

Individual Differences and Lexical Difficulty

Towards Personalised Difficulty Predictions

Camille Pirali

Supervisor: Thomas François

Co-supervisor: Núria Gala



Dissertation towards a Master's Degree in Linguistics,

Specialisation in Natural Language Processing

FACULTÉ DE PHILOSOPHIE, ARTS ET LETTRES

UCLouvain

Academic Year: 2021-2022 (June)

Acknowledgements

I would like to express my gratitude to those whose help and support contributed to this dissertation.

To my supervisors, Thomas François and Núria Gala, who guided me through the process from the start and offered invaluable advice. Thank you for your patience and availability, as well as for your helpful ideas and remarks.

To Hubert Naets, for helping me set up the reading platform online and answering my many questions, as well as to the students who participated in the study and to the FFL teachers at the SUL and ILV who asked them to - none of this would have been possible without your help. Special thanks go to Paul, who went out of his way to help with finding participants and who agreed to pilot the study with his own class.

To my parents, who did their best to keep my perfectionism in check and whose gentle reminders to not overdo did not go unheard. Your encouragements and celebrations of even my smallest accomplishments gave me the confidence I needed to undertake this project.

To Dan, whose unwavering support and pep talks kept me going when I was struggling to see the end. Thank you for your precious feedback and your insightful comments - they were much appreciated. And finally, to Camille, Emma and Antoine, for keeping me sane by forcing me to go outside and enjoy student life every once in a while. You made my last year of university all the more special.

Contents

Acknowledgements	ii
List of Abbreviations	vi
Introduction	1
1 Literature Review	5
1.1 Introduction	5
1.2 Fundamental Constructs	6
1.2.1 The Word	6
1.2.1.1 Definition	6
1.2.1.2 Multiword Expressions	8
1.2.2 Lexical Difficulty	10
1.2.2.1 Difficulty vs Complexity	10
1.2.2.2 Relevance of the Concept	11
1.2.2.3 Practical implications	11
1.3 Vocabulary and Foreign Language Acquisition	12
1.3.1 Vocabulary Knowledge	13
1.3.1.1 Definition	13
1.3.1.2 Assessing Vocabulary Knowledge	15
1.3.2 Foreign Language Acquisition	16
1.3.2.1 Main Factors	16
1.3.2.2 Impact of the L1	17
1.4 Measuring Lexical Complexity from Annotation Data	20

1.4.1	Before 2016	20
1.4.2	CWI-2016	21
1.4.2.1	Best ranking teams in terms of G-score	22
1.4.2.2	Best ranking teams in terms of F-score	24
1.4.2.3	Discussion	25
1.4.3	CWI-2018	26
1.4.3.1	Description of the best-performing systems	27
1.4.3.2	Discussion	29
1.4.4	LCP-2021	31
1.4.4.1	Description of the best-performing systems	33
1.4.4.2	Discussion	34
1.4.5	General Discussion: Insights from the Shared Tasks	36
1.5	Measuring Lexical Difficulty	39
1.5.1	Overview of Studies	39
1.5.1.1	Personalised Familiarity Prediction	39
1.5.1.2	Personalised Vocabulary Prediction	41
1.5.1.3	Personalised Lexical Simplification	43
1.5.2	Discussion	46
1.6	Conclusion	47
2	Methodology	50
2.1	Introduction	50
2.2	Data Collection Process	52
2.2.1	Demographic Questionnaire	52
2.2.1.1	Numerical and Ordinal Variables	53
2.2.1.2	Nominal Variables	53
2.2.1.3	Hybrid Variable	54
2.2.2	Text Annotation	54
2.2.2.1	Corpus Design	54
2.2.2.2	Random Text Selection	55

2.2.2.3	Lexical Units	55
2.2.2.4	Annotation Process	56
2.2.2.5	Annotation Scale	59
2.2.3	Comprehension Questions	62
2.2.3.1	Question Design	62
2.2.3.2	Feedback	63
2.2.4	Finding Participants	63
2.2.5	Discussion	64
2.3	Data Analysis	65
3	Results	68
3.1	Description of the Data	68
3.1.1	Participants	69
3.1.2	Annotated Texts	72
3.1.3	Comprehension Questions	74
3.1.4	Discussion	77
3.2	Agreement Scores	78
3.3	Ordinal Regression Models	83
3.3.1	Baseline Models	84
3.3.2	Learner Models	85
3.3.3	Individual Differences Models	88
3.3.3.1	Full Dataset	88
3.3.3.2	Partial Dataset	90
3.3.4	Discussion	93
	Conclusion	96
	Bibliography	100

List of Abbreviations

AIC	Akaike Information Criterion
CEFR	Common European Framework of Reference
CLM	Cumulative Link Model
CLMM	Cumulative Link Mixed Model
CWI	Complex Word Identification
FFL	French as a Foreign Language
LCP	Lexical Complexity Prediction
LS	Lexical Simplification
MCQ	Multiple Choice Question
MLM	Multinomial Log-Linear Model
MWE	Multiword Expression
NLP	Natural Language Processing
RH	Research Hypothesis
RQ	Research Question
SVM	Support Vector Machine
TLFi	Trésor de la Langue Française informatisé
US	United States
VKS	Vocabulary Knowledge Scale
WAT	Word Associates Test

Introduction

When learning a new language, reading can be both an enjoyable and efficient way of building vocabulary. This is especially true if learners are exposed to texts that are suited to their proficiency level and lexical knowledge. Indeed, it has been shown that extensively reading texts of an adequate complexity was preferable to studying written works in depth [4]. Providing learners with appropriate materials is therefore highly important in fostering their language development. If a text is too simple, the reader will not be able to acquire any new linguistic knowledge; too difficult, and their global comprehension will be insufficient for contextual guessing and incidental learning of new vocabulary to occur [46, 35]. This concern does not solely apply to foreign language learners, but also to struggling readers, low-literacy adults and children who are still acquiring their L1. In some cases, the need for accessible texts is particularly critical: people who struggle to find the motivation to read can enter into a vicious cycle which makes them less and less likely to improve as readers [10].

Considering the benefits of giving people access to the right reading materials, it is perhaps no surprise that this issue has attracted growing interest from the Natural Language Processing (NLP) community. For instance, several readability measures have been created, and could help learners or language teaching professionals select texts of an appropriate level. Additionally, much work has been undertaken on automatic text simplification systems, which can adapt a pre-existing text to render it more accessible. This is achieved by making adjustments on the lexical, morphological, syntactic or discursive levels [29]. In the case of lexical simplification,

the first step is to determine which words need to be substituted with an easier equivalent [83]. The current methods used to identify complex words are typically reader-independent, or at best adapted to learners of a certain proficiency level. The profile of the people who will be reading the simplified text is generally not taken into account by the models.

Lexical difficulty, however, is a subjective notion: a word might be hard to understand for one person, but very familiar to another. Adult native speakers, children, language learners and people with a reading impairment are likely to have different simplification needs from one another, as they differ in their levels of exposure to the target language as well as in their cognitive abilities. Even within one of those groups, considerable variability in difficulty perceptions can occur. This is due to what psychologists call “individual differences”, which can take the form of “physical abilities, mental abilities, knowledge, habits, personality, and character traits.” [98, p. 57]. From innate, biological characteristics to factors rooted in someone’s environmental and psychological backgrounds, individual differences refer to all the ways in which a person can distinguish themselves from their peers. As no two people are identical, the ideal approach would be to create fully-personalised lexical simplification tools, which would learn the needs of a particular reader and adapt written materials accordingly. While some studies have explored this possibility and obtained highly encouraging results, such an approach is not always feasible. A foreign language teacher, for instance, might want to select a text adapted to their class as a whole, without having the luxury of choosing a different text for each student. Moreover, fully-personalised models need to be trained on difficulty data from the target reader, which is not always an option. A middle-ground alternative would thus be to identify difficult words based on some profile information about the intended readers, such as their age, mother tongue, and proficiency level in the language. In other words, lexical difficulty predictions could be personalised on the basis of individual differences - this will be the focus of what follows.

The ambitions of this dissertation are twofold: 1) to provide a theoretical framework and guidelines for lexical difficulty prediction based on an extensive review of the literature, and 2) to explore the usefulness of some individual predictors. To achieve this second objective, difficulty annotations in short informative texts were gathered in a study targeting learners of French as a foreign language. The participants were asked to provide some demographic information, which will constitute our set of predictors. Statistical analyses and regression models using those predictors will allow us to uncover some individual differences which have an impact on the perception of lexical difficulty. Moreover, after each text, the readers were tasked with answering a series of multiple choice questions, in order to explore the link between lexical difficulty and global text comprehension. The research questions which this study will attempt to answer are the following:

Research Question 1

Is it worth taking demographic information into account when predicting lexical difficulty for a given reader? In other words, is there more between-group variability as regards perception of difficulty than there is within-group?

Research Question 2

If the answer to RQ 1 is “yes”, which individual variables have a demonstrable impact on lexical difficulty perception?

Research Question 3

Is there a direct link between the proportion of words perceived as difficult by a given reader and the quality of their text comprehension?

The remainder of this dissertation will be organised as follows. Chapter 1 will present some fundamental constructs underlying the notion of lexical difficulty,

before exploring the process of language learning and the factors which impact vocabulary acquisition. It will then provide an extensive overview of previous attempts to identify complex or difficult words on the basis of reader annotations. Building on the conclusions drawn from those past approaches, Chapter 2 will outline the methodological choices made when designing the study collecting difficulty annotations from French L2 learners. The way in which the collected data were analysed will also be described. Chapter 3 will then discuss the results obtained from those analyses, and will endeavour to answer the research questions presented above. Finally, a review of the conclusions drawn from this dissertation will be made, and some limitations of the study will be addressed. Concluding remarks as well as future avenues for research will then be suggested.

Chapter 1

Literature Review

1.1 Introduction

What causes a word to be perceived as difficult by someone reading in a foreign language is a complex question. Among other things, answering it requires to investigate the way lexical information is stored in the brain, the manner in which new words are processed when reading, as well as the variety of factors that affect vocabulary learning and second language acquisition. Studies that have been devoted to lexical difficulty or complexity (two terms referring to different concepts, as will be explained in section 1.2.2.1) have taken many forms, from psycholinguistic explorations of the mental lexicon to NLP-based predictive measures. This literature review aims to provide a summary of the research on lexical difficulty and the main findings that have emerged from it.

Section 1.2 will provide working definitions for the main concepts which will be used in this dissertation, namely those of “word”, “multiword expression”, “difficulty” and “complexity”. Section 1.3 will then explore the questions of vocabulary acquisition and foreign language learning, in order to achieve a better understanding of the mechanisms at play. An overview of studies on the task of Complex Word Identification (CWI) will be provided in section 1.4, as their findings should prove

useful for measures of difficulty as well. Finally, section 1.5 will be devoted to a summary of previous attempts at predicting lexical difficulty, and will show that although there have not yet been many, their results are highly encouraging.

1.2 Fundamental Constructs

1.2.1 The Word

1.2.1.1 Definition

The present study focuses on difficulty at the lexical level, *i.e.* the level of the word or lexical unit (the two terms will be used interchangeably in this dissertation). Although seemingly simple, the concept of word is one that has been much debated in the literature and has been defined in a myriad of ways. In its simplest sense, a word is a string of letters delimited by spaces or punctuation signs [9]. Although used in a variety of tasks, such as measuring the length of a text, this definition of the word as an orthographic unit is very limited. Besides the fact that not all languages have a written form, defining a word based on its spelling is an arbitrary and often irregular criterion. Under this orthographic definition, “bed frame” would be considered as two words while “bedstead” would only be seen as one, although the terms are synonymous and share a similar prosody (with the stress falling on the first syllable).

The mention of synonymy hints at a second possible definition: a word is a minimum meaningful unit [9]. However, in language, that role is already fulfilled by morphemes [15]. These can be either free or bound, depending on whether they can stand alone as a word. Hence, while all free morphemes can serve as words (“box”, “door”), not all words are made up of a singular morpheme. Indeed, they can be created by compounding free morphemes (“lunch+box”, “door+way”) or by combining any number of free and bound morphemes (“*un*+box+*ing*”, “out+door+*s*”). The definition mentioned above should thus be modified slightly: a word is a min-

imal *free* semantic unit. While this second interpretation allows to recognise that both “bed frame” and “bedstead” are lexical units, it poses a problem for function words, whose role is to structure the discourse rather than to carry meaning [9]. The status of units such as “of”, “for” and “to” is thus unclear.

A third definition which solves these limitations is the one proposed by Lyons, who conceptualised words as linguistic units showing a high degree of internal cohesion [53]. A lexical unit is thus characterised by the fact that its components cannot be reordered or interrupted by new elements without its meaning being altered. This definition holds for lexical units containing several orthographic forms as well as for function words. For instance, “in spite of” is rightfully accepted as one word, as it is internally stable (“*of spite in”, “*spite of in”) and uninterruptible (“in great spite of” is grammatical, but has a different, more literal meaning). Unfortunately, this conception of the word is also limited, as yet again examples can be found that do not meet its requirements. Phrasal verbs, for instance, can be interrupted by a pronoun (“pick it up”), while still forming a semantic unit with some level of internal cohesion.

Finding a definition that holds true for all different types of words is thus no easy task. While solving it is not the aim of this dissertation, it is still necessary to formulate a working definition in order to make clear what is meant by “word” or “lexical unit”. As the study undertaken here investigates lexical difficulty in written texts for non-native speakers of French, the semantic criterion (a word is a unit of meaning) seems to be essential. Indeed, understanding larger semantic units (the sentence, the paragraph, the text) requires being able to understand their building blocks. As shown above, these can extend wider than the orthographic word: for instance, the French expression “pierre de touche” (“touchstone”, “yardstick”) is made up of three units which many learners should be familiar with (“stone”, “of”, “touch”), and yet possesses a unique, non-compositional meaning likely to pose problems to readers who have never been exposed to the term.

With these considerations in mind, the following definition was developed: for the purposes of this study, a word is a minimal free form which plays either a semantic or functional role and which can be composed of one or several morphemes. When those morphemes correspond to several graphical words, the term will be referred to as a multiword expression (MWE). The way this definition was operationalised in order to segment the texts used in this study will be further discussed in section 2.2.2.3.

1.2.1.2 Multiword Expressions

As mentioned in the previous section, multiword expressions are defined in this work as lexical units (*i.e.* minimal semantic or functional free forms) which are made up of several graphical components. However, just as was the case for the concept of word, many different definitions of these phraseological units can be found in the literature, with a variety of names being used to refer to them. For English, Linda Verstraten defines “fixed phrases” as “any stable unit made up of at least two words” [95], and Calzolari *et al.* describe “multi-word expressions” as “a sequence of words that acts as a single unit at some level of linguistic analysis” [8]. For French, André Martinet calls “synthèmes” those complex linguistic units that function as a simplex (“unité linguistique signifiante, désignant une notion bien définie, mais où la forme permet de distinguer des éléments successifs porteurs au départ de sens distinctifs”) [57], while Mathieu Constant says of the “expressions multi-mots” that they “forment des unités linguistiques qui contiennent un certain degré de non-compositionalité lexicale, syntaxique, sémantique et/ou pragmatique” [14].

Although the terms used vary, the literature seems to agree on a series of central characteristics of multiword expressions: they are made up of several graphical words, have a unique and stable meaning (or function), and show some degree of morphological, syntactical and/or paradigmatic fixity [43, 57, 70]. Another defining characteristic of those expressions is their semantic non-compositionality: the meaning of a given sequence cannot be inferred from the sum of the meaning of each of

its constituents. For instance, the French expressions “fleur bleue” and “peur bleue” both use the adjective “blue” to evoke very different images: a sentimental, naive personality in the first case, and a deep, intense fear in the second. This is due to a process of lexical fixation, through which the words composing an expression progressively shed their individual meaning to eventually form a new, unified semantic entity [43]. Therein lies their difference with collocations (“heavy rain”) and linguistic clichés (“nice to meet you”), whose meaning remains semantically compositional despite the word association being somewhat fixed [71]. This distinction was also pointed out by Sag *et al.* [78], who used it to divide MWEs into two categories: lexicalised and institutionalised phrases. The former are at least partially non-compositional and are what is meant by “multiword expression” in this dissertation; the latter will be considered as distinct words. For example, “prendre conscience” (to realise, to become aware) would be segmented as a single lexical unit, whereas “prendre une décision” (to make a decision) would be split into three.

As discussed by Sag *et al.*, multiword expressions are a troublesome area of language for the development of natural language processing tools [78]. However, taking them into account in the definition of the lexical unit used in this study seems essential. Indeed, formulaic expressions are an integral part of language, and any learner is therefore likely to encounter them when reading. In a study investigating the proportion of “prefabricated language” in written and spoken texts, Erman and Warren found that, on average, about 55% of the words occurred as part of what they called a prefab [23]. Moreover, it has been shown that, although they present a processing advantage for native speakers of a language, multiword expressions tend to be more difficult for learners, probably because of their non-compositional nature. For instance, an eye-tracking study revealed that figurative idioms are processed faster than non-formulaic language by native readers, while they are read slower than idioms with a literal meaning by non-natives [86]. Another study implementing a self-paced reading task confirmed this processing advantage for natives and suggested that it might also benefit proficient L2 readers [13].

These elements go to show that a reading task asking participants to annotate difficult words needs to account for multiword expressions, as they are common in natural language and have a proven impact on reading (which can either be positive or negative depending on the reader’s proficiency level).

1.2.2 Lexical Difficulty

1.2.2.1 Difficulty vs Complexity

The distinction between difficulty and complexity is not always made in the literature, to the point where the two terms are often used interchangeably. Moreover, L2 studies focusing on one or the other often only provide a vague definition of the concept, if any, and use it to refer to a wide range of phenomena [7]. Among several other usages, the terms “difficulty” and “complexity” have for instance been regarded as synonyms of “readability” [60] and as ways of referring to the sophistication of L2 learner production or to the depth and breadth of their vocabulary knowledge [39, 7].

This work will follow the theory of Bulté and Housen [7], according to whom complexity in a foreign language can be envisioned either in relative or in absolute terms. Relative complexity, or *difficulty*, is defined on an individual basis: a feature is difficult for a learner if its processing costs are high. The mental effort required to process a word is determined by objective factors, one of which being the inherent complexity of that word, as well as by subjective factors such as the learner’s proficiency level in the L2, their mother tongue or their degree of motivation. As a result, difficulty is learner-dependent: a word that is difficult for one person could very well be easy for another. *Complexity*, by contrast, is absolute, speaker-independent and defined by inherent characteristics of the lexical item. Although there is a link between complexity and difficulty, the former is only one of the factors influencing the latter, and a word being complex is no guarantee that it will prove difficult for a given reader.

1.2.2.2 Relevance of the Concept

Numerous studies have shown that there is a close link between reading comprehension and vocabulary coverage [44, 34, 82]. It is broadly accepted today that there exists a lexical threshold below which comprehension is greatly hindered. Laufer and Ravenhorst-Kalovski, for instance, estimate that a reader needs to be familiar with 95% of the words of a text in order to be able to understand it (minimal threshold), and with 98% of the words for unassisted, comfortable reading (optimal threshold) [46]. This second threshold is also the one above which contextual guessing of unknown words' meanings can take place, which is instrumental in incidental vocabulary acquisition [35] as well as a fundamental strategy for reading comprehension [48]. Therein lies the beginner's paradox: to develop their vocabulary through reading, learners of a language need to already be familiar with enough words to cover 98% of the texts they read [11].

The notion of lexical difficulty is thus fundamental in the field of second language acquisition. Learners need to be presented with texts of the right difficulty level to ensure that they can enrich their vocabulary through extensive reading. A tool able to detect difficult words for a particular reader would have the potential to be a valuable help in many ways - for instance, in selecting texts of an adequate level, or in identifying words that need to be replaced with easier synonyms. The insights provided by such a tool could similarly be useful in the context of a reading aid providing definitions for difficult words, or for teachers and textbook designers who want to make sure that they are focusing on the right words for their target audience.

1.2.2.3 Practical implications

Lexical difficulty and lexical complexity are different phenomena; as such, predicting them require different approaches. Most of the literature focuses on Complex Word Identification, which consists in estimating the intrinsic complexity of a word based

on a series of measurable linguistic predictors. In recent years, two methods have primarily been used to assign a gold-standard complexity label to words in order to train predictive models. The first, often found in studies on complexity in French as a foreign language [27, 28, 30], relies on graded resources compiled based on word frequency in language textbooks, such as Manulex [49] or FLELex [26]. The second, typically used for English complexity prediction, is based on an aggregation of complexity or difficulty judgements [66, 100, 84]. Those judgements are generally collected through crowdsourcing platforms and come from a very diverse pool of annotators. This second type of Lexical Complexity Prediction task will be further discussed in Section 1.4.

Very few studies have been undertaken for lexical difficulty prediction, despite François and Fairon already pointing out that gap in the literature almost ten years ago [25]. Out of that small number of papers, the majority focus on predicting difficulty in order to simplify texts for specific groups, mainly language learners and readers with a disability. Fully-personalised predictions on an individual basis are even rarer, although they have shown some promising results [47, 89]. Section 1.5 will provide a more comprehensive overview of those studies.

Before exploring the factors that influence reading difficulty, it is helpful to understand what knowing a word entails, as well as how one can assess the degree to which a word is known. These elements will be discussed in the following section.

1.3 Vocabulary and Foreign Language Acquisition

Learning a language is first and foremost a matter of learning its vocabulary [45, 79]. As mentioned in section 1.2.2.1, difficult words are those that require more cognitive effort from the reader, because they are not easily accessible. Intuitively, we can assume that learning vocabulary means lowering the processing cost of a set of words by becoming increasingly familiar with them. This process of developing lexical knowledge will be discussed in what follows.

1.3.1 Vocabulary Knowledge

1.3.1.1 Definition

Lexical knowledge has been conceptualised in many ways [73], which is not surprising as it comprises a variety of dimensions. It has for instance been described as a sub-part of linguistic knowledge that consists in a mental word list and a set of composition and derivation rules [52]. A second and more widespread vision of lexical knowledge distinguishes two aspects of vocabulary: its depth (or quality) and its breadth (or size). Both dimensions have been shown to have an important impact on reading comprehension [72], and are thus equally interesting to measure. However, while the latter is rather straightforward to operationalise, as it often comes down to counting words whose meaning is known, the former is much trickier to pin down [81].

Depth of vocabulary knowledge is often represented as being made up of different components, each learned in an incremental manner through repeated exposure [81]. Following Paul Nation, we can divide these components into three main areas: form, meaning and use [62]. Form includes knowledge of the written and spoken representations of the word, of the different morphemes making up that word, as well as of the derivations that can be obtained from it. Meaning consists of familiarity with the different senses and referents expressed by the word form, the network of associations it enters into (synonyms, antonyms, hypernyms, related words in the L1...), and the connotations it can take. As for use, it involves having an intuition about the frequency of the word (*i.e.* the likelihood of encountering it in spoken or in written text), its syntactic behaviour, the registers it belongs to, and the collocations and fixed expressions into which it enters [62, 9, 75]. All those vocabulary knowledge components can be mastered at different speeds, as illustrated in Figure 1.1. Typically, aspects relating to use are acquired later, as they require to encounter the word in a variety of contexts in order to be learned incidentally [81].

The vocabulary knowledge (or mental lexicon) of a speaker can thus be vi-

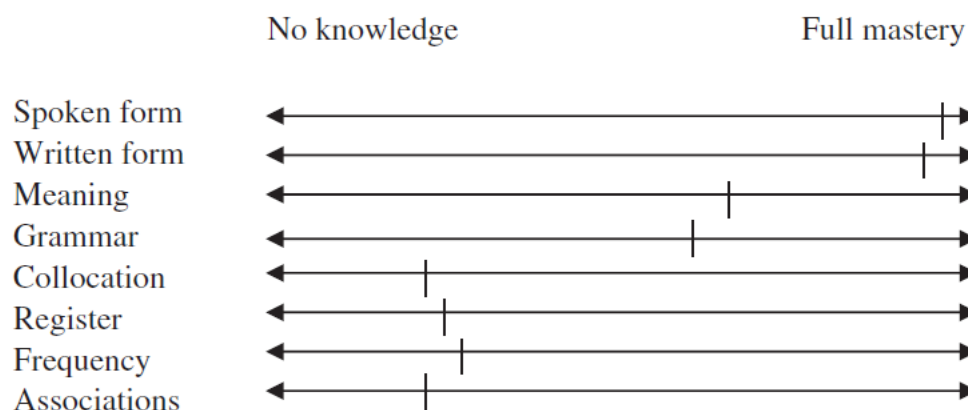


Figure 1.1: Developing Knowledge of a Word ([79, p. 38])

sualised as a complex network of words which are linked in many different ways. Breadth of vocabulary can be expressed as the number of entries in that lexicon, while depth is characterised by the quality of its organisation and the number of connections between the words it contains. If the notion of “depth” has undeniable theoretical validity, it is very difficult to operationalise in practice. Indeed, it encompasses too wide a variety of aspects to be captured by a single measure. Narrowing the scope is therefore necessary in a study investigating lexical knowledge, by specifying which aspect of vocabulary depth is targeted by a given research question [81].

The focus of this dissertation is on reading, which is a receptive task. The literature agrees that receptive (sometimes called “passive”) vocabulary knowledge is different from productive (or “active”) vocabulary knowledge. The former entails recognising a spoken or written form and mapping it to an entry in the mental lexicon (“meaning recall”), while the latter requires extracting that form from memory in order to realise a given meaning or function (“form recall”). As such, production can constitute a more difficult or advanced task for learners, which has been explained by the fact that it demands a deeper knowledge of the different word components. By contrast, being able to link a form with its meaning is often sufficient for receptive tasks, as the other types of information are present in the context. The form-meaning relationship can thus be said to be the most basic form of, as well as the

minimal requirement for, lexical knowledge [81, 9].

What, then, does it mean to know a word? To fit the purpose of this study, lexical knowledge will be understood here in a receptive sense and in a reading context: a speaker knows a word when they are able to recognise its written form and to map it to an entry in their mental lexicon which at a minimum contains meaning information. Conversely, a word will be unknown if it does not possess such an entry in the lexicon, whether it be because the reader has never encountered it before, or because they did not manage to abstract and retain its form-meaning link from previous exposure. In line with the way depth of vocabulary has been conceptualised, word knowledge will be envisioned as a continuum rather than as a binary distinction between known and unknown. Thus, lexical knowledge develops from total lack of mastery up to a stage where several (or even all) dimensions are mastered by the learner. Lexical difficulty will be measured as a factor of the quality of the association between written form and meaning, with very difficult words having no such mapping in a learner's mental lexicon. Whether the reader is aware of the collocations the word enters into, for instance, will not be considered in the present study.

1.3.1.2 Assessing Vocabulary Knowledge

The question of how to evaluate vocabulary size and quality is a core concern in the field of foreign language teaching. In a review of the literature, Schmitt [81] identifies several main ways that depth has been conceptualised in order to be measured. Several studies focused on only one component of vocabulary depth, such as knowledge of different meanings of a word, of forms that can be derived from it or of its frequent collocates, while others combined several components. Methods used to assess knowledge vary from multiple choice questionnaires asking participants to select the right synonym, antonym or paraphrase [6] to rewriting exercises based on pictures and aiming to elicit certain vocabulary items [50]. A common format used to evaluate vocabulary depth is the Word Associates Test (WAT), which assesses

two aspects of lexical knowledge (meaning and collocations) by asking participants to identify synonyms and collocates of a given word [73].

As for methods targeting the meaning component of word knowledge specifically, they often take the form of an annotation scale (see for instance [54], [85], [88]). Such studies present participants with items to annotate based on how familiar they are with their meaning. An emblematic example of that approach is Wesche and Paribakht’s Vocabulary Knowledge Scale (VKS) [96], which comprises five levels ranging from “I don’t remember having seen this word before” to “I can use this word in a sentence”.

This study will use a similar scale to assess lexical knowledge through difficulty judgements. Although such a method is first and foremost a measure of vocabulary depth, it can also give an indication of breadth through the proportion of words whose meaning is unknown. Both the qualitative and the quantitative aspects of vocabulary knowledge can thus be evaluated through this approach.

1.3.2 Foreign Language Acquisition

1.3.2.1 Main Factors

If the process of learning a foreign language was identical for everyone, personalised predictions of difficulty would not be necessary. Learners would go through the same developmental stages, would achieve the same results by using comparable methods, and would have the same needs as their peers at a similar level. In reality, however, learners of a language reach very different ultimate levels of achievement, and they do so at very different speeds. This is due to a myriad of individual and contextual differences that all impact the learning process and are responsible for the high level of variability that can be observed between learners [22].

Among the many factors influencing foreign language acquisition are the social, educational, political and pedagogical contexts in which the learner finds themselves.

These are linked to variables such as the international posture of the learner, their cultural and intergroup empathy and their language attitudes. How motivated a person is to learn a target language and how much communicative or foreign language anxiety they have are also predictors of the language acquisition process. Finally, the age at which a person started learning the target language, as well as whether they have been in contact with it through staying in a country where it is spoken, both have a recognised impact on language development [18].

Similarly, several learner-internal variables have an impact on the learning process. One that has been commonly cited is ability, which is thought to encompass intelligence, language aptitude and working memory capacity [22, 18]. A series of personality traits were also found to affect foreign language learning, namely the degree of extraversion, neuroticism, psychoticism, conscientiousness, openness to experience and emotional intelligence of the individual [18]. Other factors said to have an impact on foreign language acquisition are engagement and involvement with the language learning task [45, 80], as well as the learning style and strategies of the learner and their willingness to communicate [22].

Overall, many variables affect the learning process and make predicting a learner's lexical competence very difficult. Measuring all the factors mentioned above in a single study would be too time-consuming a task. However, taking at least some of them into account in order to group together learners with similar profiles should prove useful for the task of lexical difficulty prediction.

1.3.2.2 Impact of the L1

Another important factor in foreign language acquisition that was not mentioned in the previous section is a person's mother tongue. Unlike young children developing their L1, foreign language learners have already mastered a first language when they begin learning another. This leads to what linguists call "transfer" or "crosslinguistic influence": the learners' knowledge or use of their target language is impacted by

their pre-existing knowledge of their L1. Although this type of forward transfer (from an L1 to an L2) is most commonly talked about in the literature and will be the main focus of what follows, other forms of transfer also exist. For instance, a person's L1 might be influenced by their L2 (reverse transfer), or their L2 and L3 might affect each other (lateral transfer) [37].

Although the phenomenon of transfer is often mentioned in studies analysing learner errors, it can also have a positive outcome and facilitate language acquisition [37]. It has been shown that learners employ strategies when developing an L2, one of which consisting in making use of previous linguistic knowledge. Indeed, drawing links and similarities between structures or items of two languages makes the learning process easier [76]. For instance, when learning a new word in a foreign language, some believe it could be beneficial to initially map the unknown L2 form to a pre-existing L1 meaning in the mental lexicon. Doing so would free up more cognitive resources, which could then be spent on learning the new form [80]. Learners whose L1 uses a similar writing system to the one employed by the target language could also be at an advantage compared to others when it comes to learning orthographic word forms [81]. Because it impacts which aspects of a foreign language will be easier or more difficult for a given learner, “[crosslinguistic influence] can affect not only the rate and ultimate success of learners’ second language acquisition; it can also affect the route of their acquisition” [37, p. 11].

When crosslinguistic influence is found at the word level, it is referred to as lexical transfer. It can be, among other things, semantic, morphophonological or conceptual, depending on which aspect of L1 lexical knowledge is impacting the L2 [37]. The existence of that type of transfer begs the question of how lexicons of different languages are stored and organised in the brain: are they integrated or stored separately? Several studies support the first option, as it was shown for instance that seeing a form common to two languages (such as “four” for English and French) activates both sets of meanings and semantic networks for the reader [32].

It appears that words of different languages have the potential to be interconnected in many ways, either directly or indirectly by being both linked to another form or meaning [37]. However, it has been suggested that the way languages are organised in memory might vary from person to person, depending on characteristics such as the proficiency level in each language or the distance between an L2 and the L1 [94, 76]. Characteristics of the word itself could also play a role: cognates could for instance be stored together in a more integrated manner than other words [41, 94].

This last point indicates that cognates might constitute a special case in the mental lexicon. Although they are more accurately defined as words which share a common etymological origin, cognates can be seen under a psycholinguistic perspective as words from different languages that have a similar form (phonological, orthographic or both) and meaning [1]. This second perspective is more fitting here, as L2 readers of a text are typically not aware of whether two words share a common origin, but are likely to react differently to words whose meaning is transparent to them. Indeed, several psycholinguistic studies have found that cognates have a facilitating effect on foreign language vocabulary acquisition and reading comprehension [1, 76, 9, 92].

To sum up, crosslinguistic influence, and more particularly lexical transfer, is a key factor in the process of learning and reading in a foreign language. A person's mother tongue, and possibly the set of other languages that they have learned, is likely to have an impact on which words of a text they will find difficult. Taking a reader's linguistic profile into account should therefore provide a more accurate understanding of their specific needs.

1.4 Measuring Lexical Complexity from Annotation Data

As Section 1.2.2.1 established, lexical complexity is one of the main factors that influence lexical difficulty. Therefore, looking at findings from Complex Word Identification (CWI) studies that used annotation data to learn predictive models should prove enlightening for two reasons. First, because it will give an overview of the ways complexity annotations have been collected and aggregated into a gold standard, and should allow conclusions to be drawn about good practices to follow. Second, because it should provide valuable insights into the best predictors of lexical complexity.

The way lexical complexity has been conceptualised has evolved in the past few years. Initially measured in a binary manner, the assessment of complexity now seems to be moving towards a continuous scale so as to better capture varying degrees of difficulty. This change is made evident by the evolution of three shared tasks: two on Complex Word Identification [66, 100], and one on Lexical Complexity Prediction (LCP) [84].

1.4.1 Before 2016

While the first shared task on CWI was organised in 2016, a Semantic Evaluation (SemEval) workshop held in 2012 already included a task focusing on Lexical Simplification for English [87]. Participants were given a target word in context as well as a list of substitutes to rank from simplest to most complex. The dataset used for the task was compiled using the corpus of the 2007 shared task on Lexical Substitution, hosted at SemEval-2007 [58], which consisted in 2010 contexts containing a total of 201 lexical words to replace with a synonym. Each target word was thus found in 10 different contexts, and was associated to a list of substitutes provided by native speakers of English. For the Lexical Simplification task, a total of 9 annotators (4 for the training set, 5 for the test set) were asked to rank the substitutes from

easiest to most complex. The annotators were fluent non-native English speakers coming from different backgrounds. The gold-standard rankings were determined by aggregating all annotations for a given target word and sorting them based on their average rank.

Several interesting findings of this shared task can be reported. First, inter-rater agreement scores computed using an adaptation of Cohen’s kappa index [12] were rather low for both the training and test set (0.386 and 0.398 respectively). Interestingly, the agreement was higher between annotators of a similar L1 and/or educational background. The authors report a maximal pairwise agreement of 0.52, obtained by two annotators who had the same mother tongue and education, as well as similar English proficiency levels. It is also worth noting that annotators whose L1 belonged to the same language family also achieved higher inter-rater agreement scores [87]. These results confirm what we outlined in Section 1.3.2, namely that a person’s profile has an impact on which words they consider difficult. A second interesting consideration is the fact that a baseline which ranked words according to their frequency was only outperformed by one system, and matched by a second. This goes to show that frequency is an accurate predictor of lexical complexity. The poor performance of the submitted models compared to the baseline could be due to the small size of the dataset, as well as to the limited number of annotators and the low rate of agreement between them. The top system used a context-sensitive frequency model as well as a set of psycholinguistic features such as concreteness, imageability, familiarity and age of acquisition [38], which seems to indicate that these variables are good predictors of complexity as well.

1.4.2 CWI-2016

The 2016 International Workshop on Semantic Evaluation hosted the first edition of a shared task on Complex Word Identification, with the aim of “provid[ing] reliable resources and new insights for CWI, as well as [establishing] the state of the art performance in CWI for English texts, and bring[ing] more visibility to the area of

Text Simplification” [66, p. 560].

The task used a set of 9200 sentences annotated by 400 non-native speakers of English. The sentences were extracted from three corpora, all based on Wikipedia or Simple Wikipedia. Annotators were asked to categorise lexical words as either complex (1) or non-complex (0). Each eligible word was annotated by 20 people for the training set, while the test set only contained one annotation per word. According to the organisers of the task, this was designed to simulate a real-life situation, where the needs of one person would have to be predicted based on those of the target population as a whole [66]. The training set was presented in two ways: the joint version only had one tag per word, attributing a value of 1 if at least one of the annotators marked the word as complex, while the decomposed version provided all 20 annotations for each word. Given a word in context, participants were asked to predict whether it would be understood by a non-native speaker of English, *i.e.* whether the single annotator had labelled it as complex (1) or simple (0) [66].

21 teams took part in the shared task, entering a total of 42 systems to the competition. The following sections will present the approaches and predictors used by the three best-ranking teams in terms of G-score and F-score – as three teams shared the third place with a G-score of 0.767, the one with the highest F-score was selected for this review. The results of the task will then be discussed, as well as some of its limitations.

1.4.2.1 Best ranking teams in terms of G-score

Although the F-score, based on the harmonic mean between precision and recall, is traditionally used in NLP to evaluate a system’s performance, the organisers of the shared task felt that it was not the best fit for CWI. Indeed, a lexical simplification system is successful when it simplifies as many complex words as possible while minimising erroneous replacements of simple terms. This means that a good CWI model minimises both false positives and false negatives, or in other words that it

maximises accuracy rather than precision. Thus, the G-score, calculated from the harmonic mean between accuracy and recall, was preferred for the task [66].

The best-ranking team in terms of G-score was SV000gg, who submitted two models based on ensemble methods (a machine learning system which predicts a value based on the predictions of several sub-models, or “voters”). The first model used Hard Voting, which chooses the class predicted by the majority of voters; the second employed Performance-Oriented Soft Voting, which attributes a weight to the output of each voter according to its performance. The latter system ranked first and the former second, with G-scores of 0.774 and 0.773 respectively. The models used a variety of predictors, ranging from more traditional features, such as length and presence in a vocabulary list, to information extracted from WordNet [24], like number of senses or hypernyms [67].

The second best team, TALN, also used ensemble methods in the form of a Random Forest classifier. They took into account how many annotators had marked a word as complex in their weighted system, which obtained better results than the simple (unweighted) version, with G-scores of 0.772 and 0.750 respectively. Words were modelled as feature vectors including a variety of predictors: traditional features like length and frequency, dependency tree features and, similarly to the previous team, features extracted from WordNet. While several predictors obtained using WordNet contributed well to the model, they did not outperform the two frequency features based on Wikipedia and the British National Corpus [77].

UWB ranked third in the shared task while using a single feature in their maximum entropy classifier: document frequency. They argue that this feature is a better indicator of the true frequency of a word than term frequency measures, as it limits the impact of a word being very frequent in one document but barely ever appearing in the rest of the corpus. The team had originally selected other features, including ones taking context into account through the use of n-grams; however, while they did contribute to complexity, they did not outperform the simple unigram

document frequency predictor [42].

1.4.2.2 Best ranking teams in terms of F-score

While several teams performed quite well in terms of G-score, F-score results were low for everyone who participated in the task, with the best team obtaining a score of only 0.353 [66]. A possible explanation of this phenomenon would be that most systems were built with the G-score in mind. However, this is not the case for team PLUJAGH, who optimised one of their models for F-score and consequently came first in that ranking. The team used the document frequency of a word in Simple English Wikipedia as their only predictor, as it performed better than the other features they had investigated, either alone or combined [99].

LTG came second using ensemble methods. Their approach was original in two ways. First, they experimented with different minimal thresholds of inter-annotator agreement: while the joint dataset only required one annotator to find a word complex for it to be classified as such, the team used Decision Tree classifiers with thresholds ranging from one to four annotators. Second, they included contextual language model features in the form of different n-gram probability measures. This choice was motivated by the impact of context on how a word is understood. As many words are polysemous, not taking context into account could lead to erroneous classifications [55].

The third team, MAZA, devised their system based on the assumption that complex words tend to be less frequent and longer than simple words. They included features such as word probability in a corpus, word length and n-gram probabilities in an ensemble method using Decision Trees. They submitted two systems to the competition: one using only frequency and length features, and another adding contextual features as well. The latter achieved a higher F-score, but the former did better in terms of G-score [56].

1.4.2.3 Discussion

Some methods and features were used by several of the best-performing teams, and can therefore be assumed to be a good fit for Complex Word Identification systems. Simple predictors such as frequency, length and presence in a vocabulary list performed very well, with two teams ranking third for G-score and first for F-score respectively while using a single predictor: document frequency. More sophisticated features, such as those extracted from WordNet, also proved useful. Predictors taking context into account showed somewhat promising results, but often failed to outperform simpler features like frequency and length. Two teams included the number of annotators who had marked a word as complex in their systems and obtained good results, which shows that a word found difficult by a greater number of people is more likely to constitute a problem for someone taken out of the same population.

Moreover, ensemble methods were quite popular among participants, and were used by half of the ten best performing teams in terms of G-score [67]. This would suggest that they are particularly apt at classifying words as simple or complex, potentially because they allow for a variety of predictors and models to be used. However, all teams obtained rather low F-scores, which Shardlow *et al.* [85] interpreted as a result of the complexity of the task. Indeed, the strong imbalance between training set and test set, the latter being more than forty times larger than the former, meant that participants had to be able to classify a wide number of words by training on a much smaller dataset. Furthermore, the test set relied on a single annotation per instance while each word in the training data was annotated by 20 people, which added to the difficulty of the classification task.

Shardlow *et al.* [85] raised several further limitations of the 2016 shared task. First, since complexity was represented in a binary manner, annotators had to decide on an arbitrary threshold to separate simple and complex words. Therefore, all annotations collected for the dataset might not represent the same conception of

complexity. Moreover, only single words were analysed in the task, when multiword expressions (MWEs) are likely to be a source of problems for non-native speakers. A third limitation of the shared task was that demographic data wasn't provided (though it had been collected), when factors such as native language and background are likely to influence which words will be considered simple or not. Indeed, the organisers of the task found a correlation between the age or English proficiency of annotators and the number of words they marked as complex. Similarly, the mean inter-rater agreement for the data was very low (Krippendorff's α of 0.244), which indicates that annotators were heterogeneous in their profiles [66].

1.4.3 CWI-2018

Two years after the first shared task on Complex Word Identification, a second one was hosted at the Workshop on Innovative Use of NLP for Building Educational Applications [100]. This time, the data were not exclusively in English: training and test sets were also provided for German and Spanish, as well as a French test set for multilingual annotation. These languages represented the four available tracks for the CWI task: monolingual English, German or Spanish and multilingual French, which had participants classify French data by training models on the other three languages. The English dataset comprised three domains which were evaluated separately: News, WikiNews and Wikipedia articles [85]. In contrast to the 2016 edition, the 2018 CWI shared task provided participants with more training data than test data, and instances in the test set were annotated by several people [100].

The annotation process was conducted differently as well. While the 2016 task asked participants to annotate preselected words in a sentence, this time annotators were presented with paragraphs of five to ten lines and were asked to select words or expressions that could hinder comprehension. A maximum of ten words or expressions could be selected. The pool of annotators was made up of native and non-native speakers of the chosen languages, but all were asked to choose complex words or MWEs with a target population of learners or people with reading impair-

ments in mind. This time again, complexity was represented as a binary value which would be set to 1 if at least one person found the word complex. However, a probabilistic annotation was also provided, in the form of the proportion of annotators who marked a given word as complex. For the binary sub-task, teams were ranked according to their F-score, while the probabilistic sub-task was evaluated using the mean absolute error of a system’s predictions [100].

Twelve teams took part in the shared task, submitting systems for different combinations of tracks and sub-tasks; selecting the best-performing teams was therefore not as simple as it was for the 2016 edition. Four teams were nonetheless chosen for discussion. The first one is CAMB, who ranked first in five of the six monolingual English sub-tasks. The second, TMU, ranked first in both the binary and probabilistic sub-tasks for German and Spanish, as well as in the probabilistic English task for the News corpus. Third comes CoastalCPH, who performed best for both multilingual French sub-tasks. The decision was made to also include team ITEC, who came second for both Spanish sub-tasks as well as for probabilistic English for the News domain.

1.4.3.1 Description of the best-performing systems

CAMB obtained the best results for the monolingual English track, with F-scores exceeding 0.8 in all three domains of the binary sub-task. They selected features that had proven useful during the 2016 CWI task, as well as some additional ones, such as psycholinguistic information, CEFR levels and google n-gram word frequencies [31]. The predictors were included in an ensemble method, and it was shown that they did not contribute equally for all corpora. For instance, psycholinguistic predictors proved useful for both WikiNews and News, but not for the Wikipedia corpus. The team found that the best-performing way of dealing with multiword expressions was to mark them all as complex, although they experimented with other techniques as well. The systems entered to the competition by CAMB were context-independent, which they said was one of the main sources of errors in their results [31]. This goes

to show that a CWI system benefits from using contextual features, as many words take different meanings depending on their context.

The team that performed best for the German and Spanish monolingual tracks was TMU, who used Random Forest classifiers for the binary sub-task and Random Forest regressors for the probabilistic version. They obtained good results using very simple features: length (expressed in characters), number of words for MWEs, frequency in native and learner corpora and probability in the same corpora. Their intuition was that learners use simpler words than native speakers, and thus that using a learner corpus should prove useful. Indeed, learner corpus features performed better than those extracted from native corpora [40].

For the multilingual French tracks, CoastalCPH obtained the best results with their cross-lingual model. Their system used multitask learning based on Random Forest models and used a variety of “language agnostic features”. These include log-probability of the word in Wikipedia, character perplexity, WordNet features, inflectional complexity, part of speech and target-sentence similarity using word embeddings. Their performance on the French test set without any French training data shows that their system was able to generalise well between different languages [5].

ITEC came second in several combinations of sub-tasks and tracks. Even though deep learning architectures did not perform very well in the 2016 edition of the task, the team used a neural network with a variety of predictors: frequency, word length, psychological measures, and word and character embeddings. These last two features allowed for distributional and morphological information to be captured, as well as topical information through the concatenation of embeddings at the sentence level. The team found that word embeddings performed best, followed by character embeddings. This is due to those features indirectly capturing frequency and length information respectively. Topical information, however, was found not to contribute much to the model [16].

1.4.3.2 Discussion

The organisers of the 2018 shared task on CWI pointed out that the results were better than they had been in 2016, and that cross-lingual models showed promising performances [100]. The four short presentations made in the previous section confirmed the efficiency of simple features such as frequency and length. Less traditional predictors, namely psycholinguistic measures and features extracted from a learner corpus, also yielded some interesting results. Finally, the importance of taking context into account was shown once more, as failing to do so was one of the main sources of errors for the first team. Embeddings proved an efficient way of capturing contextual features, with the advantage of indirectly providing information on frequency (word embeddings) and length (character embeddings).

The task also presented some limitations, mainly caused by the guidelines given to annotators. First, because participants were only allowed to select up to ten words per paragraph, some might not have been able to annotate all the words that they found complex. Annotators also had a tendency of only selecting the first occurrence of a word if it appeared more than once in a paragraph, which the previous restriction might have encouraged. Second, since words were not preselected, some annotators considered certain strings of words as separate when others saw them as forming multiword expressions [31]. As a result, annotations were not as consistent as one would hope. The system used in the 2016 edition of the task, with words to annotate having been preselected, might therefore be preferable. Another option would be to predefine the bounds of each lexical unit, while still allowing annotators to select any one of them. Finally, the fact that participants had to select words that could be difficult for other people, and not for themselves, might also have impacted the data. This guideline assumes that speakers of a language, native or not, are able to predict which words could pose problems to learners or people with a reading impairment, and makes it likely that some annotators selected words which they did not themselves find difficult. How reliable the annotations were is thus not

certain.

Furthermore, the annotation system had the same shortcoming as pointed out for the 2016 edition: as words were annotated in a binary manner, each annotator had to decide on an arbitrary threshold to define at what point a word becomes complex. While a probabilistic sub-task was proposed, the continuous values provided in the data were still computed on the basis of binary judgments, as they represented the proportion of annotators having found the word or expression complex. Therefore, Shardlow *et al.* [85] rightly observed that a word labelled with a value of 0.5 was not to be seen as “half complex”, but rather had been considered complex by fifty percent of people. Though it did constitute a step in the right direction, this aggregation of binary values to create a continuous scale was thus not yet optimal.

Once again, demographic data were collected but not provided to participants, although top-ranking team CAMB believe that they would have been useful in the task. An inter-rater analysis carried out on the English section of the dataset revealed that non-native annotators had a lower agreement rate than native speakers [101]. This could be due to the non-native group being made up of a variety of profiles, whether it be in terms of native language or English proficiency [31]. Once again, the lower agreement rate between speakers of different L1s and proficiency levels indicates that they probably have different simplification needs. Therefore, investigating the way factors of lexical difficulty vary between more homogeneous groups should prove beneficial to the field of Lexical Simplification.

In a recent paper, Shardlow *et al.* [85] argued that the datasets used for both shared tasks lacked certain characteristics that would make them a better fit for Complex Word Identification / Lexical Complexity Prediction. They proposed a series of recommendations for future datasets to follow:

1. The annotations should be continuous rather than binary, for example by asking annotators to rate complexity using a Likert scale;

2. Words should be presented in context as it has an influence on their meaning;
3. As words are polysemous, several occurrences of a same token should be included;
4. Each token should receive several annotations;
5. The annotators' backgrounds and proficiency levels should be as diverse as possible in order to be representative of the wider population;
6. Texts should be chosen from a variety of genres;
7. Both single words and multiword expressions should be included in the data.

These guidelines were taken into account when designing the dataset which was to be used in the third shared task focusing on lexical complexity: LCP 2021.

1.4.4 LCP-2021

The fifteenth International Workshop on Semantic Evaluation (SemEval-2021) [68] hosted its presentations in August 2021. The first task of the workshop focused on what the organisers called Lexical Complexity Prediction, or “CWI in a continuous-label setting” [85, p. 2], thus moving away from the limitations of binary annotation. The task relied on a dataset compiled with the above recommendations in mind. It had the following characteristics:

1. The corpus was annotated using a 5-point Likert scale, ranging from “very easy” to “very complex”;
2. Words were presented in context, and annotators were asked to take the contextual meaning of the instance into account when evaluating complexity;
3. Several tokens occurred multiple times;
4. On average, 25.75 annotations were gathered per word or expression;
5. Annotators were recruited in a crowd-sourcing setting, and demographic in-

formation was not recorded;

6. Data were extracted from three different genres: the Bible, Europarl and Biomedical texts;
7. Single words and MWEs were both taken into account.

Most of the guidelines laid out by Shardlow *et al.* [85] were thus followed in their dataset, with the exception of the fifth recommendation. Indeed, demographic data were not gathered during the annotation process and it is therefore impossible to assess how representative the sample was. Moreover, if different groups have different simplification needs, this would not be visible in the dataset.

The task comprised two sub-tasks: one asking participating teams to predict the complexity of single words only, the second taking both single words and multiword expressions (MWEs) into account. In the dataset used for the task, the definition of MWEs was reduced to bigrams made up of either two nouns, or one adjective followed by a noun. While the creators of the dataset were aware that this was a very limited approach and that many expressions were left out, they argued that the types of MWEs they captured are the most frequent in English [85]. Words and expressions were given a single complexity value based on the mean value of all annotations they had received, transposed in a 0-1 range. Individual annotations were not provided in the dataset, and participants were thus tasked with predicting the value obtained from the aggregation of several complexity judgements [84]. The full dataset contained 10,800 tokens, which on average each appeared in 2 different contexts. Single words had a mean complexity of 0.302, while MWEs reached an average of 0.419, confirming that they tend to be more complex (as mentioned in Section 1.2.1.2). The way the data was split was thought out carefully by the organisers of the task, to ensure that the training instances reflected the data of the test set. Participating teams were provided with a trial set (5% of the data), a training set (85%) and a test set (10%), each containing comparable data [84].

A total of 55 teams took part in the task. They were ranked using Pearson’s correlation, but other metrics were also computed, such as Spearman’s rank and Mean Absolute Error, in order to provide secondary rankings. Unlike in the previous shared tasks, the submitted systems achieved much better results than the frequency baseline. Moreover, the teams’ scores were very close together, which shows that many of them managed to model the data well. Both deep learning approaches and feature-based approaches achieved top results, indicating that neither is intrinsically more efficient than the other at predicting lexical complexity. Systems that took context into account when learning and predicting the data, however, did tend to outperform models which did not [84]. This finding confirms what had already been suggested in section 1.4.3.2.

1.4.4.1 Description of the best-performing systems

For this brief overview, it was decided that the top 3 systems of both sub-tasks would be singled out for discussion. However, as one team ranked in the top three for both sub-tasks, and as no paper was submitted for the model ranking third in the second sub-task, only four systems will be described here. JUST BLUE [3], DeepBlueAI [69] and rg.pa [74] all used pretrained language models such as BERT [17] and RoBERTa [51] and secured the top two places in both sub-tasks. As they employed a deep learning approach, those models did not rely on a series of handcrafted predictors like the systems presented in the previous sections. Not much can thus be said about the features which were helpful in predicting complexity. It is worth noting, however, that all three systems used the context in which a word or expression appeared to model and predict complexity.

The system that came third in the first sub-task was the one submitted by Alejandro Mosquera [61], who used a feature-based approach. A total of 51 predictors were used, with a combination of traditional features, such as frequency information, various measures of word length or age of acquisition, and other features aiming to capture the context in which the word appears, such as the readability score of the

sentence. Those performed very well, which the author argues could be linked to the fact that complexity measures at the sentence level can give an indication of its register and genre. Although datasets of the two previous CWI tasks were initially used along the LCP-2021 dataset in order to train the model on more data, it was found that their inclusion actually decreased the performance of the model. This is probably due to the very different annotation guidelines and methodologies used across the three tasks.

1.4.4.2 Discussion

This latest shared task saw a shift from feature engineering approaches to deep learning models, and showed that both options constitute viable ways of evaluating lexical complexity. As a result, there is less to be said about the features that were used as predictors than for the previous CWI tasks. One thing that was made evident from the systems’ performance, however, is that taking context into account when measuring complexity makes for more accurate predictions.

The good performance of many of the submitted systems is testament to the quality of the dataset, and indicates that the recommendations described above are worth following. Nonetheless, the way multiword expressions were defined and selected in the corpus has its shortcomings. First, as mentioned above, restricting MWEs to bigrams of only two types means that many other types of expressions were overlooked when compiling the dataset. Second, the organisers of the task did not distinguish between compositional and non-compositional expressions, when that feature is very likely to influence whether a word pair is perceived as difficult or not. Indeed, the authors themselves note that the complexity of compositional bigrams is likely to be a function of the individual complexity of each of its constituent, while that of non-compositional expressions will probably not be dependent on how complex the individual words are [85]. Whether a reader understands the bigram “ancient parchment”, for instance, will depend solely on their knowledge of both of those words individually. By contrast, understanding that “parchment paper” is

used for baking instead of writing demands that the reader be already familiar with the term.

This distinction should have an important impact on the most efficient way of predicting the complexity of a multiword expression. For compositional bigrams (which, under our working definition provided in Section 1.2.1.2, do not constitute MWEs but rather collocations or free associations), taking the average complexity of the two items or labelling the whole expression with the value of the most complex word should provide satisfactory results. Such a method would be unlikely to work for non-compositional bigrams, however, and treating them as single words (extracting features such as the expression’s frequency of occurrence or level of concreteness) should be preferable. As the dataset did not make the distinction between those two types of bigrams, participating teams who used feature-based approaches had to decide on a way of labelling those expressions regardless of whether they were compositional. According to the organisers of the task, some teams chose to average either the complexity scores or the features of both words, while others trained models with twice as big a feature space for multiword expressions [84]. It would be interesting to know which strategy worked best, and whether separating the two types of bigrams would have impacted the quality of the predictions.

The use of continuous values rather than binary classes to label lexical complexity yielded encouraging results. Although a Likert scale annotation still involves some subjectivity, as people might not agree on where to draw the line between “complex” and “very complex”, the organisers of the task argue that it nevertheless provides a clearer interpretation than binary judgements, in that it gives an indication about the position of a word on a complexity continuum [84]. An analysis of the data showed that annotations for a given instance were in general normally distributed, which justifies using the mean of all annotations as a complexity label. However, some sets of annotations did not follow the normal distribution, and others that did were sometimes spread over all 5 possible values of the scale [85]. This

shows that individual differences are still visible in the data, and highlights once more the value of measuring difficulty. It is therefore a shame that demographic data were not collected when compiling the dataset, as they could have provided some interesting insights into which factors have the strongest impact on lexical difficulty.

1.4.5 General Discussion: Insights from the Shared Tasks

This review of work undertaken in the past decade on Complex Word Identification or Lexical Complexity Prediction from annotation data has revealed several changes in the way the task has been carried out. From binary annotations, the field is now moving towards continuous labelling that gives a more fine-grained representation of how complex a word is for a given population. While CWI-2016 provided a test set with only one annotation per instance, words in the CompLex dataset have been annotated by an average of over 25 people, which means that labels are more likely to be representative of the inherent complexity of a word. In addition to single words, multiword expressions have been included in both the 2018 and the 2021 tasks. This is a trend worth perpetuating and refining, as readers are more likely to struggle with fixed expressions whose meaning is not compositional than with free associations. While the first two shared tasks made substantial use of Wikipedia corpora, the latest dataset collected texts from three markedly different genres, allowing for further generalisation of complexity predictors across domains.

The elements outlined above constitute improvements of the task design, and have positively impacted submitted systems' performance. Some methodological choices, however, have been inconsistent throughout the tasks and do not allow any trend to be detected. The pool of annotators, for instance, was sometimes made up of non-natives only (2012, 2016), of both natives and non-natives (2018), or of people whose profile was unknown (2021). For a future shared task, it might be worth reflecting on which profiles should be considered. If the aim of the task is to identify the intrinsic complexity of a word, thought to be fully independent

from the language skills of the reader, then having as diverse and large a pool of annotators as possible should be best. Collecting demographic data to ensure variety and representativeness would thus be necessary. By contrast, one might argue that the variability between annotators in the previous tasks was too high, and that restricting the pool of annotators to a more specific profile would be preferable. The conclusions drawn from such data would naturally only be applicable to the population queried.

A second point of variability across the tasks concerned the guidelines given to annotators. In 2016 and 2021, they were asked to provide a label to selected items depending on how well they themselves understood them. The complexity values attributed to annotated words were thus based on an aggregation of difficulty judgements. In 2018, the guidelines had annotators classify items as complex or not based on whether they thought a given word could be difficult to understand for a learner or a person with a reading impairment. This assumes that both native and non-native speakers of a language are able to predict the needs of other readers, which is unlikely given that even language teachers sometimes struggle to identify sources of difficulty for their students [59]. It is also interesting to note that all three shared tasks used judgements on the difficulty of a word, either for oneself or someone else, as a way to measure complexity.

Which words were annotated was a third area of variation between the tasks. CWI-2016 and LCP-2021 both preselected the items to be considered for annotation, while CWI-2018 let annotators select any word or expression they wished. In the first case, the items to annotate were presented in a sentence, while the second case provided short paragraphs. Although the 2018 configuration had its shortcomings, which were discussed above, the preselection of words might not be ideal either. Indeed, by only ever considering content words, some other types of complex items might be overlooked. Prepositions such as “owing to” or “apart from”, for instance, might not be understood by all learners of English. Moreover, preselecting words

might make the task less natural, and thus less accurate, than asking people to read a text and highlight words they do not understand. Indeed, singling out a word for annotation could make it artificially more salient for readers who would have otherwise glided over it. It could therefore be worth comparing annotations of the same text obtained using both methodologies, in order to assess whether they are in fact equivalent.

As for the way the datasets were presented to participants, only CWI-2016 provided individual annotations along the gold standard complexity value. Whether to include them in a new task would depend on the organisers' objectives, but it might be interesting to explore whether having access to them might help participating teams obtain better results.

Several recurrent features were found to be excellent predictors of complexity. The strongest was undeniably frequency: teams in the first two editions of the task struggled to beat frequency baselines, and two systems in CWI-2016 achieved top results while using document frequency as their only predictor. Length, measured for instance in characters, syllables or morphemes, was a close second. Other good predictors were psycholinguistic measures such as concreteness, familiarity or age of acquisition of a word, features extracted from WordNet, the presence or absence of an item in a list of simple words, and word or character embeddings. A recurrent finding of the tasks was that taking the context in which an item appeared into account improved the quality of a system's predictions.

Despite the improvements described above, several gaps in the field remain. First, only CWI-2018 used data in languages other than English, and the English dataset was still larger than the other three. While the task seemed to show that a language-agnostic approach to CWI was possible, with complexity labels in the French test set being predicted rather accurately by models trained on English, German and Spanish, this might not be the case for all combinations of languages. Indeed, French is very close genetically to Spanish, as well as historically to English,

which means that there is likely to be a greater lexical similarity between those pairs of languages than with an Asian or Slavic language, for instance. Therefore, investigating factors of lexical complexity for a wider variety of languages could prove useful. Second, the models presented above exclusively used intrinsic features of lexical complexity, *i.e.* features of the word itself or of its context. However, extrinsic factors such as a speaker’s native language or level of fluency in the target language also influence which words they will find difficult, as was made evident by the inter-rater analyses reported for LS-2012 and CWI-2016. Shifting the focus from complexity to difficulty by taking demographic data into account should therefore allow for even more accurate predictions.

1.5 Measuring Lexical Difficulty

Lexical difficulty, as mentioned before, is a phenomenon that takes place on an individual level. What will be cognitively costly to process for one reader might be easy for another; therefore, a one-size-fits-all approach to difficulty prediction would be doomed to fail [59]. Yet, only a handful of studies have looked into personalised measures of difficulty [90]. An overview of the research on lexical difficulty prediction will be provided in what follows.

1.5.1 Overview of Studies

1.5.1.1 Personalised Familiarity Prediction

In psycholinguistic studies, the notion of word familiarity describes the “relative ease of perception attributed to every word” [91, p. 96]. The concept is therefore close to that of lexical difficulty, although it has often been defined as a unique, person-independent measure based on the aggregation of subjective judgements (much like lexical complexity). In 2005, however, one study investigated the possibility of using demographic information to better predict the familiarity of words related to the medical field. They used a multiple-choice questionnaire testing whether 21

participants understood 34 health-related terms in order to obtain familiarity data. The annotators were all adult speakers of English who did not work in the medical field, and they were asked to provide a series of demographic information such as their age, gender, L1 and education level [102].

Individual answers were transformed into binary labels: 0 for words with which a given annotator was not familiar, and 1 otherwise. An average familiarity score was attributed to each word based on the mean of the individual answers it received. Two baselines were created using common features for readability measures: word length (expressed in syllables) and presence in a list of frequent words. Those only gave a coarse estimation of which health-related terms could be difficult for a public of non-specialists. An SVM model was also created, with the aim of predicting the average familiarity scores. It used features extracted from health-related corpora, such as the frequency of the term in each of them, as well as the two features used for the baselines. The model reportedly achieved moderate success [102].

More interestingly, a linear regression model was also created, this time taking demographic data into account in addition to word features. The considered characteristics were gender, age (20 to 35, 36 to 50 or 51+), education level, native language (coded as “English” or “Non-English”), “race” (“White” or “Non-White”), ethnicity (“Hispanic” or “Non-Hispanic”), and employment (“Professional” or “Non-Professional”). Gender, age and ethnicity were found not to contribute significantly to the model and were thus discarded. Despite the obvious limitations of those categories, the model achieved satisfactory results, as it successfully classified 80.4% of the items for all participants (574 out of 714) [102]. Using demographic features as predictors in a model thus seems to be a viable way of achieving accurate individual predictions of lexical difficulty.

1.5.1.2 Personalised Vocabulary Prediction

In 2012, Ehara *et al.* [20] investigated the possibility of learner-specific vocabulary prediction. They argued that the issue with previous models predicting the vocabulary knowledge of learners was that each word was attributed a single difficulty measure, thereby assuming that all individuals had the same perception of word difficulty. Since many factors impact vocabulary knowledge, such an approach can only be limited in its performance.

To build their dataset, the authors asked sixteen English L2 learners to annotate 12,000 words based on how familiar they were with them. Participants were for the most part graduate students from the University of Tokyo whose L1 was Japanese. The annotators had to rate the difficulty of each word by using a 5-point scale with the following levels: 1) Never saw the word before, 2) Probably saw the word before, 3) Absolutely saw the word before but don't know its meaning / tried to learn the word before but forgot its meaning, 4) Probably know the word's meaning / able to guess the word's meaning, 5) Absolutely know the word's meaning [21].

The difficulty levels were then converted to a binary label: all levels were mapped to “difficult”, except for the fifth one. This decision begs the question of why a 5-point scale was used in the first place, as it does not appear that the levels given by annotators were used in the study. Moreover, one might question the validity of considering words labelled with 4 as equally difficult as those which a learner had never seen before (label of 1). Nevertheless, a model was trained using individual annotations and predictors based on frequency information in several corpora (negative log of the unigram probability of a word). According to the authors, the model obtained competitive results compared to similar systems which did not account for learner-specific lexical difficulty [20].

An analysis of the words for which the highest variation in annotations was observed was also carried out, and main causes for such high variability were suggested.

A common reason was when the English word had been adopted as a loanword in Japanese, but with a difference in spelling. Annotations would thus vary depending on whether each learner was able to recognise the link between the two words. One could also wonder whether some learners might not have indeed recognised this similarity and thus selected the fourth level on the scale, which would then be interpreted as them finding the word difficult. Another source of variability proposed by the authors was errors, in that learners might have confused certain words with other English terms that they were more familiar with. Lastly, topic specificity was found to play a role in the perceived difficulty of items, indicating that some learners with a low proficiency level might have been familiar with certain words belonging to a domain that they are interested in [20].

Asking participants to annotate 12,000 words is extremely costly, and such an approach is not often an option. Instead, a more feasible method is to use annotations made on a small subset of words as a way of predicting someone's vocabulary knowledge. This is referred to as word sampling, and assumes that the words selected for annotation are representative of a set of similar words. Until recently, the most common way of sampling words was to group items based on their frequency, and then randomly draw a given number of words from each frequency band [19]. If a learner knows words belonging to a certain frequency group, it is assumed that they also know the other terms belonging to that group, and thus that all words of a similar frequency have the same relative difficulty. Building on the observation mentioned above that the topic to which a word belongs can also impact difficulty, Ehara *et al.* [19] proposed a new approach to word sampling, using graph-based active learning. Words of a language were represented as nodes in a graph, which were linked to other nodes in a way that indicated domain specificity. The edges between word pairs were given a specific weight representing the similarity in difficulty between the words, estimated based on frequency information from several corpora. Sampling was done directly from the graph, in a way that also took topic information into account. Learners were then asked to annotate more representative

sets of words than with previous approaches. A quantitative analysis of the results obtained using this method showed that the quality of predictions was improved compared to a more traditional, frequency-based sampling [19].

1.5.1.3 Personalised Lexical Simplification

In 2018, Lee and Yeung addressed the fact that most lexical simplification systems do not take individual differences into account, although the most optimal substitution for a word - or whether the word needs to be simplified in the first place - could differ from reader to reader. They used personalised CWI models at two stages of the pipeline: detection of words which might be too difficult for a given reader, and ranking of possible substitutes. They also explored the possibility of creating several CWI models based on graded vocabulary lists and representing different proficiency levels, in order to match the most suitable model to each reader after having them annotate a small sample of words [47].

To build their models, the authors made use of two datasets. For lexical difficulty annotations, they used the corpus presented in section 1.5.1.2. Like Ehara *et al.* [20], they converted the difficulty levels into binary values: words with a level of 1 to 4 were labelled as complex, and those with a level of 5 as non-complex. For lexical substitutions, they adapted the BenchLS dataset [65], which provides a gold-standard ranked list of substitutes for over 900 target words. A personalised version of BenchLS was created for each annotator in the first dataset, by setting the gold answer to “null” for all words that the participant found easy, and removing difficult words from the gold list of substitutes for the rest. Results obtained with the personalised models were compared to a baseline using the original version of BenchLS, which showed that reader-dependent CWI consistently outperformed user-independent lexical simplification. The proficiency-level models were also found to improve the precision and accuracy of predictions [47].

In a similar approach, Tack *et al.* [90] explored different models predicting

the lexical knowledge of learners of French as a foreign language. They collected binary difficulty annotations through a custom-made online reading platform which asked participants to read texts one sentence at a time and to click on any word whose meaning they did not know. Lexical items to annotate had been tokenised and lemmatised using TreeTagger, and multiword expressions were identified with FLELex [26]. Participants were all Dutch-speaking Belgians whose proficiency level in French was A2 or B1.

A model predicting the vocabulary knowledge of learners based on their fluency level served as a baseline to which the other systems would be compared. The authors used FLELex to attribute a CEFR level to each word, based on the first level into which the word occurred. The model would then predict a word to be complex if its level was higher than that of the learner. The authors pointed out several shortcomings of that approach, such as the fact that no other learner variable than their proficiency level was taken into account, as well as the use of graded manuals rather than annotation data to produce lexical complexity labels. The other two models were created using a set of linguistic predictors inspired by Gala *et al.* [27]. The features consisted in the word's part of speech, frequency in each of the six CEFR levels, first occurrence in FLELex, length in letters, number of orthographic neighbours, number of complex orthographic patterns it contained, and number of meanings according to BabelNet [63]. All predictors were retrieved for each target word as well as for their preceding and following word. Both models took the form of neural networks [90].

The first of the two systems modelled the lexical knowledge of all learners with the same proficiency level, based on the mean of their annotations rounded to 0 or 1. The second modelled each learner's vocabulary individually, using only their own annotations. An analysis of the results showed that the personalised models considerably outperformed the baseline for all learners, while only the system modelling the knowledge of B1 learners managed to beat it. This was interpreted by

the authors as a consequence of the larger heterogeneity in A2 readers' annotations compared to those of the B1 level. In such a case, personalised models thus seem all the more interesting. The authors argued that their learner models could have been further improved by taking individual predictors into account, such as the person's age or L1. They indeed noticed that certain words which were actually very similar in French and Dutch were labelled as difficult for the learners, based on intrinsic complexity predictors. Integrating a dictionary of cognates in the models could therefore prove useful [90].

In her PhD thesis, Tack extended the work described above to further explore the value of personalised difficulty prediction. Using the same methodology as before, further data were collected from annotators spanning four proficiency levels (A2 to C1). A first trial involved 9 participants with different mother tongues and education levels. They varied greatly in their perceptions of lexical difficulty, as evidenced by the low inter-rater agreement score obtained from the trial (Krippendorff's α of 0.26). Even when grouped by proficiency, the participants still obtained low agreement scores (0.23 for A2 learners, 0.30 for B1). The values are reminiscent of the score obtained in CWI-2016 (α of 0.244), and show that learners with different backgrounds will not have homogeneous vocabulary knowledge, even at similar proficiency levels. By contrast, the agreement scores obtained in the second trial, which involved 47 native Dutch speakers, were much higher: α of 0.40 for A2 readers, 0.36 for B1, 0.51 for B2 and 0.47 for C1. Those higher scores confirm the impact of the L1 on vocabulary knowledge, and show once more the need to take individual differences into account when predicting difficulty [89].

The data collected in both trials were aggregated and used to train two types of predictive models. First, a generalised linear mixed-effects model was created, using a total of 18 word predictors and 3 learner-specific features. Word surprisal, number of previous exposures to the word in the reading task and the term's morphosyntactic function were found to have the most explanatory power among the fixed factors. As

for the learner-specific effects, those were also found to increase the performance of the model: compared to a model making averaged, learner-independent predictions, the system achieved an increase of 8% in predictive power. However, that predictive power remained quite low (25%). As a result, deep learning approaches were investigated: four neural networks were created, using word and character embeddings as inputs. Two of the models were feedforward neural networks that did not take context into account; the other two were bidirectional long-short term memory networks and made contextualised predictions. One of those two was also personalised, while the other used averaged difficulty labels. The contextualised and personalised model performed consistently better than the others, with a significantly higher discriminatory power (64% to 74% on average). Personalising the predictions also led to a significantly higher correlation with learner judgements, although the same cannot be said of contextualised models compared to non-contextualised ones. Surprisingly, learners' L1 was not found to be a significant predictor of difficulty. This can perhaps be explained by the small variety in mother tongues among annotators, as most of them were native speakers of Dutch. Indeed, only five annotators had a different L1: one was a native speaker of Chinese, another of Japanese, and three of Spanish [89]. Those small samples might thus not have been enough to explore the predictive power of a reader's L1. The learners' proficiency level and unique identifier, however, were found to significantly improve predictions.

1.5.2 Discussion

Although the studies presented above used a variety of methods, they all led to the same conclusion: using personalised predictions of difficulty leads to better-performing systems. When word-related features were used, they were usually similar to those employed by the CWI models described in section 1.4, which shows that factors of lexical complexity are also valuable for difficulty prediction. An interesting finding from the studies was the impact of the topic on individual difficulty judgements: a person interested in music, for instance, is more likely to be familiar

with common music-related terms than someone who does not share that interest. Moreover, the proficiency level of a learner was found to be correlated with difficulty judgements, to the point that making averaged predictions for learners of the same proficiency level also proved an improvement compared to a learner-independent baseline. This seems to apply especially for more advanced learners, who tend to be less heterogeneous in their difficulty judgements than lower-proficiency learners. The impact of accounting for readers' L1 on the accuracy of predictions was not clear, and requires further analyses. Finally, a last interesting finding was the effect of word exposure on perception of difficulty: a word that had already been previously encountered in the reading task tended to appear less difficult.

While this overview of studies on personalised difficulty predictions outlined some interesting results, it is not yet clear which individual differences have the strongest impact on lexical difficulty judgements in a reading task. Exploring the relationship between different individual features and difficulty annotations should therefore prove useful. Similarly, it would be interesting to investigate whether accurate predictions of difficulty could conceivably be made based on the demographic profile of a reader alone. Indeed, while fully-personalised models show very promising results, they require prior annotation data from each particular reader for whom to make difficulty predictions. As a result, such an approach will not always be possible, and predicting difficulty for a target audience could be a valuable alternative.

1.6 Conclusion

This chapter has laid the foundations on which the remainder of this dissertation will be built. Working definitions for a set of essential yet ambiguous constructs were provided. The concept of word was defined as a minimal free form with a unified semantic or grammatical role, and that of the multiword expression as a word which, in its written form, is made up of several graphical units. We refer to the processing costs of a word for a specific individual as lexical difficulty, and to

the inherent complexity of a word as lexical complexity. The former is a relative concept, impacted by individual characteristics, whereas the latter is absolute and reader-independent.

Lexical difficulty was shown to be a fundamental notion that has repercussions on both vocabulary development and reading comprehension. It is related to the concept of the mental lexicon, which can be represented as a vast network of words linked together by a variety of relationships, corresponding to the different aspects of word knowledge. The focus of this dissertation is on the aspect of meaning, as reading is a receptive task which, in its most basic form, comes down to mapping a form to its signification. This is referred to as “meaning recall”. Word knowledge is thus conceptualised as a continuum going from a total lack of meaning recall (unknown word) to an advanced level of recall (familiar word). Learning the vocabulary of a language amounts to progressing along that continuum, as well as developing the other aspects of lexical knowledge. It is a highly variable process, which changes from person to person based on factors related to the context of acquisition (educational, social, political etc.), the personality and attitude of the learner, and their prior linguistic knowledge. Indeed, lexical transfer occurring from a learner’s L1 to their L2, or from their L2 to their L3, has the potential to facilitate the vocabulary acquisition process. Reading in a foreign language is equally impacted by transfer, with cognates being a prime example of the facilitating effect of crosslinguistic influence.

Predicting lexical complexity on the basis of annotated data has been the object of three shared tasks in the past few years. The way those tasks have been carried out has seen some interesting developments, among which the shift from binary annotation data to continuous difficulty labels, as well as the inclusion of MWEs in the set of target words. The three CWI datasets were collected using different approaches, and their distinct implications and limitations highlighted the importance of defining a methodology that best serves one’s objectives when conducting

a study. Frequency and length were consistently found to be the best-performing predictors of lexical complexity. Moreover, taking context into account resulted in predictions of a higher quality.

The benefits of moving the focus of predictions from complexity to difficulty were defended throughout the chapter. The few studies having explored the possibility of personalised lexical difficulty models all reported an increase in performance compared to reader-agnostic models trying to predict the needs of an individual based on those of a heterogeneous group. Proficiency level in the target language was shown to be a strong predictor, especially so among advanced learners. Elements such as the topic of a text or whether a learner had already been exposed to a target word in the reading task were also found to impact the perception of difficulty. The effects of the L1 and potential L2s of annotators were not investigated much. Similarly, the studies focused mainly on making fully-personalised predictions of difficulty, rather than attempting to anticipate the needs of a new reader based on their profile. As a result, those questions still remain to be explored.

Chapter 2

Methodology

2.1 Introduction

The insights obtained from the literature review make it possible to formulate hypotheses concerning the research questions presented in the introduction. Several studies have found that more homogeneous groups achieved higher inter-rater agreement scores than groups with different L1s and proficiency levels. Moreover, the few systems that have been designed with demographic information as predictors consistently achieved better results than reader-independent models. As such, the following hypotheses can be put forward for research questions 1 and 2:

Research Question 1

Is it worth taking demographic information into account when predicting lexical difficulty for a given reader? In other words, is there more between-group variability as regards perception of difficulty than there is within-group?

RQ 1: Hypothesis

The within-group variability in difficulty judgements should be lower than the one found between groups.

Research Question 2

If the answer to RQ 1 is “yes”, which individual variables have a demonstrable impact on lexical difficulty perception?

RQ 2: Hypothesis

The individual factors expected to have the strongest impact on the perceived difficulty of a word are a learner’s mother tongue (plus any other language in which they are fluent) and proficiency level. Their age, education level and time spent learning the language both in a native and a non-native context could also be found to have an effect on the words that they will perceive as difficult.

As for research question 3, section 1.2.2.2 has allowed us to put forward our third hypothesis:

Research Question 3

Is there a direct link between the proportion of words perceived as difficult by a given reader and the quality of their text comprehension?

RQ 3: Hypothesis

The number of words in a text that a reader is familiar with is expected to have a positive impact on how well they can answer comprehension questions after reading. This should be especially true if they go below the threshold of 95% of known words, in which case comprehension will be hindered.

In order to test these hypotheses, individual judgements of lexical difficulty were collected using a custom-made online reading interface. Participants of a level B1 or B2 in French were asked to provide demographic information, then to read,

annotate and answer questions about two texts, selected at random. The different methodological decisions made when designing the study are outlined in the following sections.

2.2 Data Collection Process

2.2.1 Demographic Questionnaire

The aim of the present study is to determine whether taking a reader's profile into account has the potential to help better predict their needs. As a result, deciding which demographic data to collect was an important step of the study design. In an article devoted to L1 influence on a learner's interlanguage, Jarvis [36] identified a series of individual variables that should be taken into account in such a study. Those variables were:

1. A learner's age;
2. Their personality, motivation and language aptitude;
3. Their background (social, educational and cultural);
4. Their linguistic background (L1 and L2s);
5. Their language exposure (type and amount);
6. Their proficiency in the target language.

Estimating the factors outlined in the second point would be too time-consuming, and they were thus not used in this study. This should not be too much of an issue, however, as that kind of information is not likely to be available when devising reading tools for a target audience. Similarly, information about the readers' social background were not considered. All the other variables were included in the study.

The selected demographic information can be split into two main types: numerical or ordinal variables, and nominal variables. The former are expected to correlate with difficulty judgements in a linear fashion, while the latter will be used to group readers with similar profiles. The demographic questions that learners were

asked to answer are presented below, according to the variable type to which they correspond.

2.2.1.1 Numerical and Ordinal Variables

- **Age** (in years);
- **Proficiency level in French**: Expressed based on the CEFR scale;
- **Education level (current or highest achieved)**: Chosen from one of the following five options: primary education, secondary education, higher education (undergraduate), higher education (postgraduate), PhD;
- **Time spent learning French in a native context**: In weeks, months or years;
- **Time spent learning French in a non-native context**: In weeks, months or years.

The first two variables (age and target language proficiency) were already found by Paetzold and Specia [66] to correlate with the number of words marked as complex by annotators in their study. The third acquired information about the reader's educational background, while the last two correspond to the language exposure variable recommended by Jarvis [36].

2.2.1.2 Nominal Variables

- **Country of origin**: This variable provided some information about the cultural background of the reader. It was selected as factors such as the nature of the education system or the prevalent attitude towards a target language in a country are likely to have an impact on the language learning process of its inhabitants;
- **Mother tongue(s)**: This was an important variable as, for reasons outlined in section 1.3.2.2, a reader's L1 (or L1s) is likely to affect their lexical knowledge. The languages gathered with this question were expressed in 2 additional ways: language family and language sub-family;

- **Reading or visual impairments:** This factor was included to explore the difference in needs between readers with a reading or learning disability and those without. As the only impairment which was selected by some annotators was dyslexia, the variable was converted to a binary feature indicating whether a participant was dyslexic.

2.2.1.3 Hybrid Variable

- **Other languages known and proficiency level in each of those languages:** As mentioned in section 1.3.2.2, crosslinguistic influence can also occur between L2s, and could have a facilitating effect on vocabulary acquisition in a foreign language, which is why this predictor was included.

2.2.2 Text Annotation

After they had provided the relevant demographic information, the participants were presented with two randomly selected texts to read and annotate by highlighting words that they found difficult, according to a predefined annotation scheme. The following sections will further describe the process and methodological choices made concerning this annotation step.

2.2.2.1 Corpus Design

It was decided early on that each participant would be asked to annotate two texts, to avoid making the experiment unreasonably long to complete. This then begged the question of how many texts in total should be used in the study. Too small a number would limit the amount of words annotated as well as the diversity in topics, whereas too many would entail the risk of not receiving enough annotations per text. A total of four texts, each containing around 500 words, thus seemed like a safe middle ground, as it would amount to a dataset of close to 2000 words while only requiring a total of about 20 annotators in order to obtain a minimum of 5 annotations per text.

ID	Title	Level	Words
B1_A	<i>Les chenilles, les abeilles et l'effet papillon</i>	B1	432
B1_B	<i>C'est prouvé : rôler est bon pour la santé</i>	B1	472
B2_A	<i>Un jardin partagé crée du lien en plein Paris</i>	B2	564
B2_B	<i>Et si on arrêta de se mettre la pression autour des séries qu'il "faut" regarder ?</i>	B2	448

Table 2.1: Corpus Description

Texts were extracted from a selection of teaching manuals targeted at intermediate learners of French (level B1 or B2 on the CEFR scale). They had to meet several criteria, namely being of an informative nature, containing about 500 words, and being varied in their topics so as to maximise the diversity in vocabulary. Table 2.1 presents the main characteristics of the four texts chosen for the study. The CEFR levels reported were derived from the manuals in which each text appeared and were validated by two teachers of French as a foreign language. The word totals include the title and are based on a segmentation of the texts into lexical units, which will be described in section 2.2.2.3.

2.2.2.2 Random Text Selection

The participants were presented with one text of each level. Which texts would be annotated by a given participant and in what order was determined at the launch of each experiment. First, one text was randomly selected from both levels using JavaScript's `Math.random()` function. The selected texts were then shuffled and presented to the participant in a randomised order.

2.2.2.3 Lexical Units

Section 1.2.1.1 discussed the difficulty of formulating a satisfactory definition of what a word is, and proposed the following working definition: a word is a minimal free form which can convey either semantic or functional information. When a word is made up of several graphical units, it is referred to as a multiword expression.

Texts went through some preprocessing steps in order to identify the units that participants would ultimately be able to annotate. A simple script was used to tokenise orthographic words and separate them from any punctuation, apostrophes and hyphens included. Multiword expressions were then manually identified based on a simple inclusion criterion: their presence as an entry or sub-entry in the Trésor de la Langue Française informatisé (TLFi) [93]. Defining the existence of a word based on whether it can be found in a dictionary is a controversial practice. Indeed, there are a variety of reasons why a word would not be used as an entry, such as it being a neologism, an inflected form, or being too specialised for a general purpose dictionary. Moreover, lexicographers tend to define entries based on written forms, and finding a multiword expression as a headword is thus uncommon [15]. The use of an electronic dictionary, however, should circumvent those limitations to a certain degree: as they are not limited by size and can be updated more regularly than a paper version, electronic dictionaries are more likely to contain newer, rarer or more specialised entries. Additionally, they make it possible to search directly for an inflected form or a multiword expression, and thus do not limit access to entries solely from headwords. The inclusion criterion for MWEs used in this study should therefore prove a valid way of reducing subjectivity in the text segmentation process.

The multiword expressions identified in the four texts of the corpus are presented in table 2.2. All met the requirement presented above, with the exception of “binge watching” and “binge watcher” (to binge-watch), which did not exist in the TLFi. The expressions are both neologisms formed on the basis of English words, and therefore only exist in French as a unit. Including them as MWEs thus seemed the correct course of action.

2.2.2.4 Annotation Process

During the experiment, the segmented texts were reconstructed and displayed for the participants to read. When hovered over, the background colour of units that

Text	N°	Multiword Expression	Translation
B1_A	1	En effet	Indeed
	2	Révolution industrielle	Industrial revolution
	3	Par conséquent	Therefore
B1_B	1	Sautes d'humeur	Mood swings
	2	Espérance de vie	Life expectancy
	3	A la longue	Over time
	4	En moyenne	On average
	5	Liberté d'expression	Freedom of speech
	6	(se faire) Marcher dessus	(be) Walked over
	7	En effet	Indeed
	8	En permanence	Constantly
	9	De plus en plus	Increasingly
	10	Broyer du noir	Mope
	11	Cercle vicieux	Vicious cycle
B2_A	1	Entendu parler	Heard about
	2	Au prix de	At the cost of
	3	Parcours du combattant	Obstacle course
	4	A l'origine	Originally
	5	Du coup	Thus
B2_B	1	Binge watching, binge watcher	Binge-watching, binge-watch
	2	A part	Except
	3	A l'image de	Like
	4	Tant pis	Never mind
	5	Prendre conscience	Realise
	6	Avant tout	First and foremost
	7	Prendre part	Take part
	8	Faire la part des choses	Distinguish (between)
	9	Faire pareil	Do the same
	10	Sortir par les yeux	Exasperate

Table 2.2: MWEs in the Corpus

could be annotated changed, in order to mark their boundaries. The readers could then click on the words that they personally found difficult in order to cycle through the different levels of annotation described in section 2.2.2.5, each identified by a different background colour and border style. The annotators had access to a given text in its entirety, and were thus free to read back previous passages for context. This made the reading task more natural, but also made it possible for participants to go back and change the annotation of a word later on, although they were asked not to. Indeed, some words initially found difficult by a given reader might be made clearer by the rest of the text, and one could therefore be tempted to go back on previous decisions. Nonetheless, the risk of having some participants change an annotation later on seemed preferable to that of readers not being able to correct an erroneous annotation.

Unlike with most other studies presented above, participants in this experiment were able to click on any lexical unit, including grammatical words, and could highlight as many or as few words as they wanted. This meant that an annotation scale such as the ones presented in sections 1.4.4 and 1.5.1.2, for instance, could not conceivably have been used in this study: as they include several levels for words whose meaning is known by the reader, those scales could not be applied to a free annotation setting without essentially asking participants to annotate all the words of a text. As a result, the methodology employed here is a combination of the studies reviewed previously. On the one hand, like in Yimam *et al.* [100] and Tack [89], participants were asked to select words that they found difficult in context, and simple words were thus those that were not annotated by a reader. Following Ehara *et al.* [47] and Shardlow *et al.* [20], on the other hand, annotations were made using a continuous scale distinguishing between several degrees of difficulty. The scale used in this study is presented in the following section.

2.2.2.5 Annotation Scale

Several lexical difficulty scales have been used in the literature, some of which were presented in the previous chapter. The scales encountered while preparing the present study all shared one important characteristic: they defined words that had never been seen before by a reader as corresponding to the highest level of difficulty. While it is true that new words are likely to be difficult, they are not always guaranteed to pose a problem for readers. A new word's similarity to a term in another language or the context in which it appears, for instance, can have a facilitating effect on its comprehension. In fact, inferring the meaning of unknown words is a central part of the reading process [48]. It therefore seemed important for our scale of lexical difficulty to account for those transparent words, and to avoid defining a word's level of difficulty by whether or not it has been seen before.

The scale used in this study is presented in table 2.3, along with four other difficulty scales included for comparison. All four scales used five levels, which have been labelled as going from “very easy” to “very difficult”, following Shardlow *et al.* [85]. For the reason outlined in the previous section, our scale only includes one level for non-difficult words. Those consist in the words that did not receive a difficulty annotation from a given reader; therefore, no distinction is made between “very easy” and “easy” terms. As this dissertation focuses on the identification of words which might hinder comprehension for certain readers, knowing that a word is easy for a target group should prove enough. Exactly how easy a given word is, although interesting, is not essential information in this case.

The scales presented in table 2.3 all base the definition of their difficulty levels on whether the annotator knows the meaning of a given word, which makes sense in a reading context. However, they also include more debatable factors. For one, both the VKS and Sugiyama's revised version look at the ability of an annotator to use a word in a sentence or to define it as a way to determine how familiar they are with it [96, 88]. Although it is true that being able to produce a word requires more



Figure 2.1: Example of Text Annotation

in-depth lexical knowledge, it is probably preferable to focus on receptive skills in the context of a reading task. On a more practical level, asking annotators to write a sentence or a definition in order to assess their knowledge of a word would make a reading task less natural, and would not have been feasible in our study design. Moreover, all four scales made use of whether the reader had seen a word before as an indication of difficulty. As mentioned above, previous exposure to a word is not necessarily always correlated with perceived difficulty.

The scale designed for the purposes of this study only uses comprehension of a word's meaning as the defining feature of lexical difficulty. The easiest words for a person are those with which they were already familiar before reading the text. Previously unknown words that are made transparent by context or because they are recognised as the cognate of another word represent the next level in difficulty. Then come words which the reader finds vague: they have an idea of what the term might mean, but its signification is still unclear. Finally, completely opaque words, the meaning of which the reader has no idea of, embody the highest possible degree of difficulty.

Scales	Levels				
	Very Easy	Easy	Neutral	Difficult	Very Difficult
Vocabulary Knowledge Scale (VKS) [96]	“I can use this word in a sentence: (Write a sentence.)	“I know this word. It means (synonym or translation).”	“I have seen this word before, and I think it means (synonym or translation).”	“I have seen this word before, but I don’t know what it means.”	“I don’t remember having seen this word before.”
Word Knowledge Scale [21]	Absolutely know the word’s meaning	Probably know the word’s meaning / able to guess the word’s meaning	Absolutely saw the word before but don’t know its meaning / tried to learn the word before but forgot its meaning	Probably saw the word before	Never saw the word before
Revised VKS [88]	Known word (certain of the meaning)	Word probably known (correct meaning)	Word probably known (incorrect meaning)	Word encountered before (uncertain meaning)	Word never encountered before (unknown meaning)
CompLex Scale [85]	Very familiar to an annotator	Aware of the meaning	Neither difficult nor easy	Unclear of the meaning, but may have been able to infer it from the sentence	Never seen before, or very unclear
Present study	Easy word (know the meaning)		Transparent word (can guess the meaning)	Vague word (unsure of the meaning)	Opaque word (meaning unknown)

Table 2.3: Comparison of Lexical Difficulty Scales

By default, all the words of the texts presented to annotators were set to the first level (easy). Clicking on a word once would then mark it as transparent, twice as vague, and three times as opaque. A fourth click would cancel the annotation, by cycling back to the original level. Figure 2.1 provides an illustration of a text being annotated by a participant.

2.2.3 Comprehension Questions

None of the text annotation studies presented in the first chapter used comprehension questions about the text afterwards to check that participants were reading for meaning, which Tack said could be a limitation [89]. Such questions were included here, to obtain insights into the impact of lexical difficulty on global text comprehension. After reading a text, participants were asked to answer 5 multiple choice questions (MCQs) about some key information it contained.

Participants could take as long as they needed to complete the study. The time they took for each text was recorded, measured from the start of the annotation step to the moment when they submitted their answers to the MCQs.

2.2.3.1 Question Design

Each question was followed by three possible answers: the correct answer, an answer that was incorrect but was based on information that was present in the text, and another incorrect answer which was not mentioned in the text. The first type was expected to be selected by somebody who did not fully understand what they read, while the second would indicate that the participant either did not read the text, or understood the relevant passage very poorly. The questions focused on key information, as readers no longer had access to the text when answering. The aim was not to test their memory, but to evaluate whether they had been able to understand the main points of what they read. All questions and proposed answers were reviewed and approved by a teacher of French as a foreign language.

2.2.3.2 Feedback

Once participants submitted their answers, they were given some feedback on how well they did. Their answers to each question were displayed, along with a correction whenever they got a question wrong. They were also shown a short message of encouragement, which depended on the proportion of correct answers they selected.

2.2.4 Finding Participants

In a crowdsourcing setting, ensuring the quality of the data collected is a challenge. Knowing whether the demographic information we gathered was accurate, for example, would have been impossible if the study had been open to anyone online. It thus seemed preferable to keep some control over the identity of our participants, by reaching out to people who we knew for a fact were learners of French as a foreign language (FFL).

We asked for help from FFL teachers from the SUL (Aix-Marseille University) and the ILV (UCLouvain), who kindly accepted to send out the link of the study to their B1 or B2 students of French. All participants to our study were thus currently enrolled in a language course in a French-speaking context, which might impact the results. Another advantage of this approach was that we did not have to rely on self assessments of French proficiency from the annotators, as they had taken a placement test in the beginning of the academic year.

Participation was done on a voluntary basis. To encourage teachers to ask their students to complete the study, we made it possible for them to access their classes' results. Offering some form of compensation to the students who took part would have been likely to result in a higher number of annotations; however, the teachers' precious help led to a satisfactory number of participants nonetheless.

2.2.5 Discussion

The aim of the methodological choices presented above was to allow for high-quality data to be collected, according to the conclusions drawn from the literature review. Shardlow *et al.*'s [85] guidelines were followed as closely as possible in this study: a four-point annotation scale was used in order to obtain continuous difficulty data, words were annotated in context, all tokens received several annotations as each participant annotated two of the four texts making up the corpus, and both single words and multiword expressions were considered for annotation. Moreover, the ambition of the study was to collect data from annotators with a diversity of profiles; however, proficiency levels were restricted to B1 and B2 learners for practical reasons. Similarly, the texts chosen to constitute the corpus only belonged to the informative genre, although their varied subjects should allow for a high level of lexical diversity. Finally, the limited size of the corpus (about 2000 words) made the repetition of many potentially difficult words across texts unlikely, although several words were repeated within a same text.

The review of previous studies presented in chapter 1 showed certain characteristics to be valuable for a new lexical difficulty dataset. Those were taken into account in the design of this study: thorough demographic information was gathered, and the text in which each word occurs was recorded, which should allow for the impact of the topic on comprehension to be investigated. The fact that it used a language other than English is also a strong point for our dataset.

Most of the shortcomings identified in this study design could probably be solved by using a bigger, more varied corpus, and by gathering annotations from participants representing more diverse proficiency levels. The present dissertation only aims to explore the validity of the methodology proposed here in order to inform future larger-scale projects, which should remain feasible despite all recommendations not being met perfectly.

2.3 Data Analysis

Different methods were employed to draw conclusions from the data collected. Some potential correlations and trends in the data were explored using statistical tests. Inter-rater agreement scores were then computed to find out whether annotators who had a similar profile agreed more in their lexical difficulty judgements. Lastly, to achieve a deeper understanding of which factors have an impact on the perceived difficulty of a word, ordinal regression models were fitted on our dataset. Their aim was to predict the label attributed by a given participant to a word based on a set of predictors, which included the mean annotation value received by a word, text and word features and individual features. The intuition was that, since difficulty can be defined as intrinsic complexity modified by individual characteristics, using a series of personal predictors alongside a measure of complexity should uncover the factors which impact difficulty perception. Previous studies that attempted to predict complexity used an aggregation of individual judgements as gold standard for their models; as such, using the mean of annotations for a given word should allow us to efficiently capture something of its complexity. The individual features would then serve as a way to predict by how much and in which direction participants will deviate from that mean, based on their specific profiles. As for word and text features, they were added to account for potential differences in what makes a word difficult for a particular profile. For instance, certain readers might struggle more than others with words that have a high number of homographs.

The list of individual predictors was already provided in section 2.2.1. Text characteristics included the ID of the text as well as its CEFR level (B1 or B2). Word features comprised frequency measures extracted from FLELex [26] (frequency of the word in each CEFR level and in the whole corpus) and Lexique3 [64] (frequency in films and in books), part of speech, number of homographs and homophones, and number of syllables. Most features were obtained from a simple lookup in the Lexique3 dataset. The lemma of each target word was then extracted in order to

derive frequency information from FLELex. When a written form was ambiguous (in that it corresponded to different lemmas or parts of speech), the version with the highest book frequency in Lexique3 was selected. This overly simple selection criterion has obvious limitations; however, the aim here was not to build a state-of-the-art predictive model, but only to explore whether certain features have an impact on difficulty perception. As ambiguous words in our corpus were typically short, high-frequency words (such as “les”, “un” or “a”), the simplistic selection process should not significantly impact the explanatory power of our models. One last criterion, which can be qualified as being “hybrid”, was the position of a word in the text in which it appeared. It was included as it seemed possible that the same word presented at the beginning of a text, when the reader did not yet have much information about the topic, would be perceived as more difficult than that same word appearing later, when contextual guessing of its meaning could be facilitated.

Several preprocessing steps were implemented before building our regression models. All numerical variables were scaled and centred, except for the average difficulty of the words, which was already constrained to a very small range of values (0 to 2.4). The full dataset was randomly split into training and test, with 80-20% odds. This resulted in a training set of 25367 observations, while the test set contained 6417 data points. Models were fitted on the training set by a manual bidirectional stepwise selection process: starting from a baseline model, features were iteratively added if their inclusion improved the performance of the model, measured in terms of the AIC score (Akaike Information Criterion). Once no more predictors could be added, we discarded any variable whose removal from the model led to a better performance.

Originally, the plan was to use generalised linear mixed models in order to account for both fixed and random effects. However, considering the nature of our response variable (the difficulty level of a word on the annotation scale presented above), ordinal regression models were found to be more appropriate. Three types

of models were experimented with: cumulative link mixed models (CLMMs), simple cumulative link models (CLMs), and multinomial log-linear models (MLMs). The first type allows for both fixed and random effects to be used, while the second only accounts for fixed effects. Both CLMMs and CLMs consider that the levels of the response variable are ordered; by contrast, MLMs treat them as unordered. In practical terms, it means that CLMs compute a single slope coefficient for the model (based on the predictors provided) and one intercept per level of the dependent variable, whereas MLMs compute both a different slope and intercept for each level (bar one). The results obtained from the different model types will be discussed in section 3.3. To learn more about cumulative and multinomial regression models, see Hosmer *et al.* [33].

Chapter 3

Results

The data collected in this study will be analysed as follows. First, a series of descriptive statistics will be presented, in order to become better acquainted with the data and to explore some potential trends and correlations. Next, inter-rater agreement scores will be computed for different groups, in order to answer our first research question (RQ 1). Finally, some statistical models will be trained, so as to investigate which features are significant predictors of perceived difficulty.

3.1 Description of the Data

Annotations were collected in two steps. First, a pilot study was carried out in November 2021 among a B1 FFL class, in order to evaluate the design decisions outlined above. The study was then launched properly in March 2022, this time targeting students from various classes. The two studies were essentially similar, with the difference that some demographic data were missing in the pilot version. As a result, some analyses will use an aggregation of all data, while others will be based solely on those collected in the main study.

3.1.1 Participants

In total, 39 learners of French took part in the study. This number was obtained after discarding invalid results, *i.e.* cases where someone had not annotated any text, or had spent less than 1 minute on a task (text annotation and MCQ answering). Table 3.1 describes the distribution of participants across the versions and proficiency levels. Although this study originally aimed to account for both B1 and B2 learners of French, almost all participants belonged to the B1 proficiency group. As a result, comparing the two groups reliably will unfortunately be difficult.

Study	B1	B2
Pilot	16	0
Main	20	3
Total	36	3

Table 3.1: Number of Participants per Proficiency Level

The different mother tongues of participants are presented in table 3.2. The languages were grouped in main and sub-families, as the literature has shown that learners whose L1s belong to the same language family were similar in their perception of lexical difficulty (Section 1.4.1). Two participants had two L1s (Spanish and Catalan for one, English and Afrikaans for the other), which is why the numbers reported in the “B1” column add up to 38 instead of 36. In most cases, there were not enough speakers of a same L1 for an inter-rater analysis to be carried out at the language level. However, agreement scores could be computed for certain language families, mainly Indo-European ($N=15$), Germanic ($N=9$) and Sino-Tibetan ($N=7$).

The 16 speakers of Malaysian reported in table 3.2 made up the group of participants who took part in the pilot version of this study. They were a very homogeneous group in terms of age (18 years old), country (Malaysia), education level (undergraduate) and proficiency level in French (B1). They should therefore obtain a high

Mother Tongue			Proficiency	
Language Family	Sub-Family	Language	B1	B2
Indo-European	Germanic	Afrikaans	1	0
		Dutch	1	0
		English	7	0
	Italic	Catalan	1	0
		Spanish	3	0
	Balto-Slavic	Russian	2	0
Afro-Asiatic	Semitic	Arabic	0	1
Sino-Tibetan	Sinitic	Cantonese	1	0
		Chinese	4	2
Austronesian	Malayo-Polynesian	Malaysian	16	0
Other	Japonic	Japanese	1	0
	Koreanic	Korean	1	0

Table 3.2: Summary of L1s of Participants

agreement score, according to our first research hypothesis (RH 1). The group which took part in the second version of the study, on the other hand, was far less homogeneous. Figure 3.1 gives an overview of the demographic data provided by this second group. Most participants were in the 20-25 age range, and all but one were under 40 years old. Their education level ranged from secondary to doctorate, with almost 50% of annotators holding or currently studying towards an undergraduate degree. About half of participants reported having spent less than a year studying French both in a native and in a non-native contexts, while some had been learning French for several years in either type of environment. None of the numerical variables were normally distributed (Age: $W = 0.601$, $p < 0.001$, Time in a native context: $W = 0.735$, $p < 0.001$, Time in a non-native context: $W = 0.853$, $p = 0.003$). In addition to French, participants spoke between 1 and 3 other languages, with a median of 2. The foreign languages spoken by participants were English ($N=12$, A2 to C1),

German ($N=2$, A1 and A2) and Chinese ($N=1$, level C1). Two annotators had dyslexia; the rest did not report having any reading or learning disability. Finally, participants were diverse in terms of country of origin as well: they came from China ($N=6$), the United States ($N=3$), Vietnam ($N=2$), Russia ($N=2$), Japan ($N=1$), Venezuela ($N=1$), Mexico ($N=1$), South Africa ($N=1$), Syria ($N=1$), the Philippines ($N=1$), South Korea ($N=1$), Hong Kong ($N=1$), the Netherlands ($N=1$) and Spain ($N=1$).

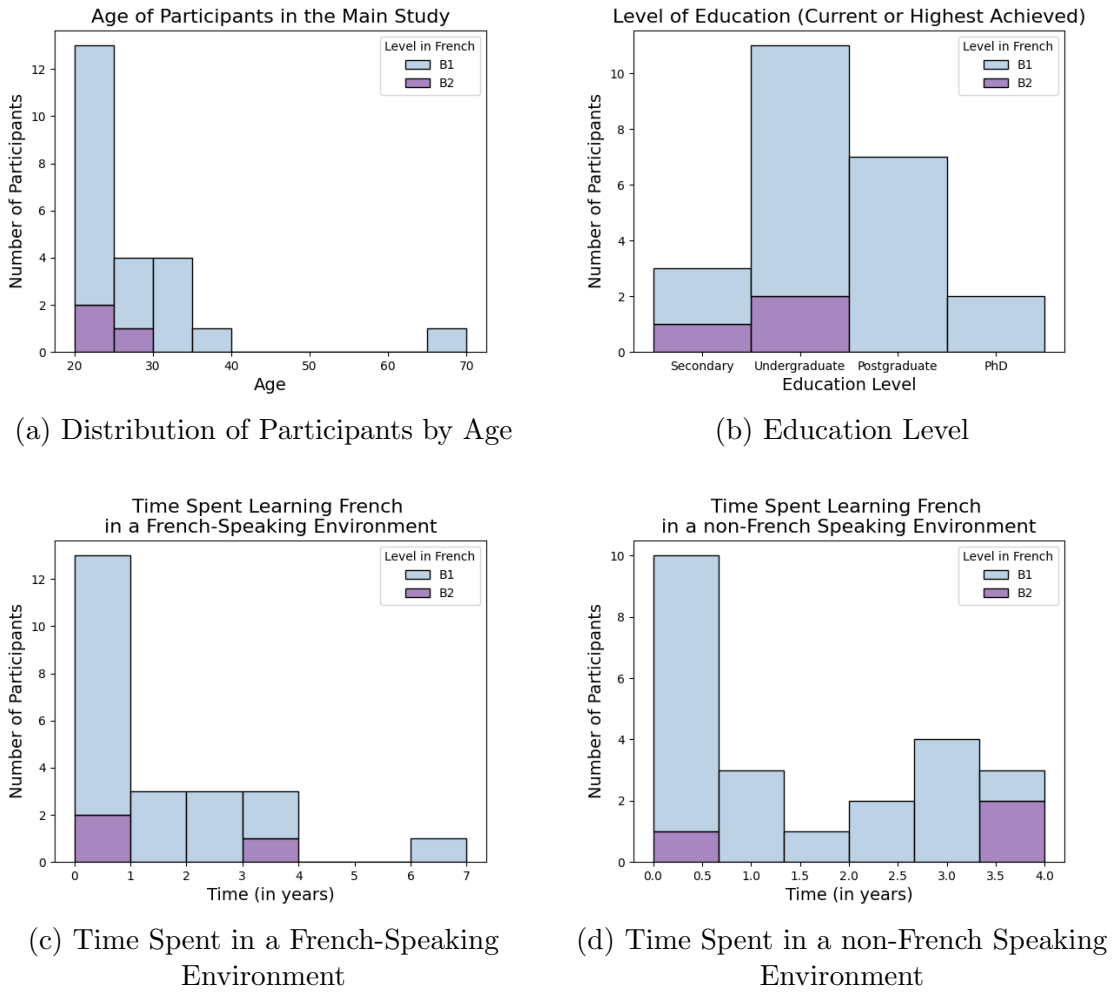


Figure 3.1: Demographic Information of Participants in the Main Study

To sum up, the pilot version of the study was completed by a very homogeneous group of 16 annotators, while the main version received answers from 23 participants with diverse backgrounds. According to our first research hypothesis (RH 1), the first group should thus obtain higher agreement scores than the second, or than both grouped together. This will be explored in section 3.2.

3.1.2 Annotated Texts

As mentioned previously, the texts that a given annotator was asked to read were randomly selected whenever they started answering the study. The pilot version only included three out of the four texts (text B1_B was not used), in agreement with the teacher who asked his class to take part. This was done so that one text (text B1_A) would be read by all annotators who completed the study, and could therefore be reused in class. As a result, the excluded text received fewer annotations in total than the others, as is shown in table 3.3. Although the number of participants per text is markedly unbalanced, each token in our corpus received at least 6 annotations, which meets the fourth recommendation outlined in section 1.4.3.2.

Characteristic	Text			
	B1_A	B1_B	B2_A	B2_B
Annotators	28	6	14	20
Number of words	432	472	564	448
Average task time (min.)	10.03	8.80	11.64	10.79
Standard deviation (min.)	4.06	2.43	6.54	5.29

Table 3.3: Number of Annotators and Time Spent on the Task per Text

Table 3.3 shows the average time that participants spent on a task (reading + MCQ answering) for each text. The slightly higher average and standard deviation for B2-level texts indicate that there might be increased variability in the time it takes to complete a task when using more advanced texts. Figure 3.2 indeed shows a very high amount of variation in the answers for text B2_A, which could also be due to the fact that it is by far the longest in our corpus. Nonetheless, text B2_B shows more variation than text B1_A, despite the two being almost equal in length. It thus seems that the texts labelled as being of the level B2 were effectively more difficult for some of the participants than those of a B1 level, and therefore required more time to read and / or answer questions about them.

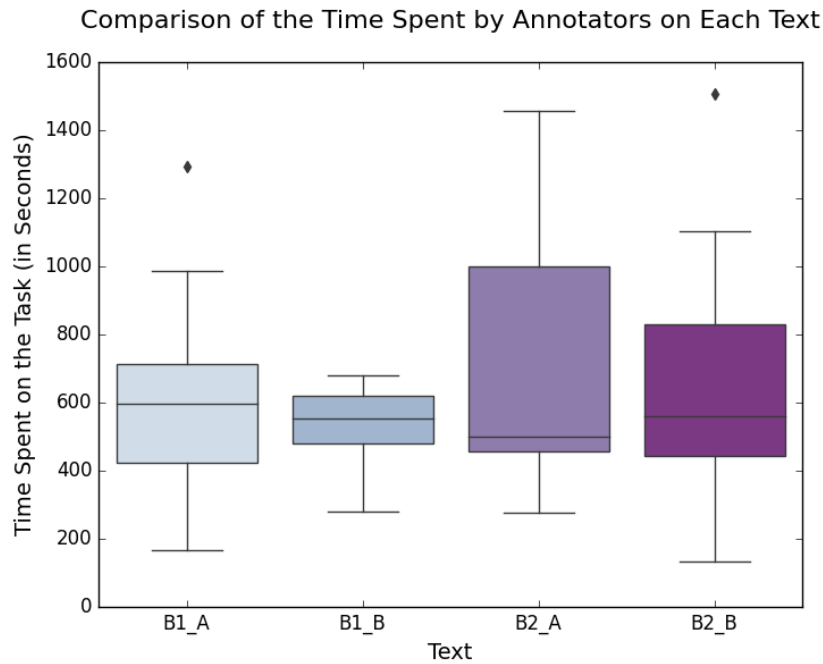


Figure 3.2: Time Spent on the Task for Each Text

Since B2-level texts were more advanced than B1-level texts, one might expect them to contain a higher proportion of words found difficult by participants. However, table 3.4 shows this not to be the case: on average, participants perceived almost the same proportion of words as easy in each text. There was nonetheless some level of variation between readers: the proportion of non-annotated words ranged between 91.44% and 99.54% in text B1_A, between 94.07% and 99.15% in text B1_B, between 93.62% and 99.29% in text B2_A and between 96.65% and 99.33% in text B2_B. Interestingly, this last text seems to have been perceived as slightly easier in its vocabulary than the others, as all of its readers were above the threshold of 95% of vocabulary coverage discussed in section 1.2.2.2. This could be due to the topic of the text (TV shows) being generally more accessible to a predominantly young pool of annotators than the others (climate change, expressing negative emotions and garden sharing). The impact of topic specificity on lexical difficulty was indeed established in section 1.5.1.2.

Difficulty	Text			
	B1_A	B1_B	B2_A	B2_B
% Easy	97.48	97.0	97.42	98.01
% Transparent	1.07	1.73	1.06	0.64
% Vague	0.80	1.02	0.84	0.80
% Opaque	0.64	0.25	0.68	0.55

Table 3.4: Average Proportion of Words of Each Difficulty Level in Each Text

3.1.3 Comprehension Questions

Multiple choice questions were used to assess the global comprehension of readers for each text. Figure 3.3 shows the distribution of the number of questions correctly answered by participants to those MCQs. Overall, all four texts were well understood, with the means ranging between 3.43 (text B2_A) and 4.17 correct answers out of 5 (text B1_B). The questions for both B2-level texts seem to have been slightly more challenging than those of the B1 texts, which confirms that the former were somewhat more difficult for the readers.

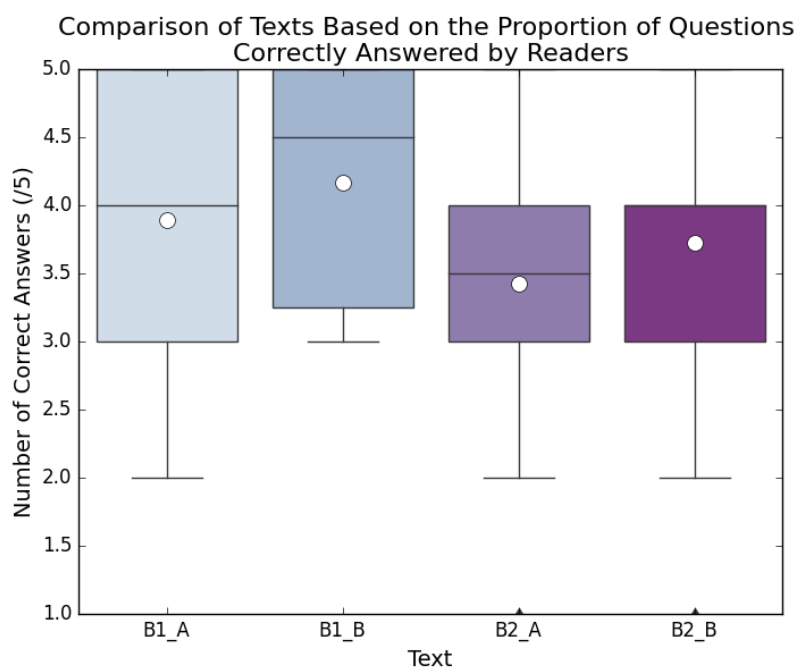


Figure 3.3: Number of Questions Correctly Answered for Each Text

Our third research hypothesis (RH 3) expected that participants who were familiar with a higher proportion of the words of a text would have answered the questions related to that text more accurately. Figure 3.4 plots the relationship between the proportion of words annotated with different values (with 1 being “transparent”, 2 “vague” and 3 “opaque”) and the number of correct answers provided by a participant. Surprisingly, figure 3.4a shows a small positive correlation between the number of annotated words and the accuracy of provided answers. The correlation was not statistically significant according to a Spearman’s rank correlation test ($\rho = 0.17$, $p = 0.17$), which was used as the independent variable did not follow a normal distribution. Further analyses revealed that there was actually a small but significant positive correlation between the proportion of words in a text annotated as transparent by a reader and the quality of that reader’s answers to the comprehension questions ($\rho = 0.28$, $p = 0.02$). Figure 3.4d also hints at a possible negative correlation between the proportion of opaque words and the number of correct answers; however, that correlation was not found to be statistically significant ($\rho = -0.18$, $p = 0.13$). The results obtained for words annotated as vague were even more inconclusive ($\rho = 0.12$, $p = 0.33$).

The striking difference in slope between figures 3.4b and 3.4d shows that it would not make sense to group all three levels of annotations as in figure 3.4a. Transparent words seem to have a significant positive impact on comprehension, while opaque words cause the opposite effect, albeit not significantly. More data would be needed to draw fully-fledged conclusions on the effects of the different types of words on the ability of readers to answer questions accurately. Indeed, the analyses reported above were based on only 68 data points, and it is possible that the trend observed with opaque words would be found to be significant in a broader study. Nonetheless, the results already allow for some interesting interpretations to be made. The apparent positive impact of transparent words can be tied back to elements that were already mentioned in this dissertation: the facilitating effect of cognates on reading comprehension and the importance of making inferences as a

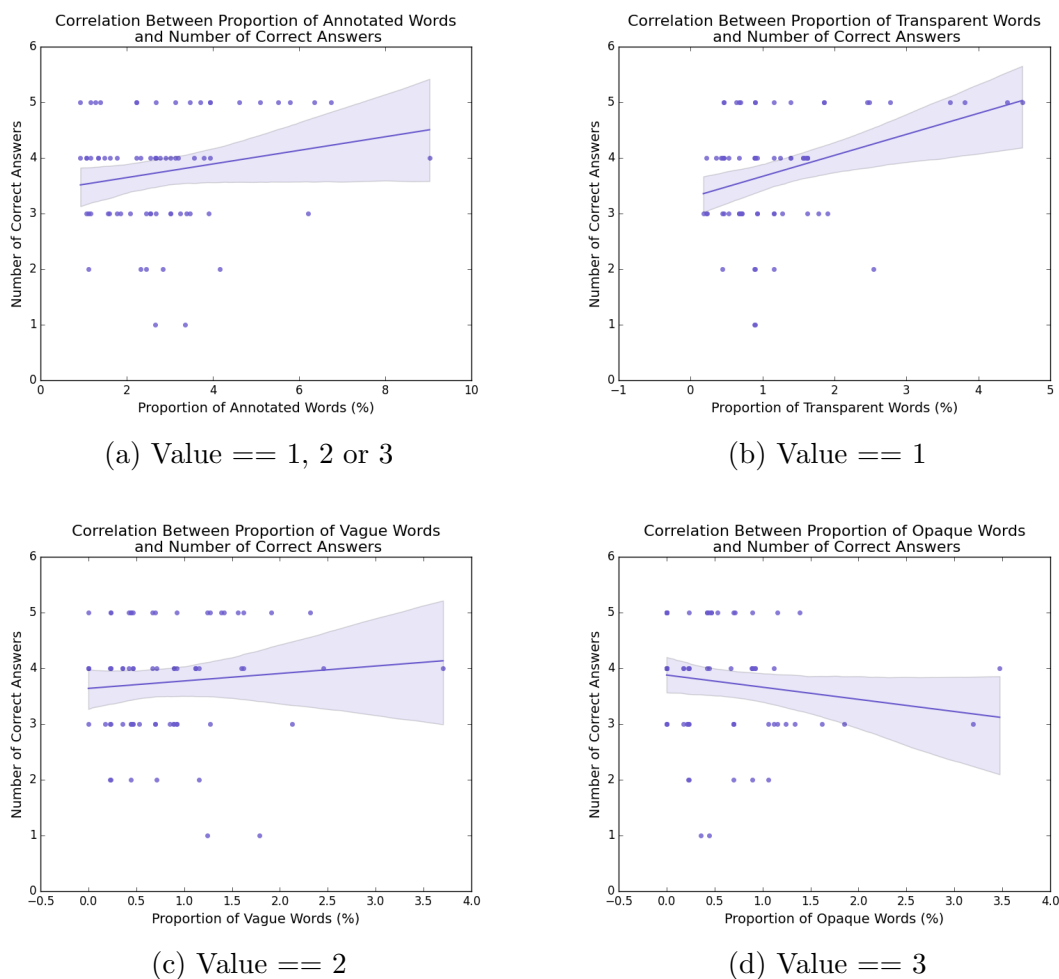


Figure 3.4: Correlation Between the Proportion of Words Annotated with Given Values and the Number of Correct Answers

reading strategy. The significant positive correlation could thus be due to the fact that people who are able to infer meaning in a text are more efficient readers, and as a result perform better on comprehension questions than those who do not make inferences as easily. Moreover, the change in slope between the three annotation levels can be seen to support the relevance of the scale used in this study: words go from having a facilitating effect on comprehension (transparent) to hindering it (opaque), which indicates that the scale does capture increasing degrees of difficulty. This highlights once more the importance of using a multi-level annotation scheme: simple binary labels would be likely to group together words that have very different impacts on the readers.

The short answer to our third research question, which was “Is there a direct

link between the proportion of words perceived as difficult by a given reader and the quality of their text comprehension?”, is that it depends on what we mean by “difficult”. Our hypothesis was that it would encompass all levels of the scale except level 0, *i.e.* all words with which a participant was not already familiar prior to reading, and that such words would hinder comprehension. However, that was found to be inaccurate: there was no significant link between the proportion of annotated words and the number of correct answers, and the observed trend was positive rather than negative. The threshold of 95% of vocabulary coverage did not seem to have an impact on comprehension under this definition. If, by contrast, we consider that difficult words are those which were labelled as “vague” or “opaque” exclusively, different conclusions might be drawn (although more data would be required to be certain). Indeed, only 2 participants annotated more than 5% of the words as difficult (level 2 or 3), and they answered 3 and 4 questions correctly respectively. Whether this means that being under the threshold had an impact on their comprehension is unclear. Although the negative slope in figure 3.4d does indicate that there was a tendency for readers who perceived more words as very difficult to provide less accurate answers, the correlation was not found to be statistically significant.

3.1.4 Discussion

To sum up, this study was answered by two pools of participants, one very homogeneous and the second much more diverse. Exploring how that impacts inter-rater agreement scores should therefore prove interesting. Two external measures which are assumed to correlate with difficulty (time spent on the task and number of comprehension questions answered correctly) showed that B2-level texts were indeed more difficult than B1-level texts, although that was not necessarily apparent at the lexical level. It might therefore be that the difficulty of those texts resided somewhere else, such as on the syntactic or grammatical levels. The impact of the topic on lexical difficulty perception was cited as a possible cause explaining why one of

the two B2 texts had a slightly higher proportion of easy words than the others. Using the ID of the texts as a predictor in a model could thus be helpful, as it would be a way to account for the differences in topic. The preliminary analyses of our data also revealed that transparent words were positively correlated with quality of comprehension, which we interpreted as a result of participants who were able to make inferences being more competent readers. Although analyses of the effect of difficult words on comprehension were inconclusive, it did appear that opaque words had a negative impact. As for vague words, the (non-significant) correlation with the number of correct answers seemed to tend towards being positive, which could be interesting to explore further in a broader study. Indeed, it is possible that participants were still able to guess the meaning of the words which they annotated as vague sufficiently to be able to understand the general gist of the sentence in which they appeared, while opaque words made that more challenging.

3.2 Agreement Scores

The review of previous studies outlined in Chapter 1 showed that readers whose profile was similar tended to agree more in their difficulty judgements than those who had different backgrounds. Factors which were found to increase agreement rates were the annotators' L1 (including when mother tongues were grouped into language families), their educational background, and possibly their proficiency. Agreement scores computed for heterogeneous groups using Krippendorff's α were around 0.25 (0.244 for CWI-2016 [66], 0.26 in Tack's thesis [89]). The more homogeneous groups of annotators analysed in the latter study, by contrast, obtained agreement rates ranging from 0.36 to 0.51.

Our first research hypothesis (RH 1) predicted that the results gathered in this study would follow the same trend. As mentioned above, the study was answered by two pools of annotators, with the first one being significantly more homogeneous than the second. It should therefore obtain higher inter-rater agreement scores, as

it was made up of people who are expected to have similar lexical needs. Table 3.5 compares the Krippendorff’s α values computed across the different texts in both versions of the study. The results confirm our hypothesis: the first group consistently achieves much higher agreement scores than the second, with their scores being over twice as high as those of the second group for texts B1_A and B2_A. It is interesting to note that the α values are in the same range as those reported in previous studies, which shows that there might be a constant trend in how much a group will agree in its perception of lexical difficulty, depending on whether or not it is homogeneous.

Study	Texts			
	B1_A	B1_B	B2_A	B2_B
Pilot	0.45 (12)	/ (0)	0.50 (5)	0.57 (9)
Main	0.22 (16)	0.32 (6)	0.23 (9)	0.35 (11)

Table 3.5: Inter-Rater Agreement per Text for Both Versions of the Study. The number of annotators is indicated between brackets.

These results seem to indicate that the answer to our first research question (RQ 1) is “yes”: there is more between-group variability than there is within-group. However, they do not yet confirm our second hypothesis (RH 2), which states that a reader’s L1 is one of the most central predictors of lexical difficulty. Indeed, the participants who made up the first group shared many characteristics besides their mother tongue (age, country of origin, educational background and proficiency level); it is therefore not yet clear which of those factors had an impact on the relative similarity of their annotations. To investigate the question further, we attempted to split the diverse pool of annotators of the second version of the study into more homogeneous language family sub-groups. The assumption was that this would result in higher scores, indicating that mother tongue indeed has a strong impact on the perception of lexical difficulty when reading.

Table 3.6 displays the results of separating the pool of annotators into L1 sub-groups. Agreement scores were only computed when there were more than 2 par-

Group	Sub-Group	Texts			
		B1_A	B1_B	B2_A	B2_B
All	All	0.22 (16)	0.32 (6)	0.23 (9)	0.35 (11)
Indo-european	All	0.31 (9)	0.32 (3)	0.12 (4)	0.36 (8)
	Germanic	0.34 (6)	/ (1)	0.19 (3)	0.44 (4)
Other	All	0.30 (7)	0.31 (3)	0.28 (5)	0.56 (3)
	Sino-Tibetan	0.24 (5)	/ (2)	/(2)	0.56 (3)

Table 3.6: Inter-Rater Agreement Scores Between L1 Sub-Groups of the Main Study. The values in bold are those that are higher or equal to the ones obtained by the entire group of participants.

ticipants for whom to compare answers. For text B1_A, splitting the group into speakers of an Indo-European language and speakers of a non-Indo-European language improves the agreement scores to 0.31 and 0.30 respectively. Removing the three speakers whose L1 is not a Germanic language from the first sub-group further raises the agreement score to 0.34. The same trend can be seen with text B2_B: isolating the three Sino-Tibetan speakers from the pool of annotators slightly improves the result for the remaining group, and yields a much higher score for the Chinese and Cantonese speakers. Once again, the Germanic sub-group obtains better results than the group of Indo-European natives. While this is also the case for text B2_A, both sub-groups still remain below the agreement score computed for the entire pool of participants. The group of non-Indo-European speakers, by contrast, achieves better results. As for text B1_B, splitting its annotators into 3 Indo-European and 3 non-Indo-European speakers has little impact on the score, which stays the same for the former and ever so slightly decreases for the latter.

All in all, these results seem to confirm the impact of a reader’s L1 on their reading comprehension in French. Splitting the pool of annotators into very broad language families (here, “Indo-European” and “Other”) tended to increase the scores slightly (5 times out of 8), while grouping participants into more fine-grained lan-

guage families (“Germanic” and “Sino-Tibetan”, in this case) outperformed those very general families in all cases but one. It is interesting to note that text B2_B consistently obtained the highest agreement scores for all groups, which can be tied back to the smaller amount of variation in the proportion of annotated words highlighted in section 3.1.2. Once again, this can perhaps be explained by the topic of the text being more universally accessible than the other three.

Using Krippendorff’s α to compute agreement scores seemed the best course of action in this study, since we had ordinal data and a limited number of annotators. Indeed, the metric is less sensitive to variations in the number of coders and coding categories than other similar measures of inter-rater agreement [2]. Nonetheless, an additional test was carried out using Fleiss’ Kappa, in order to obtain category-wise agreement scores for the entire pool of participants. The results of that test are presented in table 3.7. They show that across all texts, the difficulty label on which participants agreed the most was 0 (“Easy”); this is not surprising, as all readers annotated less than 10% of the words in the texts that they read. Interestingly, the second highest values in each column are found at the last difficulty level (“Opaque”). It therefore seems that the words on which annotators agreed the most are those that can be said to be either very simple or very complex; by contrast, words which fall towards the middle of the complexity spectrum received a wider variety of annotations. However, a correlation test using Spearman’s ρ found a very strong positive correlation between the mean difficulty value of a word and its standard deviation ($\rho = 0.9997$, $p = 0.0$), which is in line with findings reported by Shardlow *et al.* [85]. The regression between the two variables is plotted in figure 3.5.

These seemingly contradictory results are most likely a function of the design of the methods employed. Indeed, Fleiss’ Kappa does not use any weighting for ordinal data: the categories are considered as being unordered, and there is therefore no penalty for wider differences in ranks. To illustrate, let’s take the example of a

Rank	Texts			
	B1_A	B1_B	B2_A	B2_B
0	0.293	0.314	0.311	0.453
1	0.127	0.149	0.095	0.120
2	0.078	0.101	0.116	0.154
3	0.156	0.169	0.154	0.256

Table 3.7: Category-Wise Fleiss’ Kappa Scores for Each Text

population of 10 annotators. In one case, half of them annotate a word as “easy” (score of 0), and half as “opaque” (score of 3). The mean value of that word is thus 1.5, and so is its standard deviation. In a second case, half of the participants still annotate the word as easy, but the other half marks it as “transparent” (score of 1). This time, the word obtains a mean value and a standard deviation of 0.5. Fleiss’ Kappa, however, would yield the same agreement score in both scenarios, as the difference between labels 0 and 1 is not seen as smaller than the one between 0 and 3. Hence, while the results show that participants agreed slightly more on which words to annotate as 3 than as 1 or 2, their standard deviation is likely to be higher simply as a result of 0 being the most common annotation. Indeed, there was not a single word in our corpus that did not receive at least one label of 0. Only two words were annotated as easy by a single participant, namely “râler” in the title of text B1_B, and “comble” in text B2_B. Those words received a mean annotation value of 1.5 and 2 respectively. As a result, all words perceived as difficult by one reader were simultaneously marked as easy by at least one other person, which leads to high standard deviation values.

To sum up, inter-rater agreement scores showed that participants whose profile was more similar agreed more in their perception of lexical difficulty. Splitting a heterogeneous group into L1 family sub-groups generally resulted in an increase in agreement score, which indicates that the L1 of a reader is indeed an important factor for difficulty perception. The label on which all participants agreed the most was

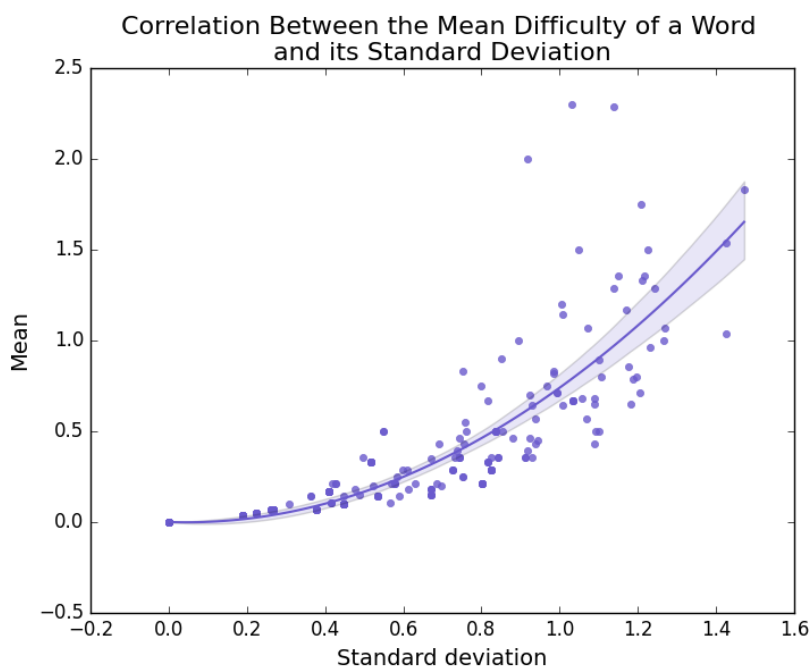


Figure 3.5: Polynomial Regression of the Standard Deviation of a Word by its Average Value

“easy”, followed by “opaque”. “Transparent” and “vague” were thus the difficulty levels for which the least amount of agreement between annotators was observed. A Spearman rank correlation test found a very significant positive correlation between the mean annotation value of a word and its standard deviation, which was interpreted as being due to all words being annotated as easy by at least one reader.

The results presented in this section confirm our first research hypothesis (RH 1; withing-group variability is lower than between groups) as well as part of the second (RH 2; the L1 of a reader has an impact on their perception of lexical difficulty). It remains to be seen whether the other variables mentioned in that second hypothesis do in fact have an effect on lexical difficulty. This is what the next section will try to uncover.

3.3 Ordinal Regression Models

As mentioned in section 2.3, three types of models were considered in order to learn our data. Two of them were cumulative link models (CLMs), one allowing mixed

effects and the other using fixed effects only. The difference between fixed factors and random factors is that the former are assumed to contain all levels of interest in a given study, whereas the latter are viewed as having been randomly sampled from a wider population of possible levels [97]. The third type of model was a multinomial model, which treats the response variable as nominal rather than ordinal. It was included in case our data did not meet the assumption of proportional odds, which assumes that the coefficient of each predictor will remain constant across the levels of the dependent variable [33]. If the four levels on our difficulty scale do not behave similarly with regard to the predictors included in the models, an MLM should be a better fit, as it will consider each level separately.

3.3.1 Baseline Models

In order to compare the quality of future models, a baseline was created for both model types (cumulative link and multinomial). They attempted to predict the value given to a word by an annotator based on the average value obtained by that word. In other words, the baselines predicted difficulty on the basis of complexity alone.

Table 3.8 reports the statistics of both baseline models, with the better result for each metric marked in bold. Interestingly, the multinomial log-linear model performed better than the cumulative link one in terms of AIC and residual deviance, but the cumulative link baseline achieved a lower percentage of classification error. The high residual deviance scores show that the average difficulty value of a word alone is not enough to predict individual annotations, which is not surprising given the low rate of agreement between the whole group of annotators. The classification error metric was computed on the test set and seems to indicate that the models performed well, as over 97% of instances were accurately predicted by both. However, this is likely due in part to the strong imbalance in values in our data: 6258 observations out of 6417 in the test set were labelled as “easy”. As a result, a naive classifier that marks all words as 0 would achieve a classification error of only 2.478,

which is only slightly worse than the score reported for the CLM baseline, and better than the one for the multinomial model.

Metrics	Models	
	CLM Baseline	MLM Baseline
AIC	4759.26	4651.83
Residual Deviance	4751.26	4639.83
Classification Error (%)	2.431	2.493

Table 3.8: Performance of the Baseline Models

3.3.2 Learner Models

Another interesting comparison to draw would be between models making fully-personalised predictions of difficulty and those predicting the needs of a reader based on annotations from people with a similar profile. The first type would use the ID of annotators as a unique individual predictor, while the second would look at characteristics such as age or L1. This section will present the results of three such learner models.

The variables considered for those models were the average difficulty value of a word, the ID of annotators, and all word and text-related features. This last set contained some variables which could not be included at the same time, because they were strongly correlated or because one variable was fully predicted by the others. As a result, only one of the different frequency measures could be used at a time, as well as one of either the text ID or text level variables. The optimal variable in those mutually exclusive sets was selected by iteratively adding them to a base model and keeping the feature which lowered the AIC score of that model the most.

The statistics of the models obtained after selecting all features that had a positive impact on their score are presented in table 3.9. Once again, the multinomial model performed best in terms of AIC score, and this time also made the most

Metrics	Models			
	CLMM	CLM	MLM	MLM Baseline
AIC	4197.74	4130.43	3985.14	4651.83
Residual Deviance	/	4006.43	3625.14	4639.83
Classification Error (%)	/	2.415	2.369	2.493

Table 3.9: Performance of the Learner Models

accurate predictions on the test set. It used the following predictors: mean difficulty value, annotator ID, text level, part of speech and word frequency in films according to Lexique3. The word index and text ID features were not informative for the model, whereas film frequency performed better than all other frequency metrics. The CLM model made use of the exact same predictors, with the only difference being that text ID did improve the AIC score, but still not as much as the text level feature. Both numerical variables and at least one level of all categorical variables were found to be statistically significant ($p < 0.05$). As for the CLMM model, only the AIC metric could be computed with the ordinal R package. It obtained the worst score out of the three models (while still outperforming the baseline), which indicates that mixed models might not be the best fit for our data. The model used the ID of annotators and the part of speech of target words as random effects, and the mean difficulty value, text level and film frequency as fixed effects. Using text level as a fixed effect resulted in a better AIC score than using text ID either as a fixed or as a random effect. All three fixed effects were once more found to be statistically significant ($p < 0.05$).

Three predictors could not be tested with those model: number of homographs, homophones and syllables. Indeed, as not all words appeared in Lexique3, some observations ended up with missing data for those features. Since this is an issue for numerical variables, some data points were dropped when fitting the models. The same problem did not arise for frequency measures, which attributed a score of 0 to

those absent words, nor for the part of speech predictor, as the empty response was treated as an additional level of the nominal variable. A second model of each type was therefore created in order to investigate the usefulness of those extra predictors. For the mixed model, all three of them were found to be useful; however, only the number of homographs and number of syllables were significant predictors of difficulty ($p_{n_syllables} < 0.001$, $p_{n_homographs} < 0.001$, $p_{n_homophones} = 0.08$). Including all three variables in the CLM model similarly resulted in the lowest AIC score, despite the contribution of the number of homophones once again not being statistically significant ($p_{n_syllables} = 0.006$, $p_{n_homographs} = 0.005$, $p_{n_homophones} = 0.09$). Finally, only the number of syllables was found to improve the AIC score for the MLM model.

All in all, the results across models were very consistent, which indicates that some of our features are reliable predictors of difficulty. Those include the ID of annotators, the mean annotation received by a word, its part of speech, the level of the text in which it appears (B1 or B2), its frequency in films, and its length expressed in syllables. The last two predictors are not surprising, as they were shown many times to be useful in the literature. The text level feature being significant was more unexpected, as section 3.1.2 found that B2-level texts did not differ much in their proportion of difficult words. Other variables which sometimes improved the performance of our models were text ID, number of homographs and number of homophones. By contrast, including word index in a model never resulted in a better AIC score.

The fact that the multinomial log-linear model performed best is also an interesting finding. It indicates that it might be preferable to treat our response variable as unordered, *i. e.* fit a different slope for each difficulty level on the annotation scale rather than assume parallel regression across levels.

3.3.3 Individual Differences Models

The models described in this section make use of profile information from the annotators as a way to predict their perception of difficulty. As such, all features of the participants were considered, except for their ID. The aim of exploring such models was to determine whether the needs of a new reader can be predicted based on their demographic information, without needing them to provide annotation data.

Once more, some variables could not be included at the same time, or the models would become rank deficient (*i.e.* one of the predictors could be explained by one or a combination of the others, or some ranks of a categorical variable would not be represented in the data). This was the case for the L1 families and sub-families, the former being dependent on the latter, as well as for the country variable, which could not be used alongside information on the L1 or education level of participants. As a result, different models had to be built and compared in order to uncover the optimal combination of predictors. Moreover, given that some demographic information was missing for participants in the pilot study, only a subset of individual predictors could be used in models trained on the whole dataset. Those models will be presented in the following sub-section, whereas others trained only on data collected in the main study will be described in section 3.3.3.2.

3.3.3.1 Full Dataset

The models fitted on the full training set had access to the following predictors: age, country, education, proficiency level in French, L1 family, L1 sub family, and number of correct answers. Just like before, the performance of the three types of models is presented in table 3.10, with the best of the two baselines included for comparison. All three models outperformed the baseline, and the multinomial model obtained the most favourable scores. Interestingly, the cumulative link model achieved the lowest error rate, like it already had in section 3.3.1.

All three models used the country predictor, which was found to improve the

Metrics	Models			
	CLMM	CLM	MLM	MLM Baseline
AIC	4228.59	4205.54	4085.36	4651.83
Residual Deviance	/	4121.54	3839.36	4639.83
Classification Error (%)	/	2.369	2.400	2.493

Table 3.10: Performance of the Individual Differences Models

AIC score more than the variables it excluded (L1 family and education). Those two variables were nonetheless shown to be statistically significant in models that did include them, with the L1 sub-family feature being more informative than the L1 family. A possible explanation for why the country predictor performed better is that it encoded L1 information as well as some other cultural factors that were not measured in our demographic questionnaire. To some extent, it might have reflected the “social, educational, political and pedagogical contexts” mentioned in section 1.3.2, which are said to impact the language acquisition process. A second, more practical reason for that performance could be the number of countries which were represented by only one reader (11, for a total of 39 participants). The predictor could therefore have acted as a proxy for the annotator ID information for about a fourth of our participants. A further look into which levels of the variable were useful showed that out of the four countries which counted 2 or more annotators (besides China, which was the level used for the intercept), 2 had a significant effect on difficulty ($p_{Russia} = 0.03$, $p_{US} < 0.001$). Moreover, 2 countries represented by a single participant did not significantly impact the predictions ($p_{Philippines} = 0.9$, $p_{SouthAfrica} = 0.3$). It is therefore possible that both explanations provided account for some part of the significance of the country predictor.

The mixed model presented in the table above used part of speech as a random effect, and the following predictors as fixed effects: text ID, word average difficulty and film frequency, annotator’s age, country and number of correct answers. The

cumulative link model was built with a very similar list of predictors: average difficulty, film frequency, part of speech, text ID, age and country. Including the number of correct answers had a slight negative effect on the model’s performance, while the part of speech predictor markedly improved its AIC score. This time, text ID was more informative than text level, despite the opposite being true for learner models. It could be that the complexity of a text has more of an impact on an individual level, whereas its topic matters more for a wider population. This trend was also observed for the multinomial model, which used the same list of predictors as the mixed model. Two features, namely the position of a word and the proficiency level of an annotator, did not contribute to the predictions of difficulty. This second variable not being useful is somewhat surprising, but is most likely due to the fact that only 3 participants in the study had a level of B2, with all the others belonging to the B1 level.

For the same reasons as outlined in the previous section, three extra predictors had to be tested on a smaller version of the dataset to avoid having missing data in numerical variables. The mixed model found all three of them (number of homographs, homophones and syllables) to be useful predictors of difficulty, and two were statistically very significant ($p_{n_syllables} = 0.001$, $p_{n_homographs} < 0.001$, $p_{n_homophones} = 0.07$). Similarly, the simple cumulative model was improved by the addition of all three, but the number of homophones was not significant ($p_{n_syllables} = 0.008$, $p_{n_homographs} = 0.005$, $p_{n_homophones} = 0.08$). The multinomial model, by contrast, was only improved when adding the number of syllables. These results are very consistent with what was reported for the same model types in the previous section, despite the models using different predictors.

3.3.3.2 Partial Dataset

To investigate the usefulness of demographic information which was not collected in the pilot study, more models were trained on a smaller version of the dataset, which contained 19748 observations in total. The same methodology as before was

employed to split the data into a training and a test set. Two baselines using only the mean difficulty predictor were created; the results of the MLM version are reported in table 3.11, as it performed better than the CLM baseline.

The additional predictors which could be used on this version of the dataset were the time spent learning French in a native and a non-native context, whether an annotator had dyslexia, and their proficiency level in each language spoken by at least two participants in the study (namely English, Spanish, Russian, Chinese and German), ranging from no knowledge at all to mother tongue. This time, we did not check for the impact of the number of homographs, homophones and syllables, as the consistent results in the previous sections indicate that they would most likely behave similarly. Once more, some predictors could not be used at the same time. In addition to the ones already mentioned above, the country variable was mutually exclusive with any language information and with the dyslexia predictor. The L1 sub-family variable could not be added to a model already including one of the language knowledge features, as they were too closely linked; however, the same problem did not arise with the broader L1 families. Lastly, knowledge of German could not be used alongside the rest of the language predictors, and was therefore dropped as it was less informative than the others.

Metrics	Models			
	CLMM	CLM	MLM	MLM Baseline
AIC	2428.35	2439.17	2334.26	2891.88
Residual Deviance	/	2395.17	2220.26	2879.88
Classification Error (%)	/	2.468	2.443	2.595

Table 3.11: Performance of the Individual Differences Models Trained on the Second Version of the Study

The results presented in table 3.11 show that, once more, the multinomial model outperformed the cumulative ones. This time, however, the mixed model achieved a lower AIC score than the non-mixed version. It was built with part of

speech as a random effect, and a large set of fixed effects: mean difficulty value, film frequency, text ID, age of the annotator, education level, knowledge of English, Chinese, Russian and Spanish, broad L1 family, time spent learning French both in a native and in a non-native context, proficiency level in French and dyslexia. The text level variable contributed less to the model than text ID. The predictors which did not improve the AIC score of the model were the index of the word, the number of correct answers provided by the participant, and the number of languages they had some knowledge of in addition to French. The information encoded by this last predictor is to a great extent covered by the language knowledge features included in the model, which might explain why it was not useful. The optimal list of predictors for the cumulative link model was the same as the one for the mixed model, with the exception of part of speech, text ID and proficiency in French, which did not improve the quality of the model. Interestingly, replacing the language knowledge information with the country predictor made those three variables as well as the number of languages known useful and statistically significant for the new model; nonetheless, the first version achieved a lower AIC score. As for the multinomial model, it used a different set of predictors than the other two: average difficulty value, film frequency of the word, age of the annotator, education level, number of languages spoken, proficiency level in English, Spanish, Russian and Chinese, dyslexia, and time spent learning French in a non-native context. Surprisingly, adding the part of speech variable had a strong negative impact on the performance of the model. Similarly, word index, L1 family, proficiency level in French, number of correct answers and time spent learning French in a native context did not improve the quality of the model. As has repeatedly been the case with the individual differences models, the text ID variable performed better than the text level feature.

To sum up, most predictors have been found to contribute to a model at least once, with the exception of the position of a word in a text. The text level variable was useful, but not as much as text ID for the individual differences models, and as such was often discarded in favour of the latter. The most consistent predictors

across our models were the mean difficulty value, film frequency and age of the annotator. Part of speech and text ID also performed well, followed by the bulk of our variables. Features which were less conclusive were the number of languages spoken by an annotator and the broad L1 families, generally outperformed by the more specific equivalents.

3.3.4 Discussion

The main objective of this section was to explore the individual variables which can help predict lexical difficulty. Our second research hypothesis (RH 2) expected these predictors to be the L1 and other languages spoken by a reader, their proficiency level in French, as well as their age, educational background and time spent learning French either in a native or in a non-native context. L1 information did indeed perform well, with more specific sub-families outperforming broad families whenever they could be included in a model. However, using the country variable generally improved the model more than using the L1 family features. This could be due to the predictor genuinely being more informative, but also to the fact that several countries were represented by a single annotator, which might have helped the model to approximate fully-personalised predictions. The language knowledge feature for English, Chinese, Russian and Spanish had a strong positive impact on the models, to the point that they outperformed the country variable. This goes to show that including foreign language knowledge information alongside the L1 helps improve the accuracy of a lexical difficulty model. The age, education level, time spent learning French in a foreign language context and dyslexia predictors also had a significant impact on difficulty annotations. Some variables were less conclusive, such as the proficiency level in French, number of languages spoken, time spent learning French in a native context, and number of questions answered correctly. A further study based on a wider dataset would be needed to more extensively explore the impact of those predictors, as well as to allow for some variables which were mutually exclusive here to be used together in a model.

This exploration yielded several other interesting findings. Training a mixed effects model on our dataset did not often prove beneficial compared to a simple CLM, except when fitting the individual differences model using the entire set of demographic variables available. The multinomial models performed best across the board, which shows that the proportional odds assumption was not met by our data. Out of all frequency metrics considered, film frequency consistently improved AIC scores the most. This could be due to all participants being enrolled in a French course in a native context (France or French-speaking Belgium): the language to which they are exposed on a daily basis is probably more similar to the one found in films than to written language found in books or textbooks. The results could therefore be different with a pool of annotators having exclusively studied French in a foreign language context. Additionally, the text level feature performed best in learner models, while text ID was a better predictor for models based on individual characteristics. It is possible that the behaviour of a given annotator will change more when going from B1 to B2-level texts, whereas topics will have a stronger impact on a heterogeneous group of readers. Lastly, other word and text predictors which were consistently useful were part of speech and length of a word in syllables. The number of homographs and, to a lesser extent, homophones was useful for CLMs, but not for multinomial models.

Learner models achieved comparatively better results than those modelling individual features. This confirms what was mentioned in section 1.5.2: while fully-personalised models are the optimal choice in theory, individual differences models are nevertheless a good alternative in situations where readers cannot be asked to provide annotation data. Combining a lexical complexity measure, here expressed as the average difficulty value attributed to a word, with a set of individual predictors should therefore be a viable way forward for NLP tasks requiring personalised difficulty predictions.

A non-negligible limitation of this study was the restricted size of the annotator

pool. Indeed, there were many opportunities for an individual variable to serve as a way for models to learn the behaviour of a particular learner rather than extract tendencies from shared characteristics. To which extent this has impacted our results is unclear, and repeating a similar study on a larger dataset would be necessary to confirm our findings.

Conclusion

The results presented in the previous chapter demonstrated that using individual characteristics as predictors is a promising alternative for NLP systems concerned with measuring or reducing lexical difficulty. As expected, individual differences models were not as powerful as fully-personalised systems; however, they consistently outperformed baselines predicting annotations on the basis of complexity alone. This implies that adding demographic variables to the models developed during the shared tasks should increase their performance even further. More accurate and flexible than reader-independent CWI systems, this middle-ground approach is also in a way less limiting than fully-personalised predictions. Indeed, individual differences models would not require the end user to provide difficulty annotations on which to train, but would estimate their areas of difficulty on the basis of their similarities with a pool of annotators. Moreover, users could define the set of individual variables representing their intended audience, whether it be an individual or a group, in order to obtain difficulty predictions that really fit their needs.

Using a custom-made online reading platform, we collected difficulty annotations from learners of French as a foreign language who were currently enrolled in a French course in a native context (France or Belgium). In total, 39 participants took part, representing twelve L1 backgrounds and fourteen countries. They provided some of their individual characteristics by answering a demographic questionnaire, then were asked to read two short informative texts and highlight any word with which they were not familiar according to a four-point annotation scale. A series

of five multiple choice questions were asked after each text, so as to get an idea of how well they had understood what they had read. Our methodology was informed by several points that were raised in the literature review, such as the importance of including multiword expressions, of collecting several annotations for each target word, of gathering demographic information from the annotators, or of accounting for the potential impact of the topic on a reader's perception of lexical difficulty.

The data collected through this study allowed us to offer some answers to our research questions. Computing inter-rater agreement scores for different groups of annotators confirmed that readers whose profile was similar tended to agree more in their difficulty judgements compared to others. Simply grouping participants by their L1 family generally led to an increase in agreement, which illustrates the interest of accounting for individual differences in difficulty predictions. Interestingly, the results obtained for the main two sub-groups of annotators were very consistent with what has been reported in the literature. The first group, whose members were of different ages, nationalities, education levels, mother tongues and proficiency levels, obtained agreement scores of about 0.25, whereas the second, much more homogeneous group of 18-year-old Malaysian undergraduates with a B1 level in French, were closer to 0.50. The similarity with previous studies indicates that those agreement scores might reflect some sort of general tendency.

To investigate the impact of different individual characteristics on difficulty perception, three types of predictive models were experimented with. Out of the three, multinomial log-linear models performed best. This implies that our data did not meet the assumption of parallel regression, and that it is preferable to fit a different slope for each level of our annotation scale. We used the mean difficulty value received by a word as a complexity metric, to stay close to the way that the CWI shared tasks defined their gold standard complexity labels. This, combined with a set of individual predictors as well as text and word features, was used to train our predictive models. All models, even the ones making the assumption of pro-

portional odds, succeeded in beating the complexity baseline, which underlines the limitations of a one-size-fits-all approach to Complex Word Identification. Among the variables that were the most useful and statistically significant for the models were L1 and L2 features, age, and education level. The time spent learning French in a foreign context was also consistently found to improve predictions, whereas time in a native context was not always useful. Whether an annotator had dyslexia also proved a good predictor of difficulty. The country variable performed equally well, but more data would be needed to determine whether that might be because it acted as a proxy for learner ID in certain cases. Some other variables that sometimes improved the performance of models were the number of languages spoken, which was most likely redundant when the other language features were included, the number of correct answers provided, and the proficiency level in French. This last variable was expected to perform better, in line with what previous studies had found; however, its moderate performance is probably due to only three participants having a level other than B1.

Our third research question related to whether a direct link could be established between the proportion of annotated words and the number of comprehension questions correctly answered. Our analyses did reveal a relationship between the two variables, albeit not the one that we expected: participants who had annotated more words as “transparent” tended to also answer the questions better. The correlation was statistically significant, and was interpreted as the sign that annotators who were able to make inferences about words that they did not previously know were more proficient readers. Although words annotated as “opaque” seemed to have the opposite effect on comprehension, the correlation was not statistically significant. More data would therefore be needed to further investigate the impact of the different annotation levels on the quality of comprehension.

This was a common conclusion of our analyses: repeating the study on a larger scale would be necessary in order for definite conclusions to be drawn. The size of

our corpus, as well as the restricted number of annotators, was indeed an important limitation of this study. One might for instance wonder whether the combination of certain variables allowed our models to approximate fully-personalised predictions of difficulty, which could have artificially inflated their significance. A second limitation was the restricted set of predictors used in the models, especially on the word and text level. Exploring the usefulness of features which performed well in the studies reported above, such as psycholinguistic measures or word embeddings, would undoubtedly have been interesting. Moreover, while our use of the mean difficulty value of a word as a reflection of its complexity allowed us to achieve our goal of identifying relevant individual predictors, it would not be sustainable in a more applied context. Future personalised models of difficulty would need to find the right combination of word-related predictors and individual differences features, since they would not have access to the mean difficulty value of a word when assigning a label to it.

Future avenues for research abound, as many paths were not explored in this dissertation. The effect of previous word exposure on comprehension, for instance, was not really investigated; neither was the question of whether MWEs were perceived as more difficult than single words. While we did account for mother tongue in our models, no information on whether a target word had a cognate in other languages was used. This could be an interesting predictor to experiment with in future work. Additionally, we remarked that, although B2-level texts appeared to be slightly more challenging for readers, this was not reflected in the annotations. It could therefore be that the complexity of those more advanced texts lied elsewhere; attempting to make personalised difficulty predictions on other levels than the lexis, such as that of syntax or the information structure, could therefore yield interesting results. Finally, individual differences take many more forms than a person's L1 and age. Finding a way to measure elements of the personality, motivation and attitude of a reader, for instance, should prove valuable for accurately anticipating their needs and presenting them with optimal reading materials.

Bibliography

- [1] D. Allen. “Cognate frequency and assessment of second language lexical knowledge”. In: *International Journal of Bilingualism* 23.5 (2019), pp. 1121–1136.
- [2] J.-Y. Antoine, J. Villaneau, and A. Lefevvre. “Weighted Krippendorff’s alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 550–559.
- [3] T. Bani Yaseen, Q. Ismail, S. Al-Omari, E. Al-Sobh, and M. Abdullah. “JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 661–666.
- [4] T. Bell. “Extensive Reading: Speed and Comprehension”. In: *The Reading Matrix* 1.1 (2001).
- [5] J. Bingel and J. Bjerva. “Cross-lingual complex word identification with multitask learning”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 166–174.

- [6] F. Boch. “Évaluer la compétence lexicale d’adultes francophones : selon quels critères ?” In: *Éducation & Didactique* 14-2 (2020), pp. 39–47.
- [7] B. Bulté and A. Housen. “Defining and Operationalising L2 Complexity”. In: *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Ed. by A. Housen, F. Kuiken, and I. Vedder. John Benjamins Publishing Company, Oct. 2012, pp. 21–46.
- [8] N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli. “Towards Best Practice for Multiword Expressions in Computational Lexicons”. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02)*. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002.
- [9] R. Carter. *Vocabulary: Applied Linguistic Perspectives*. Allen & Unwin, 1987.
- [10] C. Clark and K. Rumbold. *Reading for Pleasure: A Research Overview*. National Literacy Trust, 2006.
- [11] J. Coady. “L2 Vocabulary Acquisition through Extensive Reading”. In: *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*. Ed. by J. Coady and T. Huckin. Cambridge Applied Linguistics. Cambridge University Press, 1996, pp. 225–237.
- [12] J. Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.
- [13] K. Conklin and N. Schmitt. “Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?” In: *Applied Linguistics* 29.1 (Mar. 2008), pp. 72–89.
- [14] M. Constant. “Mettre les expressions multi-mots au coeur de l’analyse automatique de textes : sur l’exploitation de ressources symboliques externes”. PhD thesis. Université Paris-Est, Dec. 2012.
- [15] F. Cusin-Berche. “La notion d’‘unité lexicale’ en linguistique et son usage en lexicologie”. In: *Linx* 40 (1999).

- [16] D. De Hertog and A. Tack. “Deep Learning Architecture for Complex Word Identification”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 328–334.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.
- [18] J.-M. Dewaele. “Individual Differences in Second Language Acquisition”. In: *The New Handbook of Second Language Acquisition*. Vol. 2. Emerald Bingley, 2009, pp. 623–646.
- [19] Y. Ehara, Y. Miyao, H. Oiwa, I. Sato, and H. Nakagawa. “Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1374–1384.
- [20] Y. Ehara, I. Sato, H. Oiwa, and H. Nakagawa. “Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty”. In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 799–814.
- [21] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa. “Personalized Reading Support for Second-Language Web Documents by Collective Intelligence”. In: *Proceedings of the 15th International Conference on Intelligent User Interfaces*. Hong Kong, China, Feb. 2010, pp. 51–60.

- [22] R. Ellis. “Individual Differences in Second Language Learning”. In: *The Handbook of Applied Linguistics*. John Wiley & Sons, Ltd, 2004. Chap. 21, pp. 525–551.
- [23] B. Erman and B. Warren. “The Idiom Principle and the Open Choice Principle”. In: *Text* 20 (Jan. 2000), pp. 29–62.
- [24] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
- [25] T. François and C. Fairon. “Les apports du TAL à la lisibilité du français langue étrangère”. In: *Trait. Autom. des Langues* 54.1 (2013), pp. 171–202.
- [26] T. François, N. Gala, P. Watrin, and C. Fairon. “FLELex: a graded lexical resource for French foreign learners”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3766–3773.
- [27] N. Gala, T. François, D. Bernhard, and C. Fairon. “Un modèle pour prédire la complexité lexicale et graduer les mots”. In: *TALN’2014*. Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles. Marseille, France, July 2014, pp. 91–102.
- [28] N. Gala, T. François, and C. Fairon. “Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons”. In: *eLex - Electronic Lexicography*. Tallin, Estonia, Oct. 2013.
- [29] N. Gala, T. François, L. Javourey-Drevet, and J. C. Ziegler. “La simplification de textes, une aide à l’apprentissage de la lecture”. In: *Langue française* 3 (2018), pp. 123–131.
- [30] A. Garı Soler, M. Apidianaki, and A. Allauzen. “A comparative study of word embeddings and other features for lexical complexity detection in French”.

- In: *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*. Rennes, France: ATALA, May 2018, pp. 499–508.
- [31] S. Gooding and E. Kochmar. “CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 184–194.
 - [32] J. Grainger. “Visual Word Recognition in Bilinguals”. In: *The Bilingual Lexicon* (1993), pp. 11–25.
 - [33] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. 3rd ed. Wiley-Interscience Publication, 2013.
 - [34] M. Hu and I. S. P. Nation. “Unknown Vocabulary Density and Reading Comprehension”. In: *Reading in a Foreign Language* 13 (2000).
 - [35] T. Huckin and J. Coady. “Incidental Vocabulary Acquisition in a Second Language: A Review”. In: *Studies in Second Language Acquisition* 21 (1999), pp. 181–193.
 - [36] S. Jarvis. “Methodological Rigor in the Study of Transfer: Identifying L1 Influence in them Interlanguage Lexicon”. In: *Language Learning* 50.2 (2000), pp. 245–309.
 - [37] S. Jarvis and A. Pavlenko. *Crosslinguistic Influence in Language and Cognition* (1st ed.) Routledge, 2007.
 - [38] S. K. Jauhar and L. Specia. “UOW-SHEF: SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features”. In: **SEM 2012: 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 (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, July 2012, pp. 477–481.

- [39] M. D. Johnson. “Cognitive Task Complexity and L2 Written Syntactic Complexity, Accuracy, Lexical Complexity, and Fluency: A Research Synthesis and Meta-Analysis”. In: *Journal of Second Language Writing* 37 (2017), pp. 13–38.
- [40] T. Kajiwara and M. Komachi. “Complex Word Identification Based on Frequency in a Learner Corpus”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 195–199.
- [41] H. Kerkman and K. de Bot. “De organisatie van het tweetalige lexicon”. In: *Toegepaste Taalwetenschap in artikelen* 34 (1989), pp. 115–121.
- [42] M. Konkol. “UWB at SemEval-2016 Task 11: Exploring Features for Complex Word Identification”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1038–1041.
- [43] B. Lamiroy. “Le figement: à la recherche d’une définition”. In: *Zeitschrift für Französische Sprache und Literatur* 36 (2008), pp. 85–99.
- [44] B. Laufer. “What Percentage of Text-lexis is Essential for Comprehension?”. In: *Special Language: From Human Thinking to Thinking Machines* (1989), pp. 316–323.
- [45] B. Laufer and J. Hulstijn. “Incidental vocabulary acquisition in a second language: the construct of Task-Induced Involvement”. In: *Applied Linguistics* 22 (Mar. 2001), pp. 1–26.
- [46] B. Laufer and G. C. Ravenhorst-Kalovski. “Lexical Threshold Revisited: Lexical Text Coverage, Learners’ Vocabulary Size and Reading Comprehension”. In: *Reading in a Foreign Language* 22 (2010), pp. 15–30.

- [47] J. Lee and C. Y. Yeung. “Personalizing Lexical Simplification”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 224–232.
- [48] E. Léger-Zheng and O. Théophanous. “Inférence et apprentissage du vocabulaire: effet de la lexicalisation en première langue et de la tâche”. In: *Lexis. Journal in English Lexicology* 18 (2021).
- [49] B. Lété, L. Sprenger-Charolles, and P. Colé. “Manulex: A Grade-Level Lexical Database from French Elementary School Readers”. In: *Behavior Research Methods, Instruments, & Computers* 36 (Mar. 2004), pp. 156–166.
- [50] C. Lindqvist. “Vocabulary Knowledge in L3 French: A Study of Swedish Learners’ Vocabulary Depth”. In: *Languages* 6.26 (2021).
- [51] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv* (2019).
- [52] G. Lüdi. “Apprendre à construire des mots: la suffixation nominale chez des apprenants germanophones du français”. In: *Basel Electronic Working Papers in Linguistics* 1 (1997).
- [53] J. Lyons. *Introduction to Theoretical Linguistics*. Cambridge University Press, 1968.
- [54] M. Maddela and W. Xu. “A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Bruxelles, Belgique: Association for Computational Linguistics, Oct. 2018, pp. 3749–3760.
- [55] S. Malmasi, M. Dras, and M. Zampieri. “LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

- San Diego, California: Association for Computational Linguistics, June 2016, pp. 996–1000.
- [56] S. Malmasi and M. Zampieri. “MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 991–995.
- [57] A. Martinet. “Le syntème”. In: *La Linguistique* 35.2 (1999), pp. 11–16.
- [58] D. McCarthy and R. Navigli. “SemEval-2007 Task 10: English Lexical Substitution Task”. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 48–53.
- [59] A. McNeill. “Chapter 11. Native and non-native teachers’ sensitivity to language learning difficulties from a learner’s perspective: Implications and challenges for teacher education”. In: *Native and Non-Native Teachers in English Language Classrooms: Professional Challenges and Teacher Education*. Ed. by J. de Dios Martinez Agudo. De Gruyter Mouton, 2017, pp. 239–254.
- [60] J. Mesnager. “Pour une étude de la difficulté des textes. La lisibilité revisitée”. In: *Le français aujourd’hui* 137.2 (2002), pp. 29–40.
- [61] A. Mosquera. “Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 554–559.
- [62] I. S. P. Nation. *Learning Vocabulary in Another Language*. Cambridge Applied Linguistics. Cambridge University Press, 2001.
- [63] R. Navigli and S. P. Ponzetto. “BabelNet: Building a Very Large Multilingual Semantic Network”. In: *Proceedings of the 48th Annual Meeting of the*

- Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 216–225.
- [64] B. New, C. Pallier, L. Ferrand, and R. Matos. “Une base de données lexicales du français contemporain sur internet: LEXIQUE”. In: *L’année psychologique* 101.3 (2001), pp. 447–462.
 - [65] G. Paetzold and L. Specia. “Benchmarking Lexical Simplification Systems”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016.
 - [66] G. Paetzold and L. Specia. “SemEval 2016 Task 11: Complex Word Identification”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 560–569.
 - [67] G. Paetzold and L. Specia. “SV000gg at SemEval-2016 Task 11: Heavy Gauge Complex Word Identification with System Voting”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 969–974.
 - [68] A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, eds. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, Aug. 2021.
 - [69] C. Pan, B. Song, S. Wang, and Z. Luo. “DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 578–584.
 - [70] C. Pasquer. “Expressions polylexicales verbales : étude de la variabilité en corpus”. In: *TALN-RECITAL 2017*. Orléans, France, June 2017.

- [71] A. Polguère. “Non-compositionnalité : ce sont toujours les locutions faibles qui trinquent”. In: *Verbum (Presses Universitaires de Nancy)* XXXVII.2 (2015), pp. 257–280.
- [72] D. Qian. “Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective”. In: *Language Learning* 52 (2002), pp. 513–536.
- [73] D. Qian and M. Schedl. “Evaluation of an In-Depth Vocabulary Knowledge Measure for Assessing Reading Performance”. In: *Language Testing* 21.1 (2004), pp. 28–47.
- [74] G. Rao, M. Li, X. Hou, L. Jiang, Y. Mo, and J. Shen. “RG PA at SemEval-2021 Task 1: A Contextual Attention-based Model with RoBERTa for Lexical Complexity Prediction”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 623–626.
- [75] J. C. Richards. “The Role of Vocabulary Teaching”. In: *TESOL Quarterly* 10.1 (1976), pp. 77–89.
- [76] H. Ringbom. *The Role of the First Language in Foreign Language Learning*. Vol. 34. Multilingual Matters Limited, 1987.
- [77] F. Ronzano, A. Abura’ed, L. Espinosa-Anke, and H. Saggion. “TALN at SemEval-2016 Task 11: Modelling Complex Words by Contextual, Lexical and Semantic Features”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1011–1016.
- [78] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. “Multiword Expressions: A Pain in the Neck for NLP”. In: *Computational Linguistics and Intelligent Text Processing, Third International Conference*. 2002, pp. 1–15.

- [79] N. Schmitt. “Chapter 3. Key Issues in Teaching and Learning Vocabulary”. In: *Insights into Non-native Vocabulary Teaching and Learning*. Ed. by R. Chacón-Beltrán, C. Abello-Contesse, and M. del Mar Torreblanca-López. Multilingual Matters, 2010, pp. 28–40.
- [80] N. Schmitt. “Review article: Instructed second language vocabulary learning”. In: *Language Teaching Research* 12.3 (2008), pp. 329–363.
- [81] N. Schmitt. “Size and Depth of Vocabulary Knowledge: What the Research Shows”. In: *Language Learning* 64.4 (2014), pp. 913–951.
- [82] N. Schmitt, X. Jiang, and W. Grabe. “The Percentage of Words Known in a Text and Reading Comprehension”. In: *The Modern Language Journal* 95.1 (2011), pp. 26–43.
- [83] M. Shardlow. “A Comparison of Techniques to Automatically Identify Complex Words”. In: *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 103–109.
- [84] M. Shardlow, R. Evans, G. H. Paetzold, and M. Zampieri. “SemEval-2021 Task 1: Lexical Complexity Prediction”. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1–16.
- [85] M. Shardlow, R. Evans, and M. Zampieri. “Predicting Lexical Complexity in English Texts”. In: *CoRR* (2021).
- [86] A. Siyanova-Chanturia, K. Conklin, and N. Schmitt. “Adding More Fuel to the Fire: An Eye-Tracking Study of Idiom Processing by Native and Non-Native Speakers”. In: *Second Language Research* 27.2 (2011), pp. 251–272.
- [87] L. Specia, S. K. Jauhar, and R. Mihalcea. “SemEval-2012 Task 1: English Lexical Simplification”. In: **SEM 2012: 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 (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 347–355.
- [88] K. Sugiyama. “Analyse de la compétence lexicale dans la compréhension écrite des apprenants japonais en français”. In: *Revue japonaise de didactique du français, numéro spécial : Actes du IVe Congrès régional de la Commission Asie-Pacifique* (2017), 502 (12p.)
- [89] A. Tack. “Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers”. PhD thesis. 2021.
- [90] A. Tack, T. François, A.-L. Ligozat, and C. Fairon. “Modèles adaptatifs pour prédire automatiquement la compétence lexicale d’un apprenant de français langue étrangère”. In: *JEP-TALN-RECITAL 2016*. Paris, France, Jan. 2016.
- [91] K. Tanaka-Ishii and H. Terada. “Word Familiarity and Frequency”. In: *Studia Linguistica* 65 (Mar. 2011), pp. 96–116.
- [92] A. Teemant and S. E. Pinnegar. “Vocabulary Development and Language Transfer”. In: *Principles of Language Acquisition* (2019).
- [93] *TLFi : Trésor de la langue Française informatisé*. <http://atilf.atilf.fr/tlf.htm>, ATILF / CNRS – Université de Lorraine.
- [94] E. Van der Linden. “Lexique mental et apprentissage des mots”. In: *Revue française de linguistique appliquée* 11.1 (2006), pp. 33–44.
- [95] L. Verstraten. “Fixed Phrases in Monolingual Learners’ Dictionaries”. In: *Vocabulary and Applied Linguistics*. Ed. by P. J. L. Arnaud and H. Béjoint. London: Palgrave Macmillan UK, 1992, pp. 28–40.
- [96] M. B. Wesche and T. S. Paribakht. “Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth.” In: *Canadian Modern Language Review- Revue Canadienne Des Langues Vivantes* 53 (1996), pp. 13–40.
- [97] B. T. West, K. B. Welch, and A. T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. 2nd ed. Chapman and Hall/CRC, 2006.

- [98] R. S. Woodworth and D. G. Marquis. *Psychology (Psychology Revivals): A Study of Mental Life*. Psychology Press, 1963.
- [99] K. Wróbel. “PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 953–957.
- [100] S. M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. “A Report on the Complex Word Identification Shared Task 2018”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 66–78.
- [101] S. M. Yimam, S. Štajner, M. Riedl, and C. Biemann. “CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 401–407.
- [102] Q. Zeng, E. Kim, J. Crowell, and T. Tse. “A Text Corpora-Based Estimation of the Familiarity of Health Terminology”. In: *Proceedings of the 6th International Conference on Biological and Medical Data Analysis*. ISBMDA’05. Aveiro, Portugal: Springer-Verlag, Oct. 2005, pp. 184–192.