computers & Education*, *63*, 416–423. https://doi.org/10.1016/j.compedu.2013.01.005 (cit. on pp. 15, 64, 65)

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113–119. https://www.aclweb.org/anthology/W16-4113 (cit. on p. 12)

Chen, Y. (2012). Dictionary use and vocabulary learning in the context of reading. *International Journal of Lexicography*, *25*(2), 216–247 (cit. on p. 65).

Cheng, Y.-H., & Good, R. L. (2009). L1 glosses: Effects on EFL learners' reading comprehension and vocabulary retention. *Reading in a Foreign Language*, *21*(2), 119–142 (cit. on pp. 64, 65).

Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences*, *45*, 137–143. https://doi.org/10.1016/j.lindif.2015.11.018 (cit. on pp. 61, 64)

Choi, S., Kim, J., & Ryu, K. (2014). Effects of context on implicit and explicit lexical knowledge: An event-related potential study. *Neuropsychologia*, *63*, 226–234 (cit. on p. 117).

Choubey, P., & Pateria, S. (2016). Garuda & Bhasha at SemEval-2016 Task 11: Complex word identification using aggregated learning models. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1006–1010. http://www.aclweb.org/anthology/S16-1156 (cit. on pp. 128–130, 254)

Chun, D. M., & Payne, J. S. (2004). What makes students click: Working memory and look-up behavior. *System*, *32*(4), 481–503 (cit. on p. 66).

Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. *Proceedings of the 31st Annual Meeting of the Association*

*for Computational Linguistics*, 1–8. https://doi.org/10/d2psjr (cit. on p. 191)

Coady, J. (1996). L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 225–237). Cambridge University Press. https://doi.org/10.1017/CBO9781139524643.016. (Cit. on pp. 6, 29)

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284. https://doi.org/10.1037/h0076540 (cit. on pp. 9, 103)

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448–1462. https://doi.org/10.1002/asi.20243 (cit. on p. 207)

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537 (cit. on p. 294).

Constant, M., & Sigogne, A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*, 49–56. https://hal-upec-upem.archives-ouvertes.fr/hal-00621585/document (cit. on pp. 214, 263)

Council of Europe. (2001). *Common european framework of reference for languages*. Cambridge University Press. (Cit. on pp. 11, 138).

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951 (cit. on p. 104)

Critchlow, D. E., & Fligner, M. A. (1991). On distribution-free multiple comparisons in the one-way analysis of variance. *Communications in Statistics - Theory and Methods*, *20*(1), 127–139. https://doi.org/10/djkzgm (cit. on p. 183)

Crossley, S. (2013). Assessing automatic processing of hypernymic relations in first language speakers and advanced second language learners: A semantic priming approach. *The Mental Lexicon*, *8*(1), 96–116. https://doi.org/10.1075/ml.8.1.05cro (cit. on pp. 145, 199)

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334. https://doi.org/10.1111/j.1467-9922.2009.00508.x (cit. on pp. 131, 144, 174, 179, 186, 198)

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15–30. https://doi.org/10.1111/j.1540-4781.2007.00507.x (cit. on pp. 144, 179, 198)

Crossley, S. A., & Salsbury, T. (2010). Using lexical indices to predict produced and not produced words in second language learners. *The Mental Lexicon*, *5*(1), 115–147 (cit. on pp. 144, 198, 352).

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6th ed). Blackwell Pub. (Cit. on p. 146).

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*(1), 11–28 (cit. on pp. 9, 104, 137).

Danesh, T., & Farvardin, M. T. (2016). A comparative study of the effects of different glossing conditions on EFL learners' vocabulary recall. *Sage Open*, *6*(3), UNSP 2158244016669548. https://doi.org/10.1177/2158244016669548 (cit. on p. 67)

Davoodi, E., & Kosseim, L. (2016). CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 982–985. http://www.aclweb.org/anthology/S16-1151 (cit. on pp. 128–130, 283)

Day, R. R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*(2), 541–51 (cit. on p. 31).

De Hertog, D., & Tack, A. (2018). Deep learning architecture for complex word identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 328–334. https://www.aclweb.org/anthology/W18-0539 (cit. on pp. 23, 128–131, 294–296, 298, 300, 302, 303, 329)

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit*

*social cognition: Measurement, theory, and applications* (pp. 176–193). Guilford Press. (Cit. on pp. 112, 120).

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. *LREC*, *14*, 4585–92. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf (cit. on p. 264)

De Ridder, I. (2000). Are we conditioned to follow links? Highlights in CALL materials and their impact on the reading process. *Computer Assisted Language Learning*, *13*(2), 183–94. https://doi.org/10.1076/0958-8221(200004)13:2;1-D;FT183 (cit. on pp. 15, 64)

De Ridder, I. (2002). Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? *Language Learning & Technology*, *6*(1), 123–46 (cit. on pp. 15, 64).

de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21. https://doi.org/10.1017/S1366728906002732 (cit. on p. 143)

de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1–56. https://doi.org/10.1111/0023-8333.00110 (cit. on pp. 147, 148)

de Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*(1), 90–123. https://doi.org/10/bd3hwp (cit. on p. 148)

de Kleijn, P., & Nieuwberg, E. (1993). *Basiswoordenboek Nederlands*. Wolters Leuven. (Cit. on p. 169).

de Landsheere, G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. *Le Travail Humain*, *26*(1/2), 141–154 (cit. on pp. 209, 215, 233).

de Melo, G. (2014, May). Etymological Wordnet: Tracing the history of words. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of*

*the 9th Language Resources and Evaluation Conference (LREC 2014)* (pp. 1148–1154). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1083_Paper.pdf. (Cit. on pp. 189, 265)

Deneckere, M. (1954). Histoire de la langue française dans les Flandres ( 1770-1823). *Handelingen der Maatschappij voor Geschiedenis en Oudheidkunde te Gent*, *8*(1). https://doi.org/10/gg87dz (cit. on p. 200)

Deville, G., Dumortier, L., & Meurisse, J.-R. (2019). *A corpus-based context-sensitive reading tool for learners of English and Dutch*. The 27th Conference of the European Association for Computer-Assisted Language Learning (EUROCALL 2019), Louvain-la-Neuve, Belgium. (Cit. on p. 4).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423 (cit. on pp. 207, 294)

Dietterich, T. G. (2000). Ensemble methods in machine learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.), *Multiple Classifier Systems* (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1. (Cit. on p. 127)

Dirix, N., Brysbaert, M., & Duyck, W. (2019). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods*, *51*(6), 2800–2816. https://doi.org/10/gf4vfx (cit. on p. 103)

Dolch, E. W. (1936). A basic sight vocabulary. *The Elementary School Journal*, *36*(6), 456–460 (cit. on p. 104).

Dolgunsöz, E. (2016). Using eye-tracking to measure lexical inferences and its effects on reading rate during EFL reading. *Journal of Language and Linguistic Studies*, *12*(1), 63–78 (cit. on pp. 65, 66, 116, 133, 134).

Dolgunsöz, E., & Sariçoban, A. (2016). CEFR and eye movement characteristics during EFL reading: The case of intermediate readers. *Journal of Language and Linguistic Studies*, *12*(2), 238–252 (cit. on pp. 65, 115, 116, 134, 347).

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum Associates. (Cit. on pp. 19, 353).

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford Univ. Press. (Cit. on p. 55).

Duan, S. (2018). Effects of enhancement techniques on L2 incidental vocabulary learning. *English Language Teaching*, *11*(3), 88–101 (cit. on pp. 64, 65).

Dürlich, L., & François, T. (2018–May 12). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (cit. on pp. 139, 141).

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 663–679. https://doi.org/10.1037/0278-7393.33.4.663 (cit. on pp. 148, 192)

Dwass, M. (1960). Some k-sample rank-order tests. In I. Olkin, . G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 198–202). Stanford University Press. (Cit. on p. 183).

Ebadi, S., Weisi, H., Monkaresi, H., & Bahramlou, K. (2018). Exploring lexical inferencing as a vocabulary acquisition strategy through computerized dynamic assessment and static assessment. *Computer Assisted Language Learning*, *31*(7), 790–817. https://doi.org/10.1080/09588221.2018.1451344 (cit. on p. 15)

Ehara, Y. (2019). Uncertainty-aware personalized readability assessments for second language learners. *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1909–1916. https://doi.org/10/ghrq43 (cit. on pp. 18–20, 295)

Ehara, Y. (2020). Interpreting neural CWI classifiers' weights as vocabulary size. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 171–176. https://doi.org/10/ghk4px (cit. on pp. 18–20, 295)

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109 (cit. on pp. 14, 18, 61, 65, 115, 116, 284, 288, 289, 350)

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*(2), 365–414. https://doi.org/10.1111/lang.12052 (cit. on pp. 61, 62, 65, 66, 284, 288, 289)

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024 (cit. on pp. 174, 258)

Ellis, R. (1999). Factors in the incidental acquisition of second language vocabulary from oral input. *Learning a Second Language through Interaction* (pp. 35–61). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.17. (Cit. on pp. 31, 55)

Ender, A. (2016). Implicit and explicit cognitive processes in incidental vocabulary acquisition. *Applied Linguistics*, *37*(4), 536–560. https://doi.org/10.1093/applin/amu051 (cit. on pp. 14, 66, 119, 121, 122)

English Profile. (2011). *EnglishProfile: Introducing the CEFR for English*. Cambridge University Press. (Cit. on p. 138).

Eschenbach, T. G. (1992). Spiderplots versus tornado diagrams for sensitivity analysis. *Interfaces*, *22*(6), 40–46. https://doi.org/10/dgxfxc (cit. on pp. 313, 314)

Farkas, I., & Li, P. (2002). DevLex: A self-organizing neural network model of the development of lexicon. *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, 2546–2551 vol.5. https://doi.org/10/gg8zbq (cit. on p. 143)

Farvardin, M. T., & Biria, R. (2012). The impact of gloss types on Iranian EFL students' reading comprehension and lexical retention. *International Journal of Instruction*, *5*(1), 99–114 (cit. on p. 64).

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and

biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10/bxjdcg (cit. on p. 234)

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press. (Cit. on pp. 143, 257).

Filatova, K. (2010). Third language acquisition, macrocategories and synonymy. In M. Pütz & L. Sicola (Eds.), *Converging Evidence in Language and Communication Research* (pp. 85–96). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.13.08fil. (Cit. on p. 143)

Finnimore, P., Fritzsch, E., King, D., Sneyd, A., Ur Rehman, A., Alva-Manchego, F., & Vlachos, A. (2019). Strong baselines for complex word identification across multiple languages. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 970–977. https://doi.org/10/ghdhcs (cit. on pp. 13, 289, 347, 348)

Firth, J. R. (1957). Applications of general linguistics. *Transactions of the Philological Society*, *56*(1), 1–14. https://doi.org/10/fgwn6n (cit. on p. 17)

Flesch, R. (1951). *How to test readability*. Harper. (Cit. on pp. 9, 208, 209, 213–215, 233).

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532 (cit. on pp. 9, 103)

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. (Cit. on p. 291).

Francois, T., Gala, N., Watrin, P., & Fairon, C. (2014, May). FLELex: A graded lexical resource for french foreign learners. In N. C. ( Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA). (Cit. on pp. 11, 21, 139, 141, 151, 209, 258, 282).

François, T., & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–477 (cit. on p. 11).

François, T., Volodina, E., Pilán, I., & Tack, A. (2016–May 28). SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 213–219. http://www.lrec-conf.org/proceedings/lrec2016/pdf/275_Paper.pdf (cit. on pp. 139, 141)

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10/f43xkh (cit. on pp. 115, 288)

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 878–883. https://www.aclweb.org/anthology/P13-2152 (cit. on pp. 18, 125, 132)

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10/f6x56q (cit. on pp. 18, 125, 126, 132, 288)

Fraser, C. A. (2007). Reading rate in L1 mandarin chinese and L2 english across five reading tasks. *The Modern Language Journal*, *91*(3), 372–394. https://doi.org/10/cm84x2 (cit. on p. 213)

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, *35*(4), 376–403. https://doi.org/10/c5xcf2 (cit. on p. 117)

Furtner, M. R., Rauthmann, J. F., & Sachse, P. (2011). Investigating word class effects in first and second languages. *Perceptual and Motor Skills*, *113*(1), 87–97. https://doi.org/10.2466/04.11.28.PMS.113.4.87-97 (cit. on p. 116)

Gagné, C. L., Spalding, T. L., Spicer, P., Wong, D., Rubio, B., & Cruz, K. P. (2020). Is buttercup a kind of cup? Hyponymy and semantic transparency in compound words. *Journal of Memory and Language*, *113*, 104110. https://doi.org/10/gg8zj3 (cit. on p. 145)

Gala, N., François, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *Electronic Lexicography in the*

*21st Century: Thinking Outside the Paper : Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia, 2013, Págs. 132-151*, 132–151. https://dialnet.unirioja.es/servlet/articulo?codigo=4563520 (cit. on pp. 141, 142, 200)

Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 1346–1354. https://www.aclweb.org/anthology/2020.lrec-1.169/ (cit. on pp. 101, 106, 108–110, 121, 348)

Gala, N., David, C., Tack, A., & François, T. (2019, August 29). *Assessing vocabulary knowledge for learners of French as a foreign language: Accounting for L1 variability to go beyond the CEFR scale*. The 27th Conference of the European Association for Computer-Assisted Language Learning (EUROCALL 2019), Louvain-la-Neuve, Belgium. (Cit. on p. 353).

Gala, N., & Javourey-Drevet, L. (2020). Mots « faciles » et mots « difficiles » dans ReSyf : un outil pour la didactique du lexique mobilisant polysémie, synonymie et complexité. *Lidil*, (62). https://doi.org/10/gjr32k (cit. on p. 141)

Godfroid, A. (2019, October 31). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide* (1st ed.). Routledge. https://doi.org/10.4324/9781315775616. (Cit. on p. 115)

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483–517. https://doi.org/10.1017/S0272263113000119 (cit. on pp. 14, 18, 64, 66, 116, 134, 346)

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the Noticing Hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Converging Evidence in Language and Communication Research* (pp. 169–197). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.13.14god. (Cit. on p. 40)

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence.

*Language Learning*, *65*(4), 896–928. https://doi.org/10.1111/lang.12136 (cit. on pp. 67, 119)

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press. (Cit. on p. 294).

Gooding, S., & Kochmar, E. (2018). CAMB at CWI Shared Task 2018: Complex word identification with ensemble-based voting. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 184–194. http://www.aclweb.org/anthology/W18-0520 (cit. on pp. 127–130, 254)

Gooding, S., Kochmar, E., Sarkar, A., & Blackwell, A. (2019). Comparative judgments are more consistent than binary classification for labelling word complexity. *Proceedings of the 13th Linguistic Annotation Workshop*, 208–214. https://doi.org/10/ggdhz3 (cit. on pp. 18, 20)

Gougenheim, G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Didier. (Cit. on p. 260).

Graën, J., Alfter, D., & Schneider, G. (2020). Using multilingual resources to evaluate CEFRLex for learner applications. *Proceedings of The 12th Language Resources and Evaluation Conference*, 346–355. https://www.aclweb.org/anthology/2020.lrec-1.43 (cit. on pp. 200, 352)

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564 (cit. on pp. 12, 144)

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (cit. on pp. 189, 262, 299).

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation (S. Nakagawa, Ed.). *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10/f8hm9v (cit. on pp. 227, 234)

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri (cit. on p. 140)

Grinstead, W. J. (1915). An experiment in the learning of foreign words. *Journal of Educational Psychology*, *6*(4), 242–245. https://doi.org/10.1037/h0071403 (cit. on p. 29)

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill. (Cit. on pp. 9, 103, 137).

Gurzynski-Weiss, L., & Plonsky, L. (2017). Look who's interacting: A scoping review of research involving non-teacher/non-peer interlocutors. In L. Gurzynski-Weiss (Ed.), *AILA Applied Linguistics Series* (pp. 306–324). John Benjamins Publishing Company. https://doi.org/10.1075/aals.16.13gur. (Cit. on pp. 20, 34, 48)

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective – Challenges and potential solutions. In C. Bardel, B. Laufer, & C. Lindqvist (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 11–28). EUROSLA. (Cit. on p. 142).

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11–15). (Cit. on pp. 44, 100).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://doi.org/10/b2t3v6 (cit. on pp. 125, 126)

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412. https://doi.org/10/ggh2gp (cit. on p. 125)

Hamada, A. (2015). Effects of forward and backward contextual elaboration on lexical inferences: Evidence from a semantic relatedness judgment task. *Reading in a Foreign Language*, *27*(1), 1–21 (cit. on pp. 14, 64, 231).

Hamada, M., & Koda, K. (2011). Similarity and difference in learning L2 word-form. *System*, *39*(4), 500–509. https://doi.org/10.1016/j.system.2011.10.011 (cit. on p. 7)

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, *32*(2), 135–154. https://doi.org/10/bp38m8 (cit. on p. 313)

Han, Z. (2009). Interlanguage and fossilization: Towards an analytic model. In V. Cook & L. Wei (Eds.), *Contemporary Applied Linguistics: Volume 1 Language Teaching and Learning* (pp. 137–162). Continuum. (Cit. on p. 10).

Han, Z. (2013). Forty years later: Updating the Fossilization Hypothesis. *Language Teaching*, *46*(2), 133–171. https://doi.org/10/ghz752 (cit. on p. 10)

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'10, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10/ghbzf2 (cit. on p. 334)

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. https://doi.org/10/gcwbsg (cit. on p. 17)

Hartmann, N., & dos Santos, L. B. (2018). NILC at CWI 2018: Exploring feature engineering and feature learning. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 335–340. http://www.aclweb.org/anthology/W18-0540 (cit. on pp. 128–130)

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: theory and illustrations. *English Profile Journal*, *1*(01), 1–23 (cit. on p. 138).

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239 (cit. on p. 311)

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified kneser-ney language model estimation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 690–696. https://www.aclweb.org/anthology/P13-2121 (cit. on pp. 125, 254, 260)

Heck, R. H., & Thomas, S. L. (2008). *An introduction to multilevel modeling techniques* (3rd ed.). Routledge. (Cit. on p. 223).

Herman, P. A. (1987). Incidental acquisition of word meaning from expositions with varied text features. *Reading Research Quarterly*, *22*(3), 263–84 (cit. on p. 31).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on p. 304)

Holley, F. M., & King, J. K. (1971). Vocabulary glosses in foreign language reading materials. *Language Learning*, *21*(2), 213–219. https://doi.org/10.1111/j.1467-1770.1971.tb00060.x (cit. on p. 64)

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. (Cit. on p. 264).

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799 (cit. on p. 7)

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10/b53crr (cit. on pp. 10, 12)

Hu, S.-M., Vongpumivitch, V., Chang, J. S., & Liou, H.-C. (2014). The effects of L1 and L2 e-glosses on incidental vocabulary learning of junior high-school English students. *ReCALL*, *26*(1), 80–99. https://doi.org/10.1017/S0958344013000244 (cit. on pp. 15, 64, 65)

Huang, H.-T., & Liou, H.-C. (2007). Vocabulary learning in an automated graded reading program. *Language Learning & Technology*, *11*(3), 64–82 (cit. on p. 61).

Huang, S., Willson, V., & Eslami, Z. (2012). The Effects of Task Involvement Load on L2 Incidental Vocabulary Learning: A Meta-Analytic Study. *Modern Language Journal*, *96*(4), 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x (cit. on p. 32)

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, *21*(2), 181–193 (cit. on pp. 21, 32).

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 258–286). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780. (Cit. on p. 31)

Hulstijn, J. H., Alderson, J. C., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: are there links between them? In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 11–12). European Second Language Association. (Cit. on p. 138).

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal*, *80*(3), 327–339. https://doi.org/10.2307/329439 (cit. on pp. 61, 64)

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language*, *17*(1), 23–59. https://eric.ed.gov/?id=EJ689121 (cit. on pp. 21, 31)

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10/drbjhg (cit. on p. 100)

Inkpen, D., & Frunza, O. (2005). Automatic identification of cognates and false friends in French and English. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 251–257 (cit. on p. 191).

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, *17*(1), 19–28 (cit. on p. 64).

Jauhar, S. K., & Specia, L. (2012). UOW-SHEF: SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features. *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and*

*Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 477–481. http://dl.acm.org/citation.cfm?id=2387636.2387715 (cit. on p. 142)

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034 (cit. on pp. 4, 7, 33)

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10/gghpn6 (cit. on p. 311)

Jones, E., Oliphant, T., Peterson, P., et al. (2001). *SciPy: Open source scientific tools for Python*. http://www.scipy.org/. (Cit. on p. 99)

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice-Hall, Inc. (Cit. on p. 124).

Kaivanpanah, S., & Moghaddam, M. (2012). Knowledge sources in EFL learners' lexical inferencing across reading proficiency levels. *RELC Journal*, *43*(3), 373–391. https://doi.org/10.1177/0033688212469219 (cit. on p. 65)

Kaivanpanah, S., & Rahimi, N. (2017). The effect of contextual clues and topic familiarity on L2 lexical inferencing and retention. *Porta Linguarum*, (27), 47–61. https://dialnet.unirioja.es/servlet/articulo?codigo=6151249 (cit. on pp. 62, 65, 231)

Kajiwara, T., & Komachi, M. (2018). Complex word identification based on frequency in a learner corpus. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 195–199. http://www.aclweb.org/anthology/W18-0521 (cit. on pp. 128, 129)

Kandel, L., & Moles, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers d'études de radio-télévision*, *19*, 253–79 (cit. on pp. 209, 215, 233).

Kang, E. Y. (2015). Promoting L2 vocabulary learning through narrow reading. *RELC Journal*, *46*(2), 165–179. https://doi.org/10.1177/0033688215586236 (cit. on p. 62)

Kauchak, D. (2016). Pomona at SemEval-2016 Task 11: Predicting word complexity based on corpus frequency. *Proceedings of the 10th International*

*Workshop on Semantic Evaluation (SemEval-2016)*, 1047–1051. http://www.aclweb.org/anthology/S16-1164 (cit. on pp. 128, 129)

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10/c266jz (cit. on p. 44)

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643 (cit. on p. 167)

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology (2006)*, *68*(8), 1665–1692. https://doi.org/10.1080/17470218.2015.1022560 (cit. on p. 259)

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*(5), 1155–1180 (cit. on pp. 62, 65, 114).

Kim, Y. (2006). Effects of input elaboration on vocabulary acquisition through reading by Korean learners of English as a foreign language. *TESOL Quarterly*, *40*(2), 341–373 (cit. on p. 64).

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, *78*(3), 285–299. https://doi.org/10.2307/330108 (cit. on pp. 36, 64, 66)

Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, *46*(1), 56–79. https://doi.org/10.1002/tesq.3 (cit. on p. 64)

Koda, K. (1996). Orthographic knowledge in L2 lexical processing: A cross-linguistic perspective. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 35–52). Cambridge University Press. https://doi.org/10.1017/CBO9781139524643.005. (Cit. on p. 7)

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. *Second Meeting of the North American Chapter of the Association for*

*Computational Linguistics*. https://doi.org/10.3115/1073336.1073350 (cit. on p. 147)

Konkol, M. (2016). UWB at SemEval-2016 Task 11: Exploring features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1038–1041. http://www.aclweb.org/anthology/S16-1162 (cit. on pp. 128–130, 254)

Kortmann, B., & Szmrecsanyi, B. (Eds.). (2012). *Linguistic complexity: Second language acquisition, indigenization, contact*. De Gruyter. (Cit. on pp. 9, 11).

Krashen, S. (1978). The Monitor Model for second language acquisition. In R. C. Gingras (Ed.), *Second Language Acquisition and Foreign Language Teaching* (pp. 1–26). Center for Applied Linguistics. (Cit. on p. 8).

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464 (cit. on pp. 8, 17, 31).

Krashen, S. D. (1981). The case for narrow reading. *TESOL Newsletter*, *15*, 23 (cit. on p. 6).

Krashen, S. D. (2004). The case for narrow reading. *Language Magazine*, *3*(5), 17–19 (cit. on p. 6).

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. http://repository.upenn.edu/asc_papers/43 (cit. on pp. 222, 348)

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, *66*(3), 563–580. https://doi.org/10/gjqmxt (cit. on p. 103)

Kuru, O. (2016). AI-KU at SemEval-2016 Task 11: Word embeddings and substring features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1042–1046. http://www.aclweb.org/anthology/S16-1163 (cit. on pp. 128, 130)

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10/ffh9zj (cit. on p. 117)

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10/c6ndp4 (cit. on p. 117)

Lafourcade, M. (2007). Making people play for lexical acquisition with the JeuxDeMots prototype, 7. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200883 (cit. on p. 141)

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *59*(4), 567–587. https://doi.org/10.3138/cmlr.59.4.567 (cit. on pp. 8, 31)

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1 (cit. on pp. 32, 66)

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30 (cit. on pp. 7, 313).

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, *13*(2), 549–66 (cit. on pp. 22, 65, 67).

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2019). FlauBERT: Unsupervised language model pre-training for French (cit. on p. 207).

Lee, H., & Lee, J. H. (2015). The effects of electronic glossing types on foreign language vocabulary learning: Different types of format and glossary information. *Asia-Pacific Education Researcher*, *24*(4), 591–601. https://doi.org/10.1007/s40299-014-0204-3 (cit. on p. 15)

Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, *21*(2), 32 (cit. on pp. 15, 65).

Lee, H., Warschauer, M., & Lee, J. H. (2018). Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in

a corpus-based L2 vocabulary learning experiment. *ReCALL*. https://doi.org/10.1017/S0958344018000162 (cit. on p. 65)

Lee, H., Lee, H., & Lee, J. H. (2016). Evaluation of electronic and paper textual glosses on second language vocabulary learning and reading comprehension. *Asia-Pacific Education Researcher*, *25*(4), 499–507. https://doi.org/10.1007/s40299-015-0270-1 (cit. on p. 65)

Lee, J., & Yeung, C. Y. (2018a). Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 4 pp. https://doi.org/10.1109/ICNLSP.2018.8374392 (cit. on pp. 18, 20)

Lee, J., Liu, M., Lam, C. Y., Lau, T. O., Li, B., & Li, K. (2017). Automatic difficulty assessment for Chinese texts. *Proceedings of the IJCNLP 2017, System Demonstrations*, 45–48. https://www.aclweb.org/anthology/I17-3012 (cit. on p. 20)

Lee, J., & Yeung, C. Y. (2018b). Personalizing lexical simplification. *Proceedings of the 27th International Conference on Computational Linguistics*, 224–232. https://www.aclweb.org/anthology/C18-1019 (cit. on pp. 19, 20)

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. manual. https://CRAN.R-project.org/package=emmeans. (Cit. on p. 321)

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*(3), 325–335. https://doi.org/10/cqw94v (cit. on p. 275)

Lété, B. (2004). Manulex : une base de données du lexique écrit adressé aux élèves. In E. Calaque & J. David (Eds.), *Didactique du lexique. Contextes, démarches, supports*. (pp. 241–257). De Boeck Supérieur. (Cit. on pp. 140, 258).

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: a grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(1), 156–166 (cit. on pp. 140, 141, 149).

Levine, A., & Reves, T. (1998). Interplay between reading tasks, reader variables, and unknown word processing. *TESL-EJ*, *3*(2). http://www.tesl-ej.org/wordpress/issues/volume3/ej10/ej10a1/ (cit. on p. 119)

Libben, M. R., & Titone, D. A. (2009). Bilingual lexical access in context: Evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 381–390. https://doi.org/10/drwbpb (cit. on p. 148)

Lin, D. T. A., Pandian, A., & Jaganathan, P. (2018). READ+ vs. READ: Investigating extensive reading and vocabulary knowledge development among Malaysian remedial ESL learners. *Journal of Asia TEFL*, *15*(2), 349–364. https://doi.org/10.18823/asiatefl.2018.15.2.6.349 (cit. on p. 32)

Liu, Y.-T., & Leveridge, A. N. (2017). Enhancing L2 vocabulary acquisition through implicit reading support cues in e-books. *British Journal of Educational Technology*, *48*(1), 43–56. https://doi.org/10.1111/bjet.12329 (cit. on p. 62)

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323 (cit. on p. 42).

Lowie, W., Verspoor, M., & Seton, B. (2010). Conceptual representations in the multilingual mind: A study of advanced Dutch students of English. In M. Pütz & L. Sicola (Eds.), *Converging Evidence in Language and Communication Research* (pp. 135–148). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.13.12low. (Cit. on p. 143)

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x (cit. on pp. 12, 169)

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science*. manual. https://doi.org/10.5281/zenodo.1308157. (Cit. on p. 291)

Lupton, R., & Allwood, J. (2017). Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. *Resources, Conservation and Recycling*, *124*, 141–151. https://doi.org/10/ctgr (cit. on p. 100)

Ma, T., Chen, R., Dunlap, S., & Chen, B. (2016). The effect of number and presentation order of high-constraint sentences on second language

word learning. *Frontiers in Psychology*, *7*, 1396. https://doi.org/10.3389/fpsyg.2016.01396 (cit. on p. 231)

Maddela, M., & Xu, W. (2018). A word-complexity lexicon and a neural readability ranking model for lexical simplification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3749–3760. http://aclweb.org/anthology/D18-1410 (cit. on pp. 18, 20)

Mahdavy, B. (2011). The role of topic familiarity and rhetorical organization of texts in L2 incidental vocabulary acquisition. In Z. Bekirogullari (Ed.), *Proceedings of the 2nd International Conference on Education and Educational Psychology*. Elsevier Science Bv. (Cit. on p. 65).

Makowski, Dominique, Lüdecke, & Daniel. (2019). The report package for R: Ensuring the use of best practices for results reporting. *CRAN*. https://github.com/easystats/report (cit. on p. 291)

Malmasi, S., Dras, M., & Zampieri, M. (2016). LTG at SemEval-2016 Task 11: Complex word identification with classifier ensembles. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 996–1000. http://www.aclweb.org/anthology/S16-1154 (cit. on pp. 128–130)

Malmasi, S., & Zampieri, M. (2016). MAZA at SemEval-2016 Task 11: Detecting lexical complexity using a decision stump meta-classifier. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 991–995. http://www.aclweb.org/anthology/S16-1153 (cit. on pp. 128–130)

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press. (Cit. on pp. 36, 124).

Marello, C. (2012). Word lists in reference level descriptions of CEFR (Common European Framework of Reference for Languages). *Proceedings of the XV Euralex International Congress*, 328–335 (cit. on pp. 138, 198).

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2019). CamemBERT: a tasty French language model. http://arxiv.org/abs/1911.03894 (cit. on pp. 207, 210, 211, 261, 325, 333)

Martínez Martínez, J. M., & Tan, L. (2016). USAAR at SemEval-2016 Task 11: Complex word identification with sense entropy and sentence perplex-

ity. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 958–962. http://www.aclweb.org/anthology/S16-1147 (cit. on p. 128)

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10/cs2djx (cit. on p. 311)

Mc Laughlin, G. H. (1969). SMOG grading—–a new readability formula. *Journal of Reading*, *12*(8), 639–646 (cit. on pp. 9, 103, 137).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10/djsbj6 (cit. on p. 293)

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56). (Cit. on pp. 99, 334).

McLaughlin, B. (1987). *Theories of second language learning*. Arnold. (Cit. on p. 8).

McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, *11*(1), 176–197. https://doi.org/10/f9r32s (cit. on pp. 44, 100)

Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics*, *27*(4), 620–644. https://doi.org/10.1093/applin/amlo30 (cit. on p. 143)

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, *25*(1), 107–130. https://www.aclweb.org/anthology/J99-1003 (cit. on p. 191)

Melby-Lervag, M., & Lervag, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409–433. https://doi.org/10.1037/a0033890 (cit. on pp. 21, 33)

Met Office. (2010–2015). *Cartopy: a cartographic python library with a matplotlib interface*. http://scitools.org.uk/cartopy. (Cit. on p. 100)

Meunier, F., Seigneuric, A., & Spinelli, E. (2008). The morpheme gender effect. *Journal of Memory and Language*, *58*(1), 88–99. https://doi.org/10.1016/j.jml.2007.07.005 (cit. on p. 13)

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. *Proceedings of EMNLP 2004*, 404–411. https://www.aclweb.org/anthology/W04-3252 (cit. on p. 44)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119 (cit. on pp. 294, 298).

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition*, *17*(2), 141–147. https://doi.org/10.3758/BF03197064 (cit. on pp. 131, 174)

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). European Second Language Association. (Cit. on pp. 138, 199, 200).

Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary Studies in First and Second Language Acquisition* (pp. 194–211). Palgrave Macmillan UK. https://doi.org/10.1057/9780230242258_12. (Cit. on p. 199)

Mitkov, R., Pekar, V., Blagoev, D., & Mulloni, A. (2007). Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, *21*(1), 29. https://doi.org/10.1007/s10590-008-9034-5 (cit. on pp. 148, 191, 192)

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*(1), 118–141 (cit. on p. 6).

Montero Perez, M., Van den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739. https://doi.org/10.1016/j.system.2013.07.013 (cit. on p. 6)

Mukherjee, N., Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 Task 11: Identifying complex words in a sentence. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 986–990. http://www.aclweb.org/anthology/S16-1152 (cit. on pp. 104, 128–130, 145)

My, H. N. T., Suzuki, S., & Miyazaki, Y. (2017). Building personalized readability equation and personalized english vocabulary list for continued study. *Proceedings of the 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 791–5. https://doi.org/10.1109/IIAI-AAI.2017.135 (cit. on pp. 19, 20)

Nagata, N. (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals*, *32*(4), 469–479. https://doi.org/10.1111/j.1944-9720.1999.tb00876.x (cit. on p. 64)

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233. https://doi.org/10/chqptj (cit. on p. 29)

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models (R. B. O'Hara, Ed.). *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10/f4pkjx (cit. on p. 275)

Nandiegou, M., & Reboul, S. (2018). *La simplification lexicale comme outil pour faciliter la lecture des enfants dyslexiques et faibles lecteurs*. Mémoire en vue de l'obtention du Certificat de capacité en Orthophonie, Aix Marseille Université. Marseille, France. (Cit. on p. 107).

Nathan, P. (2016). *PyTextRank, a Python implementation of TextRank for text document NLP parsing and summarization*. https://github.com/ceteri/pytextrank/. (Cit. on p. 44)

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759. (Cit. on p. 252)

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy*. Cambridge Univ. Press. (Cit. on p. 138).

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10/c586m8 (cit. on p. 107)

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/S014271640707035X (cit. on pp. 255, 256, 258)

Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience*, *4*(1), 15–22. https://doi.org/10/cck9bk (cit. on p. 117)

O'Donnell, M. E. (2009). Finding middle ground in second language reading: Pedagogic modifications that increase comprehensibility and vocabulary acquisition while preserving authentic text features. *Modern Language Journal*, *93*(4), 512–533. https://doi.org/10.1111/j.1540-4781.2009.00928.x (cit. on p. 64)

Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar*. Paul Treber. (Cit. on p. 104).

Paetzold, G., & Specia, L. (2016a). SemEval 2016 Task 11: Complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 560–569. http://www.aclweb.org/anthology/S16-1085 (cit. on pp. 13, 18–20, 42, 68, 103, 104, 122, 127, 207, 219, 222, 229, 232, 310, 330, 345)

Paetzold, G., & Specia, L. (2016b). SVooogg at SemEval-2016 Task 11: Heavy gauge complex word identification with system voting. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 969–974. http://www.aclweb.org/anthology/S16-1149 (cit. on pp. 127–130)

Paetzold, G. H., & Specia, L. (2016c–December 17). Understanding the lexical simplification needs of non-native speakers of English. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 717–727 (cit. on pp. 103, 282).

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.5526 (cit. on p. 16)

Palakurthi, A., & Mamidi, R. (2016). IIIT at SemEval-2016 Task 11: Complex word identification using nearest centroid classification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1017–1021. http://www.aclweb.org/anthology/S16-1158 (cit. on pp. 128, 129)

Palmero Aprosio, A., Menini, S., & Tonelli, S. (2020). Adaptive complex word identification through false friend detection. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 192–200. https://doi.org/10/ghdhbs (cit. on p. 265)

Paribakht, T. S. (2005). The influence of first language lexicalization on second language lexical inferencing: A study of Farsi-speaking learners of English as a foreign language. *Language Learning*, *55*(4), 701–748. https://doi.org/10.1111/j.0023-8333.2005.00321.x (cit. on p. 14)

Paribakht, T. S., & Wesche, M. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *Canadian Modern Language Review*, *52*(2), 155–178 (cit. on pp. 12, 31).

Parry, K. (1991). Building a vocabulary through academic reading. *TESOL Quarterly*, *25*(4), 629–653 (cit. on p. 6).

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400m neural response. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 38–46. https://www.aclweb.org/anthology/U11-1007 (cit. on pp. 125, 206)

Paulussen, H., Macken, L., Trushkina, J., Desmet, P., & Vandeweghe, W. (2006). Dutch Parallel Corpus: a multifunctional and multilingual corpus. *Cahiers de l'Institut de Linguistique de Louvain*, *32* (cit. on p. 193).

Pawlik, M., & Augsten, N. (2015). Efficient computation of the tree edit distance. *ACM Trans. Database Syst.*, *40*(1), 3:1–3:40. https://doi.org/10/f66sb3 (cit. on p. 191)

Pawlik, M., & Augsten, N. (2016). Tree edit distance: Robust and memory-efficient. *Information Systems*, *56*, 157–173. https://doi.org/10/f76hvp (cit. on p. 191)

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation*. Dulau and Co. (Cit. on pp. 311, 331).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830 (cit. on pp. 100, 334, 339).

Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, *11*(2), 36–58 (cit. on pp. 52, 67).

Peters, E. (2012a). The differential effects of two vocabulary instruction methods on EFL word learning: A study into task effectiveness. *International Review of Applied Linguistics in Language Teaching*, *50*(3), 213–238 (cit. on p. 32).

Peters, E. (2012b). Learning German formulaic sequences: The effect of two attention-drawing techniques. *Language Learning Journal*, *40*(1), 65–79. https://doi.org/10.1080/09571736.2012.658224 (cit. on pp. 64, 67, 232)

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proc. of NAACL*. https://doi.org/10/gft5gf (cit. on p. 207)

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare*, *13*(3), 141–146. https://doi.org/10.1097/XEB.0000000000000050 (cit. on pp. 34, 40)

Pham, M. T., Rajić, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5*(4), 371–385. https://doi.org/10.1002/jrsm.1123 (cit. on p. 34)

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28 (cit. on pp. 14, 61).

Pilán, I., Vajjala, S., & Volodina, E. (2016, March 29). *A readable read: Automatic assessment of language learning materials based on linguistic complexity*. arXiv: 1603.08868 [cs]. http://arxiv.org/abs/1603.08868. (Cit. on p. 11)

Pilán, I., Volodina, E., & Borin, L. (2016). Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *TAL*, *57*(3), 67–91 (cit. on p. 141).

Pilán, I., Volodina, E., & Zesch, T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, 2101–2111 (cit. on p. 141).

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2020, December 30). *Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program)*. arXiv: 2003.12206 [cs, stat]. http://arxiv.org/abs/2003.12206. (Cit. on p. 333)

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer. (Cit. on p. 249).

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 58–81. https://doi.org/10/dcscfv (cit. on p. 274)

Pintard, A., & François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 85–92. https://www.aclweb.org/anthology/2020.readi-1.13 (cit. on pp. 198, 200, 352)

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25 (cit. on p. 66)

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior*, *19*(2), 221–243. https://doi.org/10.1016/S0747-5632(02)00015-8 (cit. on pp. 64, 66)

Popović, M. (2018). Complex word identification using character n-grams. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 341–348. http://www.aclweb.org/anthology/W18-0541 (cit. on pp. 128, 130, 254)

Porter, M. F. (2001, October). *Snowball: A language for stemming algorithms*. http://snowball.tartarus.org/texts/introduction.html. (Cit. on p. 253)

Postma, M. C., Miltenburg, E., Segers, R., Schoen, A., & Vossen, P. T. J. M. (2016). Open Dutch WordNet. *Proceedings of the Eigth Global Wordnet Conference* (cit. on p. 153).

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, *53*(2), 233–284 (cit. on pp. 14, 44, 65).

Pulido, D. (2004a). The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix*, *4*(2), 20–53 (cit. on pp. 44, 65).

Pulido, D. (2004b). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, *54*(3), 469–523. https://doi.org/10.1111/j.0023-8333.2004.00263.x (cit. on p. 65)

Pulido, D. (2007). The effects of topic familiarity and passage sight vocabulary on L2 lexical inferencing and retention through reading. *Applied Linguistics*, *28*(1), 66–86. https://doi.org/10.1093/applin/amlo49 (cit. on pp. 14, 65)

Pulido, D. (2009). How involved are American L2 learners of Spanish in lexical input processing tasks during reading? *Studies in Second Language Acquisition*, *31*(1), 31–58. https://doi.org/10.1017/S0272263109090020 (cit. on p. 65)

Pulido, D., & Hambrick, D. Z. (2008). The virtuous circle: Modeling individual differences in L2 reading and vocabulary development. *Reading in a Foreign Language*, *20*(2), 164–190 (cit. on pp. 14, 65, 66).

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *56*(2), 282–307. https://doi.org/10.3138/cmlr.56.2.282 (cit. on p. 142)

Quijada, M., & Medero, J. (2016). HMC at SemEval-2016 Task 11: Identifying complex words using depth-limited decision trees. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1034–1037. http://www.aclweb.org/anthology/S16-1161 (cit. on pp. 128–130)

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal: Promoting communications on statistics and Stata*, *2*(1), 1–21. https://doi.org/10/ghnj3m (cit. on p. 275)

Rabinovich, E., Tsvetkov, Y., & Wintner, S. (2018). Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, *6*, 329–342. https://doi.org/10/gfzcgp (cit. on pp. 147, 189)

Raptis, H. (1997). Is second language reading vocabulary best learned by reading? *Canadian Modern Language Review*, *53*(3), 566–580 (cit. on pp. 8, 21, 31).

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.10.15rea. (Cit. on pp. 142, 143)

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50 (cit. on pp. 299, 334).

Reynolds, B. L. (2016). The effects of target word properties on the incidental acquisition of vocabulary through reading. *TESL-EJ*, *20*(3), 1–31 (cit. on pp. 49, 62).

Reynolds, B. L., & Bai, Y. L. (2013). Does the freedom of reader choice affect second language incidental vocabulary acquisition? *British Journal of Educational Technology*, *44*(2), E42–E44. https://doi.org/10.1111/j.1467-8535.2012.01322.x (cit. on p. 66)

Riba, P. (2010). *La spécification et la certification pour le francais des niveaux C1 et C2 du Cadre européen commun de référence pour les langues* (Thèse de doctorat). Équipe d'accueil Didactique des langues, des textes et des cultures. Paris, France. (Cit. on p. 198).

Rinsland, H. D. (1945). *A basic vocabulary of elementary school children*. Macmillan. (Cit. on p. 140).

Roberts, T. A. (2008). Home storybook reading in primary or second language with preschool children: Evidence of equal effectiveness for second-language vocabulary acquisition. *Reading Research Quarterly*, *43*(2), 103–130. https://doi.org/10.1598/RRQ.43.2.1 (cit. on p. 6)

Rodríguez-Gómez, P., Martínez-García, N., Pozo, M. A., Hinojosa, J. A., & Moreno, E. M. (2018). When birds and sias fly: A neural indicator of inferring a word meaning in context. *International Journal of Psychophysiology*, *123*, 163–170. https://doi.org/10.1016/j.ijpsycho.2017.09.015 (cit. on p. 117)

Ronzano, F., Abura'ed, A., Espinosa Anke, L., & Saggion, H. (2016). TALN at SemEval-2016 Task 11: Modelling complex words by contextual, lexical and semantic features. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1011–1016. http://www.aclweb.org/anthology/S16-1157 (cit. on pp. 104, 128–130, 145, 283)

Rose, M., & Kitchin, J. R. (2019). Scopus: Scriptable bibliometrics using a Python interface to Scopus. *SSRN Electronic Journal*. https://doi.org/10/gfv4h3 (cit. on p. 95)

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619 (cit. on pp. 14, 62, 288).

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, *57*(2), 165–199. https://doi.org/10.1111/j.1467-9922.2007.00406.x (cit. on pp. 62, 67)

Rouhi, A., & Mohebbi, H. (2012). The effect of computer assisted L1 and L2 glosses on L2 vocabulary learning. *Journal of Asia TEFL*, *9*(2), 1–19 (cit. on p. 64).

Rouhi, A., & Mohebbi, H. (2013). Glosses, spatial intelligence, and L2 vocabulary learning in multimedia context. *3L: The Southeast Asian Journal of English Language Studies*, *19*(2), 75–87 (cit. on p. 15).

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. MIT Press. (Cit. on p. 293).

Sagot, B., & Fišer, D. (2008). Building a free French WordNet from multilingual resources. https://hal.inria.fr/inria-00614708/document (cit. on p. 257)

Saltelli, A., & Annoni, P. (2011). Sensitivity analysis. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1298–1301). Springer. https://doi.org/10.1007/978-3-642-04898-2_509. (Cit. on p. 313)

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72–78 (cit. on pp. 29, 31, 49).

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of challenges in the management of large corpora 3 (CMLC-3)*. IDS. http://rolandschaefer.net/?p=749. (Cit. on pp. 254, 260, 298)

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. C. ( Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)* (pp. 486–493). European Language Resources Association (ELRA). http://rolandschaefer.net/?p=70. (Cit. on pp. 254, 260, 298)

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, 44–49 (cit. on pp. 209, 263).

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780.003. (Cit. on pp. xii, 8, 19, 349, 410, 414)

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x (cit. on p. 7)

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 523–530. https://doi.org/10.3115/1219840.1219905 (cit. on p. 207)

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 57–61 (cit. on p. 99).

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209–31 (cit. on p. 10).

Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2016). Improving text-to-pictograph translation through word sense disambiguation. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 131–135. https://doi.org/10/gf5npr (cit. on p. 153)

Shardlow, M. (2013a). A comparison of techniques to automatically identify complex words. *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 103–109. http://www.aclweb.org/anthology/P13-3015 (cit. on pp. 13, 18, 103, 104, 122, 128, 129, 310)

Shardlow, M. (2013b). The CW Corpus: A new resource for evaluating the identification of complex words. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 69–77. http://www.aclweb.org/anthology/W13-2908 (cit. on pp. 19, 20, 105, 106)

Shardlow, M. (2014). Out in the open: Finding and categorising errors in the lexical simplification pipeline. *LREC 2014*, 1583–1580. https://pdfs.

semanticscholar.org/558d/d65fc9d200cfob5d63b6437e49b782f95cde.pdf (cit. on p. 16)

Shardlow, M., Cooper, M., & Zampieri, M. (2020). CompLex — A new corpus for lexical complexity prediction from likert scale data. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READing DIfficulties (READI)*, 57–62. https://www.aclweb.org/anthology/2020.readi-1.9 (cit. on pp. 122, 127)

Shen, H. H. (2008). An analysis of word decision strategies among learners of Chinese. *Foreign Language Annals*, *41*(3), 501–524. https://doi.org/10.1111/j.1944-9720.2008.tb03309.x (cit. on pp. 14, 67)

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330 (cit. on p. 126)

Simard, M., Foster, G. F., & Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing - Volume 2*, 1071–1082. http://dl.acm.org/citation.cfm?id=962367.962411 (cit. on p. 191)

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press. (Cit. on p. 8).

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10/b6pms8 (cit. on p. 47)

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33). https://escholarship.org/uc/item/69s3541f (cit. on pp. 125, 206)

Soares, A., Medeiros, J., Simões, A., Machado, J., Costa, A., Iriarte S., A., Almeida, J., Pinheiro, A., & Comesaña, M. (2013). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior research methods*, *46*. https://doi.org/10/f6c5dv (cit. on p. 140)

Sobol', I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1-3), 271–280. https://doi.org/10/dkxtdn (cit. on p. 313)

Sonbul, S., & Schmitt, N. (2010). Direct teaching of vocabulary after reading: Is it worth the effort? *ELT Journal*, *64*(3), 253–260. https://doi.org/10.1093/elt/ccp059 (cit. on p. 32)

sp, s., Kumar, A., & K P, S. (2016). AmritaCEN at SemEval-2016 Task 11: Complex word identification using word embedding. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1022–1027. http://www.aclweb.org/anthology/S16-1159 (cit. on pp. 128–130)

Štajner, S., Popovic, M., Saggion, H., Specia, L., & Fishel, M. (2016). Shared task on quality assessment for text simplification. *Proceeding of the Workshop on Quality Assessment for Text Simplification-Lrec*, 22–31 (cit. on p. 310).

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10/b3f9hz (cit. on p. 7)

Steel, R. G. D. (1960). A rank sum test for comparing all pairs of treatments. *Technometrics*, *2*(2), 197–207. https://doi.org/10/br69k3 (cit. on p. 183)

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33–48. https://doi.org/10/gfwsm7 (cit. on p. 4)

Sun, H. (2014). The effects of exposure frequency and contextual richness in reading on Chinese EFL learners' vocabulary acquisition. *Chinese Journal of Applied Linguistics*, *37*(1), 86–106. https://doi.org/10.1515/cjal-2014-0006 (cit. on pp. 62, 64, 231)

Sung, Y.-T., Lin, W.-C., Dyson, S. B., Chang, K.-E., & Chen, Y.-C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, *99*(2), 371–391. https://doi.org/10/f7mwhd (cit. on p. 11)

Swaffar, J. K. (1988). Readers, texts, and second languages: The interactive processes. *Modern Language Journal*, *72*(2), 123–49 (cit. on pp. 30, 33).

Tabatabaei, O., & Shams, N. (2011). The effect of multimedia glosses on online computerized L2 text comprehension and vocabulary learning of Iranian EFL learners. *Journal of Language Teaching and Research*, *2*(3), 714–725. https://doi.org/10.4304/jltr.2.3.714-725 (cit. on p. 15)

Tack, A., François, T., Desmet, P., & Fairon, C. (2017, February 10). *Introducing NT2Lex: A machine-readable CEFR-graded lexical resource for Dutch as a foreign language*. Computational Linguistics in the Netherlands 27 (CLIN 2017), Leuven, Belgium. (Cit. on pp. 137, 141).

Tack, A., François, T., Desmet, P., & Fairon, C. (2018a, July 4). *Making sense of L2 lexical complexity with NT2Lex, a CEFR-graded lexicon linked to Open Dutch WordNet*. The XIXth International Computer Assisted Language Learning (CALL) Research Conference, Bruges, Belgium. (Cit. on p. 137).

Tack, A., François, T., Desmet, P., & Fairon, C. (2018b). NT2Lex: A CEFR-graded lexical resource for Dutch as a foreign language linked to Open Dutch WordNet. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 137–146. https://www.aclweb.org/anthology/W18-0514 (cit. on pp. xii, 21, 137, 141, 164, 352, 410, 414)

Tack, A., François, T., Desmet, P., & Fairon, C. (2019, July 2). *The role of cognate vocabulary in CEFR-based word-level readability assessment*. Vocab@Leuven International Conference, Leuven, Belgium. (Cit. on p. 137).

Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016a–May 28). Evaluating lexical simplification and vocabulary knowledge for learners of French: Possibilities of using the FLELex resource. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 230–236. http://www.lrec-conf.org/proceedings/lrec2016/pdf/544_Paper.pdf (cit. on pp. 18, 104, 134, 141, 164, 208)

Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016b–July 8). Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. *Actes de La 23ème Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN'16)*, 221–

234. https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/Papers/T22.pdf (cit. on pp. 141, 282, 295)

Taylor, A. (2006). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, *23*(2), 309–318. https://doi.org/10.1558/cj.v23i2.309-318 (cit. on pp. 4, 15, 21, 33)

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10/ggwdf2 (cit. on p. 206)

Tharp, J. B. (1939). The measurement of vocabulary difficulty. *The Modern Language Journal*, *24*(3), 169–178. https://doi.org/10.1111/j.1540-4781.1939.tb02893.x (cit. on p. 142)

Tjur, T. (2009). Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, *63*(4), 366–372. https://doi.org/10/d99t5j (cit. on pp. 286, 312, 331, 339)

Tsai, A. (2017). Conceptualizations of vocabulary knowledge in second language reading. *Reading Matrix: An International Online Journal*, *17*(2), 16–39 (cit. on pp. 21, 30, 33).

Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning*, *27*(1), 1–25. https://doi.org/10.1080/09588221.2012.692384 (cit. on pp. 15, 67)

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to english texts based on textual features. *Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference (APCLC 2018)* (cit. on p. 11).

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*. https://doi.org/10/gf4p27 (cit. on pp. 32, 284)

Vahedi, V. S., Ghonsooly, B., & Pishghadam, R. (2016). Vocabulary glossing: A meta-analysis of the relative effectiveness of different gloss types on L2 vocabulary acquisition. *Teaching English with Technology*, *16*(2), 3–25 (cit. on pp. 15, 33).

van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). http://ilk.uvt.nl/downloads/pub/papers/tadpole-final.pdf. (Cit. on p. 152)

Van Eynde, F. (2004). *Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands*. Centrum voor Computerlinguïstiek. KU Leuven, Belgium. (Cit. on p. 152).

Vandenberghe, B., Montero Perez, M., Reynvoet, B., & Desmet, P. (2019). The role of event-related potentials (ERPs) as sensitive measures in L2 vocabulary acquisition research. *Journal of the European Second Language Association*, *3*(1), 35–45. https://doi.org/10/gg467d (cit. on p. 117)

Vanhauwaert, R. (2017). *Identification et analyse de la difficulté lexicale auprès de lecteurs non natifs du français: Le cas des expressions polylexicales* (Master's Thesis). KU Leuven. (Cit. on pp. 218, 231).

Velleman, E., & van der Geest, T. (2014). Online test tool to determine the CEFR reading comprehension level of text. *Procedia Computer Science*, *27*, 350–358. https://doi.org/10.1016/j.procs.2014.02.039 (cit. on p. 11)

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4. (Cit. on p. 291)

Volodina, E., Borin, L., Pilán, I., François, T., & Tack, A. (2017, April). SVALex. En andraspråksordlista med CEFR-nivåer. In E. Sköldberg, M. Andréasson, H. Adamsson Eryd, F. Lindahl, J. Prentice, S. Lindström, & M. Sandberg (Eds.), *Svenskans beskrivning* (pp. 369–382). Göteborgs Universitet. https://gupea.ub.gu.se/bitstream/2077/52211/1/gupea_2077_52211_1.pdf. (Cit. on pp. 139, 141)

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016). SweLLex: second language learners' productive vocabulary. *Proceedings of the Joint 5th NLP4CALL and 1st NLP4LA Workshops (SLTC 2016)*, 76–84. http://www.ep.liu.se/ecp/130/ecp16130.pdf (cit. on p. 141)

Vossen, P., Görög, A., Izquierdo, R., & den Bosch, A. V. (2012). DutchSemCor: Targeting the ideal sense-tagged corpus. In N. C. ( Chair), K. Choukri, T.

Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 584–589). European Language Resources Association (ELRA). (Cit. on p. 153).

Vossen, P., Maks, I., Segers, R., van der Vliet, H., Moens, M.-F., Hofmann, K., Sang, E. T. K., & de Rijke, M. (2013). Cornetto: A combinatorial lexical semantic database for Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch* (pp. 165–184). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_10. (Cit. on p. 153)

Wang, M., & Koda, K. (2005). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning*, *55*(1), 71–98. https://doi.org/10.1111/j.0023-8333.2005.00290.x (cit. on p. 6)

Wang, Y.-H. (2016). Promoting contextual vocabulary learning through an adaptive computer-assisted EFL reading system. *Journal of Computer Assisted Learning*, *32*(4), 291–303. https://doi.org/10.1111/jcal.12132 (cit. on pp. 12, 62, 121)

Wani, N., Mathias, S., Gajjam, J. A., & Bhattacharyya, P. (2018). The whole is greater than the sum of its parts: Towards the effectiveness of voting ensemble classifiers for complex word identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 200–205. http://www.aclweb.org/anthology/W18-0522 (cit. on pp. 128–130, 145)

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2) (cit. on pp. 14, 67).

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *19*(3), 287–307 (cit. on pp. 14, 64, 66).

Watts, M. L. (2008). Clause type and word saliency in second language incidental vocabulary acquisition. *The Reading Matrix*, *8*(1), 1–22 (cit. on p. 65).

Weaver, W. (1955). Translation. *Machine translation of languages*, *14*(15-23), 10 (cit. on p. 17).

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/amlo48 (cit. on pp. 14, 37, 49, 62, 288)

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245. https://eric.ed.gov/?id=EJ815123 (cit. on pp. 14, 64, 231)

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40 (cit. on pp. 55, 62).

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data*. manual. https://CRAN.R-project.org/package=tidyr. (Cit. on p. 291)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6 (cit. on pp. 207, 333, 334)

Woodsend, K., & Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 409–420. https://www.aclweb.org/anthology/D11-1038 (cit. on p. 105)

Wróbel, K. (2016). PLUJAGH at SemEval-2016 Task 11: Simple system for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 953–957. http://www.aclweb.org/anthology/S16-1146 (cit. on pp. 127, 129, 130)

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, *3*(0), 283–297. https://transacl.org/ojs/index.php/tacl/article/view/549 (cit. on p. 105)

Yancey, K., & Lepage, Y. (2018). Korean L2 vocabulary prediction: Can a large annotated corpus be used to train better models for predicting un-

known words? In N. C. ( chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). (Cit. on p. 20).

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48–67 (cit. on pp. 15, 119).

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, *31*(3), 251–267. https://doi.org/10/gj9zzt (cit. on p. 4)

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10/cfqrdn (cit. on p. 255)

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A report on the complex word identification shared task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78. https://www.aclweb.org/anthology/W18-0507 (cit. on pp. 13, 18, 20, 68, 122, 123, 127, 207, 219, 229, 232, 310, 330, 345)

Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017a). CWIG3G2 - Complex word identification task across three text genres and two user groups. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 401–407. http://www.aclweb.org/anthology/I17-2068 (cit. on pp. 19, 20, 128–130)

Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C. (2017b). Multilingual and cross-lingual complex word identification. *Proceedings of Recent Advances in Natural Language Processing*, 813–822. https://doi.org/10.26615/978-954-452-049-6_104 (cit. on pp. 20, 128–130)

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*(3), 85–101 (cit. on pp. 15, 37, 64).

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*(1), 39–58. https://doi.org/10.1080/09588221.2010.523285 (cit. on pp. 4, 12, 15, 21, 33)

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *57*(4), 541–572. https://doi.org/10.3138/cmlr.57.4.541 (cit. on p. 284)

Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, 59–63. http://www.aclweb.org/anthology/W17-5910 (cit. on pp. 13, 289, 347, 348)

Zampieri, M., Tan, L., & van Genabith, J. (2016). MacSaar at SemEval-2016 Task 11: Zipfian and character features for complex word identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1001–1005. http://www.aclweb.org/anthology/S16-1155 (cit. on pp. 128, 129, 254)

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide* (Touchstone Applied Science Associates). (Cit. on p. 140).

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 649–657 (cit. on p. 300).

Zhao, A., Guo, Y., Biales, C., & Olszewski, A. (2016). Exploring learner factors in second language (L2) incidental vocabulary acquisition through reading. *Reading in a Foreign Language*, *28*(2), 224–245 (cit. on pp. 14, 65).

Zimmerman, C. B. (1997). Do reading and interactive vocabulary instruction make a difference? An empirical study. *TESOL Quarterly*, *31*(1), 121–140. https://doi.org/10.2307/3587978 (cit. on p. 32)

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley. (Cit. on p. 168).

# SUMMARIES

# SUMMARY IN FRENCH

## NOTEZ MES MOTS ! SUR LA PRÉDICTION AUTOMATIQUE DE LA DIFFICULTÉ LEXICALE POUR LES LECTEURS DE LANGUES ÉTRANGÈRES

L'objectif de cette recherche doctorale est la prédiction automatique des mots difficiles dans un texte pour les locuteurs non natifs. Cette prédiction est cruciale car une bonne compréhension d'un texte est fortement déterminée par le vocabulaire. Si un texte contient un pourcentage élevé de mots inconnus, le lecteur aura probablement des difficultés à comprendre le texte. Afin de fournir un bon soutien au lecteur de langue étrangère, nous devons d'abord être en mesure de prédire le nombre de mots difficiles. En général, nous le faisons manuellement en nous basant sur notre expertise ou sur des tests de vocabulaire antérieurs. Cependant, ces méthodes ne sont pas pratiques lorsque nous lisons dans un environnement informatique tel qu'une tablette ou une plateforme d'apprentissage en ligne. Dans ces cas, nous devons automatiser correctement les prédictions.

La thèse est divisée en trois parties. La première partie contient un examen systématique de la littérature scientifique pertinente. La synthèse comprend 50 ans de recherche et 140 publications évaluées par des pairs sur la prédiction statistique de la compétence lexicale chez les lecteurs non natifs. Les analyses montrent, entre autres, que le champ scientifique est divisé en deux domaines de recherche peu connectés. D'une part, il existe une longue tradition de recherche expérimentale en matière d'acquisition de langues étrangères (SLA) et d'apprentissage des langues assisté par ordinateur (CALL). Ces études expérimentales testent principalement l'effet de certains facteurs (par exemple, la répétition de mots difficiles ou l'ajout de glosses électroniques) sur l'apprentissage de mots non familiers pendant la lecture. D'autre part, des études récentes sur le traitement du langage naturel (NLP) s'appuient sur l'intelligence artificielle pour prédire automatiquement les mots difficiles.

En outre, l'étude de la littérature met en évidence certaines limites qui ont été approfondies dans le cadre de cette recherche doctorale. La première limite est le manque de mesures et de prédictions contextualisées. Bien que la recherche nous ait appris que le contexte dans lequel un mot apparaît est un facteur important, les prédictions sont souvent faites sur la base de tests de vocabulaire isolés, entre autres. La deuxième limite est le manque de mesures et de prédictions personnalisées. Bien que la recherche sur l'acquisition des langues étrangères ait montré qu'il existe de nombreuses différences entre les lecteurs non natifs, des études récentes en intelligence artificielle font des prédictions basées sur des données agrégées. La dernière limite est que la majorité des études (74%) se concentrent sur l'anglais en tant que langue étrangère. L'objectif de cette recherche doctorale est donc une approche contextualisée et personnalisée et une focalisation sur le néerlandais et le français comme langues étrangères.

La deuxième partie examine deux mesures de la difficulté lexicale pour les lecteurs non natifs. D'une part, elle étudie la manière dont les mots sont introduits dans les matériels de lecture didactique étiquetés avec les niveaux du CECR. Cette étude introduit une nouvelle base de données lexicale graduée pour le néerlandais, à savoir NT2Lex (Tack et al., 2018b). La caractéristique innovante de cette base de données est que la fréquence par niveau de difficulté a été calculée pour le sens de chaque mot, désambiguïsé sur la base du contexte de la phrase. Cependant, les résultats montrent qu'il existe d'importantes incohérences dans la manière dont les traductions étymologiquement liées apparaissent dans les bases de données néerlandaise et française. Par conséquent, cette mesure de difficulté ne semble pas encore valable comme base pour un système automatisé.

D'autre part, on étudie comment les locuteurs non natifs perçoivent les mots difficiles pendant la lecture. La perception de la difficulté est importante à prévoir car l'attention de l'apprenant est un facteur déterminant dans le processus d'apprentissage (Schmidt, 2001). L'étude introduit de nouvelles données pour les lecteurs du français. Un objectif important de ces données est de faire des prédictions correctes pour tous les mots du texte, ce qui contraste avec les études sur l'acquisition des langues étrangères qui se concentrent sur un nombre limité (*Mdn* = 22) de mots cibles dans le texte. De plus, les analyses

montrent que les données peuvent être utilisées pour développer un système personnalisé et contextualisé.

La dernière section examine deux types de modèles prédictifs développés sur les données susmentionnées, à savoir les modèles à effets mixtes et les réseaux neuronaux artificiels. Les résultats valident l'idée que la perception de la difficulté lexicale peut être prédite principalement sur la base de la "surprise des mots", un concept central de la théorie de l'information. En outre, les analyses montrent que les statistiques de performance couramment utilisées (telles que la précision et le F-score) sont sensibles aux différences individuelles dans les taux de difficulté. Comme ceux-ci ne sont donc pas appropriés pour comparer les prédictions correctes pour différents apprenants, les coefficients D et Phi sont utilisés. De plus, les résultats montrent clairement qu'un modèle personnalisé fait des prédictions nettement meilleures qu'un modèle non personnalisé. D'autre part, les résultats montrent qu'un modèle contextualisé peut mieux discriminer la difficulté, bien que ces améliorations ne soient pas toujours significatives pour chaque apprenant.

# SUMMARY IN DUTCH

## LET OP MIJN WOORDEN! OVER DE GEAUTOMATISEERDE VOORSPELLING VAN LEXICALE MOEILIJKHEID VOOR LEZERS VAN VREEMDE TALEN

Het doel van dit doctoraatsonderzoek is het automatisch voorspellen van moeilijke woorden in een tekst voor anderstaligen. Deze voorspelling is cruciaal omdat een goed tekstbegrip sterk wordt bepaald door woordenschat. Als een tekst een te hoog percentage onbekende woorden bevat, zal de lezer deze tekst waarschijnlijk met moeite begrijpen. Om de anderstalige lezer een goede ondersteuning te bieden, moeten we eerst het aantal moeilijke woorden kunnen voorspellen. Meestal doen we dit handmatig op basis van expertise of voorafgaande woordenschattesten. Dergelijke methoden zijn echter niet praktisch wanneer we lezen in een computergebaseerde omgeving zoals bijvoorbeeld een tablet of een online leerplatform. In deze gevallen moeten we de voorspellingen op een correcte manier automatiseren.

De scriptie is opgedeeld in drie delen. Het eerste deel bevat een systematische studie van de relevante wetenschappelijke literatuur. De synthese omvat 50 jaar onderzoek en 140 peer-reviewed publicaties over het statistisch voorspellen van lexicale competentie in anderstalige lezers. De analyses tonen onder meer aan dat het wetenschappelijk bereik opgedeeld is in twee onderzoeksvelden die weinig met elkaar verbonden zijn. Enerzijds is er een lange traditie van experimenteel onderzoek in vreemdetaalverwerving (SLA) en computerondersteund taalonderwijs (CALL). Deze experimentele studies toetsen voornamelijk het effect van bepaalde factoren (bv. het herhalen van moeilijke woorden of het toevoegen van elektronische glossen) op het leren van ongekende woorden tijdens het lezen. Anderzijds zijn er recente studies in natuurlijke taalverwerking (NLP) die beroepen op artificiële intelligentie om moeilijke woorden automatisch te voorspellen.

Bovendien wijst de literatuurstudie op enkele beperkingen die in dit doctoraatsonderzoek verder bestudeerd werden. De eerste beperking is het tekort aan gecontextualiseerde maten en voorspellingen. Hoewel we weten uit onderzoek dat de context waarin een woord voorkomt een belangrijke factor is, worden voorspellingen vaak gemaakt op basis van onder meer geïsoleerde woordenschattesten. De tweede beperking is het tekort aan gepersonaliseerde maten en voorspellingen. Hoewel onderzoek in vreemdetaalverwerving aangetoond heeft dat er veel verschillen zijn tussen anderstalige lezers, maken recente studies in artificiële intelligentie voorspellingen op basis van geaggregeerde data. De laatste beperking is dat het merendeel van studies (74%) focust op Engels als vreemde taal. Het doel van dit doctoraatsonderzoek is bijgevolg een gecontextualiseerde en gepersonaliseerde aanpak en een focus op Nederlands en Frans als vreemde taal.

Het tweede deel bekijkt twee maten van lexicale moeilijkheid voor anderstalige lezers. Enerzijds wordt er onderzocht hoe woorden worden geïntroduceerd in didactisch leesmateriaal gelabeld met ERK niveaus. Deze studie introduceert een nieuwe gegradueerde lexicale databank voor Nederlands, namelijk NT2Lex (Tack e.a., 2018b). Het vernieuwende aan deze databank is dat de frequentie per moeilijkheidsniveau werd berekend voor de betekenis van elk woord, gedisambigueerd op basis van de zinscontext. De resultaten tonen echter aan dat er belangrijke inconsistenties zijn in hoe etymologisch verwante vertalingen voorkomen in de Nederlandse en Franse databanken. Daarom lijkt deze moeilijkheidsmaat nog niet valide als basis voor een geautomatiseerd systeem.

Anderzijds wordt er onderzocht hoe anderstaligen zelf moeilijke woorden percipiëren tijdens het lezen. De perceptie van moeilijkheid is belangrijk te voorspellen want de aandacht van de leerder is een bepalende factor in het leerproces (Schmidt, 2001). De studie introduceert nieuwe data voor lezers van Frans. Een belangrijk doel van deze data is om correcte voorspellingen te doen voor alle woorden in de tekst, wat contrasteert met veel studies in vreemdetaalverwerving die focussen op een beperkt aantal (*Mdn* = 22) doelwoorden in de tekst. Bovendien tonen de analyses dat de data kunnen gebruikt worden om een gepersonaliseerd en gecontextualiseerd systeem te ontwikkelen.

Het laatste deel bekijkt twee types voorspellende modellen die op voorgenoemde data werden ontwikkeld, namelijk mixed-effects modellen en artificiële neurale netwerken. De resultaten bekrachtigen de idee dat de perceptie van lexicale moeilijkheid voornamelijk kan worden voorspeld op basis van "word surprisal", een centraal begrip in de informatietheorie. Verder tonen de analyses aan dat de veelgebruikte prestatiestatistieken (zoals accuraatheid en F-score) gevoelig zijn aan individuele verschillen in percentages van moeilijkheid. Omdat deze daarom niet gepast zijn om voorspellingen correct te vergelijken voor verschillende leerders, worden de D en Phi coëfficiënten gebruikt. Bovendien tonen de resultaten duidelijk aan dat een gepersonaliseerd model significant betere voorspellingen maakt dan een niet-gepersonaliseerd model. Anderzijds tonen de resultaten aan dat een gecontextualiseerd model moeilijkheid beter kan discrimineren, alhoewel deze verbeteringen niet altijd significant zijn voor elke leerder.