David Alfter

Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective

Data linguistica
<https://www.gu.se/svenska-spraket/data-linguistica>

Editor: Lars Borin

Språkbanken Text
Department of Swedish
University of Gothenburg

31 • 2021

David Alfter

Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective

Gothenburg 2021

Data linguistica 31
ISBN 978-91-87850-79-0
ISSN 0347-948X
E-publication <http://hdl.handle.net/2077/66861>

Printed in Sweden by
Stema Specialtryck AB 2021

Typeset in $\LaTeX\,2_\varepsilon$ by the author

Cover design by Kjell Edgren, Informat.se

Cover layout by Sven Lindström

Front cover illustration composed by the author; source material by
OpenClipart-Vectors and Gordon Johnson on Pixabay ©

Author photo on back cover by the author

ABSTRACT

In this thesis, we investigate how natural language processing (NLP) tools and techniques can be applied to vocabulary aimed at second language learners of Swedish in order to classify vocabulary items into different proficiency levels suitable for learners of different levels.

In the first part, we use feature-engineering to represent words as vectors and feed these vectors into machine learning algorithms in order to (1) learn CEFR labels from the input data and (2) predict the CEFR level of unseen words. Our experiments corroborate the finding that feature-based classification models using 'traditional' machine learning still outperform deep learning architectures in the task of deciding how complex a word is.

In the second part, we use crowdsourcing as a technique to generate ranked lists of multi-word expressions using both experts and non-experts (i.e. language learners). Our experiment shows that non-expert and expert rankings are highly correlated, suggesting that non-expert *intuition* can be seen as on-par with expert *knowledge*, at least in the chosen experimental configuration.

The main practical output of this research comes in two forms: prototypes and resources. We have implemented various prototype applications for (1) the automatic prediction of words based on the feature-engineering machine learning method, (2) language learning applications using graded word lists, and (3) an annotation tool for the manual annotation of expressions across a variety of linguistic factors.

SAMMANFATTNING

I den här avhandlingen undersöker vi hur språkteknologiska verktyg och tekniker kan appliceras på ordförrådet hos andraspråksinlärare av svenska genom att klassificera lexikala enheter utifrån inlärares olika färdighetsnivåer.

I avhandlingens första del undersöker vi olika språkliga särdrag och deras kombinationer vilka används för att representera ord som vektorer. Dessa vektorer matas in i maskininlärningsalgoritmer för att (1) identifiera färdighetsnivåer enligt CEFR-skalan utifrån indata och (2) predicera färdighetsnivåer hos okända ord. Våra experiment visar att när det gäller att avgöra ett ords komplexitet är särdragsbaserade klassifikationsmodeller som utgår från "traditionell" maskininlärning fortfarande överlägsna de nyare, allt populärare djupinlärningsmetoderna.

I andra delen använder vi crowdsourcing för att låta både experter (dvs språklärare) och icke-experter (dvs språkinlärare) rangordna flerordsuttryck. Våra experiment visar att experternas och icke-experternas rangordningar korrelerar starkt med varandra, vilket tyder på att icke-experternas intuition ligger i linje med experters kunskap, åtminstone med avseende på de variabler som har kunnat testas givet experimentets förutsättningar.

Den forskning som redovisas i denna avhandling har även genererat två typer av praktiskt tillämpbara resultat: prototyper – i form av datorapplikationer – och dataresurser. Vi har implementerat flera prototyper: (1) applikationer som automatiskt kan predicera ord med hjälp av särdragsbaserad maskininlärning, (2) språkinlärningsapplikationer som använder graderade ordlistor som informationskälla, och (3) en applikation för manuell annotering av språkliga uttryck utifrån en mängd lingvistiska faktorer.

ACKNOWLEDGEMENTS

There are many people I would like to thank, and while I cannot name every single one by name, I hope you know that I did not exclude anyone on purpose.

First and foremost, I would like to thank my supervisors Elena Volodina and Lars Borin for their unending patience, constructive feedback, helpful insights, but also for their support when I thought I couldn't pull through.

I would like to thank the discussion leader for my final seminar, Robert Östling, for providing very constructive feedback on the first draft of my thesis that helped in improving it.

I am grateful to Therese Lindström Tiedemann for providing feedback on the first draft on my thesis, although she didn't have to, for very interesting conversations and for the ongoing collaborations.

I would also like to thank everyone at Språkbanken Text for creating such a welcoming and fostering environment.

I would like to thank all the PhD students at the department of Swedish for creating a convivial, lively and (especially pre-2020 but even after that) very social environment. I would especially like to thank Anders Agebjörn, a fellow PhD student. We started at the same time and yet it seems I finish earlier. The journey certainly wouldn't have been the same without you. Thanks also for translating the abstract to Swedish.

I would like to thank Herbert Lange, who told me that he got a PhD position in Gothenburg during a computational linguistics' students' meeting (Tagung der Computerlinguistik Studenten; TaCoS) before I knew I would also end up in Gothenburg half a year later. Over the years, we had many interesting beer talks about everything and nothing, and I finally managed to read (his copy of) Gödel-Escher-Bach.

I would like to thank Sven Lindström for helping with the cover design and getting the thesis to print.

I would also like to thank all the wonderful people I had the privilege to meet throughout my studies. I especially thank the EuroCALL community for making me feel welcome right from the start. It was a pleasure to run into so many of you year after year, and it felt like we knew each other for so much longer.

I am also thankful to all the funding opportunities that made traveling across the world possible. I would especially like to thank Språkbanken Text, Kungliga Vitterhetsakademien, Filosofiska fakulteternas gemensamma donationsnämnd, Adlerbert Scholarships. Further, I am grateful for the different networks that enabled me to visit different host institutions during my studies, namely the ENel and enetCollect COST networks. I would also like to thank the L2 profiles project[1], funded by Riksbankens Jubileumsfond, grant P17-0716:1, in which most of my work was carried out.

Last, but certainly not least, I would like to thank my husband Stephan for having my back, carrying me when I was exhausted, pushing me when I was despairing, and helping me to see trees when all I saw was forest.

[1] https://spraakbanken.gu.se/en/projects/l2profiles

CONTENTS

Appendices

Part I

Introduction and overview

1 Introduction

1.1 Motivation

Vocabulary plays a major role in language learning (see for example Laufer and Nation 1999; O'Dell et al. 2000; Meara 2002; Gu 2003; Nation 2013), as is also expressed in the following quotes:

> "while without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (Wilkins 1972: pp. 111-112).

> "lexical knowledge is central to [...] the acquisition of a second language" (Schmitt 2000: p. 55).

With the past and ongoing advances in technology, new possibilities have been opened up that transcend the traditional classroom setting and printed textbooks. Computer technology has found its way into various areas such as language assessment and language teaching (Chapelle and Douglas 2006). There has been a rise in computer-assisted language learning (CALL) and intelligent computer-assisted language learning (ICALL) platforms, which additionally incorporate natural language processing (NLP), opening up a whole new field of opportunities. Traditional CALL platforms such as Moodle, a free open source content management system for educational content, tend to produce static and deterministic content, meaning that content authored through such tools is not changing once created. By using natural language processing, it is possible to enrich textual data for example by automatically identifying part-of-speech classes, syntactic relationships, named entities or compounds. Given this plethora of information, it is possible to devise exercises that are dynamically generated.

One of the practical outcomes of vocabulary research are word lists. Many vocabulary lists are created from native speaker material and thus might not reflect a learner's reality or needs (François et al. 2014: p. 3767). Basing vocabulary lists on language learner material ensures that the lists reflect the language that learners are confronted with. While vocabulary lists are often criticized as unnatural learning resource, they have their worth (Meara 1995).

Graded vocabulary lists are important resources in L2 contexts as evidenced by their use in language assessment tests (e.g. Coxhead 2011) or as a vocabulary learning strategy (e.g. LaBontee 2019). They are also used in text assessment such as in the recently released Duolingo CEFR checker[2], a tool that strongly resembles our prior release of Texteval[3]. Both tools are used to assess overall text complexity, but in addition also highlight words of different proficiency levels.

In this thesis, we look at how vocabulary can be characterized in terms of lexical complexity. We look at vocabulary from two perspectives: *receptive* knowledge and *productive* knowledge. Receptive knowledge concerns vocabulary that is readily understood by a language learner, although it does not necessarily mean that a learner would be able to actively use or produce said vocabulary. Receptive vocabulary is best exemplified by reading texts in textbooks or graded readers, i.e. books adapted to certain proficiency levels. In such texts the learner is expected to understand (most of) the text, even if understanding is reliant on illustrations or context. Productive knowledge concerns vocabulary that is actively used by a language learner. A source of productive vocabulary knowledge are for example learner essays. In the essays, one can see what vocabulary a learner can produce.

From a more theoretical point of view, there is an ongoing debate about what it means to *know* a word (e.g. Read 1988, 2004; Schmitt 2010).Words can be seen as multi-faceted entities, having multiple different aspects such as pronunciation(s), spelling(s), meaning(s) and different senses, synonyms, possible inherent constraints such as collocational patterns. Further, vocabulary knowledge can be seen as broad (i.e. having a diverse vocabulary) versus deep (i.e. having a better grasp of the different aspects of a word). Thus, does one *know* a word if one knows one of its possible senses? We acknowledge the existence of such theoretical debates but we will not dive into them in this thesis.

However, a general problem is polysemy (Parent 2009); word lists often tend to conflate different senses of words into a single entry. This is problematic, as not all senses of a word are learned at the same time (Crossley, Salsbury and McNamara 2010). In recent years, the focus has shifted from word-based to sense-based view of vocabulary, a tendency that is obvious not only in vocabulary lists (e.g. Tack et al. 2018) but also for example in word embeddings (Nieto Piña 2019). Thus, our initiative to work on sense-based graded word

[2] `https://cefr.duolingo.com`

[3] `https://spraakbanken.gu.se/larka/texteval`

lists indicates a timely development.

In this thesis, we focus on vocabulary as a construct in second language learning and specifically how NLP and other methods can be applied in ICALL contexts in order to improve the language learning experience.

1.2 Research questions

In the first part of the thesis, we look at single-word lexical complexity and ask the following questions:

(1) How can frequency distributions across proficiency levels be used to derive target levels?

(2) How can we assign target proficiency levels to words?[4]

(3) How can we assign target proficiency levels to unseen words?

(4) How can we check the validity of the assigned levels?

Concerning multi-word expressions, we ask the following questions:

(5) Does compositionality correlate with complexity?

(6) Can crowdsourcing techniques be used to create graded lists?

Research question 5 is only answered in the kappa; the experiment and result were planned to be included in publication 4 (see below) but were later removed.

1.3 Contributions

The main aim of this thesis is to investigate vocabulary from a second language learning perspective using natural language processing techniques. The research is based on two different types of corpora, a textbook corpus and a learner essay corpus. From each of these two corpora, words and their frequencies were extracted. In contrast to purely frequency based word lists, however, these lists also contain distributions of frequencies over different proficiency levels.

The first contribution concerns insights into how to best project these frequency distributions to single levels. In publication number 1 (see section

[4] In this context, *target proficiency level* is to be understood as the minimum proficiency level one has to have reached in order to be able to understand and/or produce a word.

1.4), we explore a threshold approach, not unsimilar to Hawkins and Filipović (2012), and find that this method produces more plausible target levels than other approaches, at least for learner-based data, which is also the case in Hawkins and Filipović (2012). In publication number 2, we also use textbook-based data and another, simpler projection technique based on the first occurrence of an expression; an expression is simply given the level of the text or essay it was first observed at. This is similar to Gala, François and Fairon (2013) who have also experimented with more involved projection techniques for textbook-based French data but have found the easy method to perform almost equally as well, if not better. This finding is further corroborated by our own findings that the majority of assigned levels are the same across word lists, regardless of the projection method used (see section 5.5.1).

The second contribution concerns the evaluation of the automatically assigned levels. We use methodological triangulation, i.e. various different methods such as comparison between projection techniques and custom semantic space embeddings (publication 1), 10-fold cross-validation (publication 2), crowdsourced data (publication 4) and multilingual aligned comparisons (publication 6). Unsurprisingly, there are outliers and other artifacts present in the data, the reasons being manifold, ranging from OCR and transcription errors to subjective and idiosyncratic language use. However, overall, we can see that the majority of assigned levels seem plausible as evidenced by positive results from different testing methods.

The third contribution concerns lexical complexity prediction. We use linguistic features to characterize words, and machine learning to learn target proficiency levels (publication 2). The output of this research is a system capable of predicting target proficiency levels for unseen words, i.e. words not present in the word list. We also adapt our pipeline to English for a shared task (publication 3). Overall, we find that our results are in line with lexical complexity prediction results for other languages (publications 2 and 3).

The fourth contribution concerns learner intuitions in comparison to teacher and assessor judgments as to the (perceived) difficulty of multi-word expressions (publication 4). Our results show a high degree of correlation between learner intuitions and teacher and assessor judgments, which indicates that internal language development and external language assessment are aligned. While this is a more indirect approach to linking expressions to levels, it opens up interesting new research questions. The study suggests that crowdsourcing can be (a first step as) an alternative to more direct level assignment; the study further suggests that language learners' contributions (in this experiment) can be seen as on par with expert knowledge.

The fifth contribution concern the conception and implementation of a lexicographic annotation tool that allows for rich manual annotation to complement automatic enrichment by linking different resources (publication 5).

Finally, to demonstrate the practical value of the current research, we use the obtained the word lists, techniques and algorithms in practical applications such as text evaluation and various exercise prototypes (publications 7 and 8). We surmise that data collected through such exercises may prove valuable for future research.

While publications 5, 7 and 8 are not directly connected to research questions, they are nonetheless of importance. In publication 5 we introduce a custom tool for manual annotation of vocabulary items. The aim of this tool is to create a new resource; resource creation (e.g. dictionary compilation) is often undervalued, yet such processes are necessary and enable further research. In publications 7 and 8 we show how the research output (i.e. graded vocabulary) can be deployed in practice. In addition, data collected through practical applications can be used for further research.

1.4 Overview of publications

The thesis contains the following published articles:

1. Alfter, David and Yuri Bizzoni and Anders Agebjörn and Elena Volodina and Ildikó Pilán 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016 (No. 130, pp. 1-7). Linköping University Electronic Press. [chapter 9]

2. Alfter, David and Elena Volodina 2018. Towards single word lexical complexity prediction. Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications (pp. 79-88). [chapter 10]

3. Alfter, David and Ildikó Pilán 2018. SB@GU at the Complex Word Identification 2018 Shared Task. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 315-321). [chapter 11]

4. Alfter, David and Therese Lindström Tiedemann and Elena Volodina In press. Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. Northern European Journal of Language Technology. [chapter 12]

5. Alfter, David and Therese Lindström Tiedemann and Elena Volodina 2019. LEGATO: A flexible lexicographic annotation tool. In NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland (No. 167, pp. 382-388). Linköping University Electronic Press. [chapter 13]

6. Graën, Johannes and David Alfter and Gerold Schneider 2020. Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. Proceedings of The 12th Language Resources and Evaluation Conference (pp.346-355). [chapter 14]

7. Alfter, David and Lars Borin and Ildikó Pilán and Therese Lindström Tiedemann and Elena Volodina 2019. Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. In Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018 (No. 159, pp. 1-14). Linköping University Electronic Press. [chapter 15]

8. Alfter, David and Johannes Graën 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?. In Proceedings of the 22nd Nordic Conference on Computational Linguistics (pp. 321-326). [chapter 16]

For publications numbered 1, 2 and 3, I was the main contributor regarding ideas, implementation and analysis.

For publication number 4, the methodology and design are partially based on a previous study (currently unpublished). I did the data preparation (extraction of items from the corpus and assigning levels), the technical implementation and the quantitative result analysis. The co-authors selected the items to be included in the experiment, selected definitions, checked the item senses and provided analyses of the results.

For publication number 5, the desired functionality and the guidelines were created in cooperation with the co-authors. I did the data preparation and extraction, automatic interlinking of different resources, the technical implementation, and the evaluation of the pilot phase.

For publication number 6, I did the data preparation and merging the multilingual lists. The other co-authors did the multilingual alignment, experiments and analysis of results. The evaluation methodology and discussion were done in cooperation with the co-authors.

For publication number 7, the original infrastructure is not my own work. However, I did partially re-implement and modernize the whole front end, extended the back end, added graphical user interfaces for Hitex and Texteval, and added new exercise types and tools.

For publication number 8, the co-author did the extraction of translation equivalents and example sentence selection as well as the technical implementation. The functionality was elaborated in cooperation with the co-author. I did the extraction of particle verbs. I had also started my own implementation of an exercise prototype but for time reasons we decided to go with an alternative implementation by the co-author.

Publications may have been visually altered to fit the current page format. No changes were made to content.

1.5 Structure of the thesis

The thesis is structured as follows. Part I contextualizes the work and summarizes the main points elaborated in the articles in part II.

Chapter 2 raises some key issues, introduces certain key notions, and describes related work in different areas adjacent and central to the main topic of the thesis. First, we discuss the notion of *word* and *multi-word expression*. We then discuss the notion of *complexity* from different angles, moving from text-based complexity research to single-word and multi-word lexical complexity. We then discuss the notion of *proficiency*. Finally, we describe different resources that have been used in complexity research.

Chapter 3 describes in more detail the source data and resources used in this thesis, the problems present in those resources, and how we address these problems in the future by re-creating the resources with the inclusion of sense distinctions and semi-automatic enrichment.

Chapter 4 gives a general overview of different methods used in this thesis.

Chapter 5 summarizes how we derived target proficiency levels from the resources, how we evaluated the level assignments, how we built an automatic proficiency prediction system based on this data, and how one could address the problem of predicting levels that were not included in the original data.

Chapter 6 moves from single words to multi-word expressions. We discuss the difference between single and multi-word expressions and why they cannot be treated by the same pipeline. We also check how well the automatic MWE recognition works by inspecting a manually corrected selection of texts. We explore the potential link between compositionality in MWEs and complexity. Further, we explore the use of crowdsourcing techniques for ranking MWEs by difficulty.

Chapter 7 describes applications of the presented thesis work. Some of the described application scenarios have been implemented, while others discuss possible future implementations. The chapter also discusses some more theoretical applications of this work.

Chapter 8 concludes part I. Here, we discuss the main findings and limitations, and elaborate on possible future work.

Part II consists of a compilation of publications.

Chapter 9 describes how we derive target proficiency levels from distributions of frequency over different proficiency levels.

Chapter 10 describes how we use the data from the previous chapter in order to train a classification algorithm that is able to predict the proficiency level of unseen words.

Chapter 11 describes our system entered in the 2018 Complex Word Identification Shared Task, where we adapted the proficiency prediction pipeline to English. For the non-English tasks, we used simple n-gram language models.

Chapter 12 describes an experiment on using crowdsourcing techniques to rank MWEs by perceived complexity in an L2 context.

Chapter 13 describes the lexicographic annotation tool Legato that was developed to facilitate (1) the correction of automatically assigned information and (2) the addition of lexicographic information.

Chapter 14 describes a comparison between English, Swedish and French word lists through the alignment via parallel multi-lingual corpora.

Chapter 15 describes the Lärka platform. Lärka is an experimental web platform offering different functionalities such as automatically generated exercises aimed at learners of Swedish and students of linguistics, a text evaluation tool, or the lexicographic annotation tool Legato (chapter 13). Most of the prototypes implemented as a result of my research are deployed in Lärka.

Chapter 16 describes a particle verb exercise which uses word lists and multilingual aligned corpora.

The appendix A contains a list of publications not included in this thesis. Appendix B contains an alphabetical list of abbreviations used in this thesis. Appendix C contains a list of resources both mentioned and used in this thesis.

2 Background and related work

In this chapter, we introduce the main theoretical notions used in this thesis and contextualize our work.

2.1 What is a word?

This question might seem trivial, but people with a background in linguistics know that that is far from the truth. The concept of "word" is multi-faceted and even to this day ill-defined (Jensen 1990; Dixon et al. 2002; Haspelmath 2011); the working definition often depends on the question to be answered by asking "What is a word?" (Nation and Meara 2013). Furthermore, the word "word" can represent more concrete units such as units appearing in spoken or written form, or a more abstract concept of units stored in the mental lexicon of speakers (Jensen 1990). Such abstract units are then called *lexemes* (Jensen 1990). As this thesis focuses on "single-word"- and "multi-word"- complexity, it is inevitable to explore how the concept of "word" can and has been defined across various disciplines, as well as to define and justify the choices we have made regarding the definition of "word" in this work. For a more extensive discussion on the notion of *word*, the interested reader is encouraged to read Langacker (1972: pp. 36-55).

Historically, words have been defined on the basis of meaning. For example, Zedler (1749) defines a word as "[...] ein vernemlicher Laut, der etwas bedeutet." '[...] a perceptible sound that means something.' (as cited in Haspelmath (2011)). Brugmann (1892: p. 3) differentiates between *composita* and *simplex* (i.e. *words*), defining a compositum as a unit composed of two or more simplex, and defining the simplex as being the result of separating a compositum into its constituting parts, whereby each of the parts loses its relative independence. Simplex in this definition can be single-morpheme words (i.e. units that bear meaning and can stand on their own, such as "cat") and affixes (i.e. units that bear meaning but cannot stand alone but have to be attached to other words, such as the prefix "un-", often implying the negation of the word it is attached to). However, he also points out that, from a historical perspective, units that are perceived as simplex nowadays could actually be considered composita in earlier times, thus blurring the distinction between simplex and compositum (Brugmann 1892: p. 5). The Princeton WordNet defines *forms* as a sequence of either phonemes (i.e. spoken) or characters (i.e. written), and *words* as *forms* having meaning (Miller 1995).

Another possible definition of words concerns orthography. Many languages written in the Greek, Latin or Cyrillic alphabets and variations thereof use *spaces* to separate meaningful units (Haspelmath 2011). In this definition, a word is a sequence of characters delimited by spaces on either side or by a space on one side and a punctuation mark on the other side.

Another definition of words concerns phonology. Therein, words are characterized as being delimited by pauses or other phonological or prosodic features such as final-consonant devoicing in German or Russian, vowel harmony borders in languages with vowel harmony such as Turkish or Hungarian, or stress patterns in languages with fixed stress patterns such as French or Spanish (Dixon et al. 2002; Haspelmath 2011).

There are certain schools of thought that question the naturalness or the existence of words. According to Bloomfield (1914) "words" are the product of theoretical reflection, and that such a subdivision cannot be taken for granted. He argues that the 'original datum of language' (i.e. the smallest possible subdivision) is the sentence (p. 65).

In contrast to the above definitions of words, construction grammar sees words, which are parts of the mental lexicon, as a combination of phonological, syntactic and semantic information (Croft 2007). The notion of *construction* encompasses a variety of linguistic concepts which tend to be separate in other grammar theories: constructions can be morphemes, words, multi-word expressions, fixed expressions and idioms, but also more abstract grammatical concepts such as passive constructions (Goldberg 2006: p. 5).

From a more computational perspective, a text can be seen as a sequence of tokens. However, dividing a text into separate tokens is a non-trivial task (Grefenstette and Tapanainen 1994; Nation and Meara 2013). Units in a text can be measured by counting *tokens*, *types*, *lemmas* or *word families*, among others; tokens are the actually occurring forms in a text, types are the *set* of tokens, i.e. actually occurring forms without counting duplicates, lemmas are dictionary forms of the actually occurring forms, and word families are words with a common root regardless of part-of-speech (e.g. happy, happiness, unhappy) (Bauer and Nation 1993; Nation and Meara 2010). This still leaves the problem of how to count *Multi-word units* (MWUs), also called *Multi-lexeme units* (MLUs) or *Multi-word expressions* (MWEs) among others.

The Merriam-Webster dictionary (Merriam-Webster) lists twelve distinct definitions under “word” as a noun. Of these, the first definition is subdivided into two parts, each of which is further subdivided into two parts. They are

1. (a) “a speech sound or series of speech sounds that symbolizes and communicates a meaning usually without being divisible into smaller units capable of independent use”
   (b) “the entire set of linguistic forms produced by combining a single base with various inflectional elements without change in the part of speech elements”
2. (a) “a written or printed character or combination of characters representing a spoken word”
   (b) “any segment of written or printed discourse ordinarily appearing between spaces or between a space and a punctuation mark”

Of these definitions, point 1a clearly targets spoken language, while point 2a indirectly characterizes spoken language through written language. Definition 1b is an intensional definition of a word as the set of all possible forms for a given base form without changing the part-of-speech. Finally, definition 2b coincides with the notion of orthographic word.

All of these definitions have their own advantages and shortcomings. Phonetic representations have the advantage of decoupling speech and writing, although one needs access to spoken material. Orthographic representations are easy to use but have some difficulties when it comes to (for example Swedish) compounds (are they *one* word?) or entities such as ‘New York’ (is it one word or two words?). There can also not be applied to languages that do not use spaces in written material. Construction grammar is a more all-encompassing theory although one loses the distinction between linguistic categories, should one wish to keep it, as everything is treated as a construction.

In this work, we investigate written Swedish texts from a computational perspective, and therefore we regard as “word” any sequence of orthographic characters delimited by spaces or by space and punctuation marks, thus effectively operationalizing “word” as *orthographic* word. This is in part justified by the use of computational techniques to treat the texts, in part by conventions used by other resources (Saldo), and in part also “simply because it’s convenient, without implying that the orthographic representation has any theoretical status” (Haspelmath 2011: p. 69).

2.2 What is a multi-word expression?

After having established our working definition of a word, we also have to clarify what we mean by multi-word expression. Analogous to the definition of *word*, the definition of multi-word expression is equally multi-faceted. To start with, different names have been given to these units that are larger than words, for example multi-word expressions (Sag et al. 2002), multi-word lexical units (Cowie 1992), collocations (e.g. Bhalla and Klimcikova 2019), phraseological units (e.g. Paquot 2019), lexicalized phrases (e.g. Sag et al. 2002), fixed expressions (e.g. Moirón 2005), formulaic language (e.g. Paquot and Granger 2012), lexical bundles (Chen and Baker 2010; Ädel and Erman 2012; Granger 2014), words-with-spaces (e.g. Sag et al. 2002), formulaic sequences (e.g. Wray and Perkins 2000), prefabricated units ("prefabs", Bolander 1989). While not perfectly synonymous, they all describe, in a certain sense, units that are bigger than single words and that "form a single unit of meaning" (Fazly and Stevenson 2007). In this work, we will use the term MWE.

MWEs can be characterized in different ways; a typical characterization of MWEs concerns their idiosyncrasy across one or multiple dimensions; Baldwin and Kim (2010: p. 2) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)" where the MWE can be decomposed into multiple simplex words. MWEs can exhibit

- lexical idiosyncrasy, meaning that the MWE does not allow for parts of itself to be replaced by (near-)synonyms without changing the meaning of the whole
- syntactic idiosyncrasy, meaning the parts combine in a syntactically unexpected way
- semantic idiosyncrasy, meaning that the meaning of the whole expression is not derivable from the meaning of its constituents
- pragmatic idiosyncrasy, meaning the expression is used in certain pragmatic contexts and not in others
- statistical idiosyncrasy, meaning that the combination of words occurs more often than expected

The degree and amount of idiosyncrasy can vary, thus creating a continuum of MWEs ranging from semantically transparent productive constructions to semantically totally opaque and syntactically fixed idioms (Howarth 1998: p. 28; Calzolari et al. 2002: p. 1934).

Linguistic characterizations such as those by Cowie (1992) or Burger (1998) subdivide MWEs into communicative phrasemes (pragmatic idioms), collocations, partially idiomatic expressions, proverbs, syntagmatic idiomatic expressions and routine formulae. Bauer (1983) divides MWEs into lexicalized and institutionalized phrases, with lexicalized phrases being further divisible into fixed expressions, semi-fixed expressions and syntactically flexible expressions. Another feature often discussed by linguists is semantic decomposability (Nunberg, Sag and Wasow 1994) which relates the meaning of an expression to the meanings of its constituents. The more compositional an expression is, the less idiomatic it is, and vice versa (Baldwin and Kim 2010).

While semantic compositionality and semantic transparency are often used interchangeably, they are different concepts (for an in-depth discussion, see Bourque 2014). Bourque (2014) defines semantic transparency as a scalar and multi-faceted property, meaning that (1) there is no binary distinction in transparency (transparent/opaque) but rather a continuum with varying degrees of transparency, and that (2) multiple factors influence semantic transparency; in this definition, compositionality is one of the multiple factors in transparency.

On the other hand, more computational and data-driven approaches tend to characterize MWEs by statistical measures such as association measures and corpora (Howarth 1998; Evert and Krenn 2005; Zhang et al. 2006; Villavicencio et al. 2007; Ramisch, Villavicencio and Boitet 2010). A popular association measure is point-wise mutual information (PMI). This is a common measure to find collocations. It measures how often two items x and y occur together, as opposed to occurring separately and is expressed as:

$$PMI(x,y) \equiv \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{1}$$

with $p(x|y)$ the probability of x given y and $p(x)$ the probability of x. However, purely statistical measures cannot differentiate between “true” MWEs and statistically significantly often co-occurring free combinations; neither can they differentiate between more semantically transparent and semantically more opaque expressions (Squillante 2014).

In-between the linguistic and the data-driven approaches, there are hybrid classifications which take into account both linguistic features and statistical measures (Baldwin and Villavicencio 2002; Van de Cruys and Moirón 2007; Gurrutxaga and Alegria 2013; Squillante 2014). It is argued that because linguistic measures are – especially from a computational point of view – rather vaguely defined, care must be taken when using linguistic features to ensure that the measurements are replicable and reliable (Sag et al. 2002; Laporte 2018). Features in this category are for example *substitutability*, i.e. whether words in an MWE can be switched for (near-)synonyms without losing or changing the meaning, or *interruptability*, i.e. whether an MWE allows for the insertion of other words between its constituents.

Given that this work is done from a computational linguistic point of view and, given that in this work, we use the Swedish Associative Thesaurus *Saldo* (Borin, Forsberg and Lönngren 2013) extensively, both as a pivot to connect different language resources as well as as a main resource of information, we have opted to follow Saldo's definition of MWE – which coincides with the definition given in Sag et al. (2002) except for the part concerning Swedish orthography – as "lexicalized (or even conventionalized) expressions containing spaces in their written form according to the standard orthography of Swedish." (Borin, manuscript). While on the surface, this seems to be the definition of *words-with-spaces* – an approach criticized for not taking into account syntactic flexibility (Baldwin and Kim 2010: p. 2) – Saldo's definition also covers particle verbs and other more flexible expressions which allow for the insertion of words or changing the order of words (e.g. *ha händerna fulla* 'have one's hands full').

As an aside, the definition of both a *word* and a *multi-word expression* has practical implications on what goes into a vocabulary list. Should vocabulary lists contain phrases and clauses? Answering such questions is beyond the scope of this thesis and we merely acknowledge the existence of such methodological concerns; for the purposes of the research presented in this thesis, we presuppose the existence of vocabulary lists (e.g. SVALex) or at least a previously established methodology for calculating such vocabulary lists from corpus data.

2.3 What is linguistic complexity?

The next point we have to clarify is *complexity*. Complexity itself –again– is a multi-faceted phenomenon. Generally speaking, complexity refers to a system containing a collection of objects that interact in multiple ways (Johnson 2009: p. 13). When referring to language, complexity can mean one of two things: it can either refer to the complexity of the language itself as a system (e.g. "Hungarian is a complex language") (Miestamo, Sinnemäki and Karlsson 2008), or to the complexity of sub-parts of the language *in* the language, such as text-level complexity (e.g. "This text is complex") (Housen and Kuiken 2009). The former is called *absolute* or *typological* complexity, while the latter is called *relative* complexity (Pilán and Volodina 2018).

Research on language complexity can target either complexity from a native-speaker (L1) perspective (e.g. this text is understandable by 7-year-old children; this text is understandable by first-year university students) or complexity from a second language learner (L2) perspective (e.g. this text is understandable by intermediate learners of Swedish; this text is understandable by beginner learners of Swedish whose mother tongue is German).

With regard to second language learners, complexity is also one of the dimensions in the *complexity, accuracy and fluency* (CAF) framework (Skehan and Foster 1999; Ellis 2003), wherein it is broadly defined as "the extent to which the language produced in performing a task is elaborate and varied" (Ellis 2003: p. 340). From a second language acquisition (SLA) perspective, complexity can mean different things: task complexity (properties of task) or L2 complexity (properties of L2 performance and proficiency) (Robinson 2001; Skehan 2001). L2 complexity can further be divided into cognitive complexity, i.e. the relative difficulty with which language features are processed in L2 performance and acquisition, and linguistic complexity.

In this work, we are interested in the relative complexity of single words and multi-word expressions in Swedish for second language learners. While complexity and *difficulty* can sometimes be understood as synonymous, it should be borne in mind that they are different concepts (Jensen 2009: p. 62); just as compositionality can be seen as a factor in semantic transparency, so can difficulty be seen as a factor in complexity. The following paragraphs elaborate on previous work with regard to different kinds of complexity that are relevant to this work before going into more detail about single-word and multi-word lexical complexity.

2.3.1 Linguistic complexity on text and sentence level

Text-level complexity, also called readability, is concerned with judging whether a piece of written material is understandable by a certain group of readers (Klare 1974: p. 1). It is based on the intelligibility of the writing and the ease of understanding, but rather than being focused solely on textual elements, another important factor is the interest of the reader (Bhagoliwal 1961; Mc Laughlin 1969; Council of Europe 2001). One of the most encompassing definitions of text-level complexity is given by Dale and Chall (1949):

> The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.

Besides human judgments as to the complexity of a text, several formulae have been developed to predict the complexity given a certain number of text-based features such as average sentence length and average word length. One of the most well-known text-level complexity formulae is the Flesch-Kincaid score (Flesch 1948). However, there is a plethora of other text-level complexity formulae such as the SMOG score (Mc Laughlin 1969), the LIX score (Björnsson 1968), nominal ratio Hultman and Westman (1977) or a lexically-enriched

LIX variant (Volodina 2008). Most early text-level complexity formulas target native language learner knowledge or native speakers with cognitive impairment. The Flesch-Kincaid scale for example indicates how many years of school, according to the American school system, one has to have finished in order to be able to understand a text. Work on text-level complexity in recent years has focused on using a more extensive feature set, taking advantage of advances in automatic text processing and tagging and advances in machine learning. The Coh Metrix system (Graesser et al. 2004) for example takes into account over 200 different text features.

Readability formulae are useful in the automatic assessment of text-level complexity, as they can process longer text segments such as books and they can process input faster than humans. However, it has been shown that human judgments concerning complexity judgments often perform better than judgments calculated by formulae (Klare 1974: p. 1). Since Klare's statement, a lot of work has been done on refining existing formulae and to create new complexity measures.

Text-level complexity can be assessed at different levels of granularity, starting with a binary distinction (easy-to-read versus hard-to-read) to more fine-grained scales such as the six-point[5] scale of the Interagency Language Roundtable (ILR) used predominantly in the United States or the six-point[6] scale of the Common European Framework of Reference (CEFR) (Council of Europe 2001) more commonly used in Europe.

The reasons for research on text-level complexity are manifold. A main aim is to increase the comprehensibility, especially in the legal domain (e.g. LoPucki 2014; Curtotti et al. 2015), in the medical domain (e.g. Deléger and Zweigenbaum 2009), for children (e.g. De Belder and Moens 2010), language learners (e.g. Petersen and Ostendorf 2007), or people with disabilities (e.g. Devlin 1998; Chung et al. 2013). Especially in a second language learner context, it is also used to select suitable reading material (e.g. Uitdenbogerd 2005; Ozasa, Weir and Fukui 2007; Kasule 2011; Xia, Kochmar and Briscoe 2016) or to assess learner productions (e.g. Nystrand 1979; Cohen and Ben-Simon 2011).

While text-level complexity has received a lot of attention, sentence-level and word-level complexity research is scarce at best and quite focused on specific points. While they are related, they are quite different; for example Oakland and Lane (2004) argue that text-level complexity formulae should not be used for passages of text shorter than paragraphs. This finding has been empirically corroborated by Sjöholm (2012) who, inter alia, compared how

[5] If taking into account sub-levels, the scale has 11 possible scores

[6] The scale is sometimes (arbitrarily) subdivided further

different text-level metrics work on sentence complexity prediction compared to sentence-based metrics and found that sentence-based metrics outperform text-level metrics in sentence complexity prediction in all cases.

There have been a few studies investigating sentence-level complexity. In contrast to text-level complexity, classification tends to be binary, i.e. distinguishing between *easy-to-read* and *hard-to-read* sentences (e.g. Dell'Orletta, Montemagni and Venturi 2011; Sjöholm 2012; Karpov, Baranova and Vitugin 2014; Vajjala and Meurers 2014), or distinguishing sentences of a certain level such as B1 from non-B1 sentences (e.g. Ahmad, Hussin and Yusri 2018; Pilán, Volodina and Johansson 2013). There has also been research into classifying sentences into three difficulty classes via the proxy of relative ranking (Howcroft and Demberg 2017). Moving from texts to sentences, one loses information about the broader context and co-reference. In order to make up for this loss of information, all of the above studies make use of the full feature set available at the sentence level, including lexical, morpho-syntactic and syntactic features.

Overall, the reasons for sentence-level complexity research overlap with reasons for text-level complexity research. For example in a text simplification scenario, text-level complexity evaluation can give an indication as to the overall complexity of the text; however, the simplification process itself targets sentences (Dell'Orletta et al. 2014). Sentence-level complexity is also used to find suitable sentences in automatic exercise generation (e.g. Pilán, Volodina and Borin 2017).

2.3.2 What is lexical complexity?

Lexical complexity focuses on the lexis (lexicon) of a language, i.e. on the vocabulary. Cutler (1983: p. 44) defines lexical complexity as the opposite of lexical simplicity, where lexical simplicity is defined as "the case when a phonetic representation of a word evokes a single lexical entry which contains only a single word class representation and a single semantic representation". This is not to say that "lexically simple" words cannot be complex; as with most factors, there is a spectrum between simplicity and complexity. Thus, if one were to group together all "lexically simple" words, one would still be able to subdivide this group into "more simple" and "less simple" groups to the desired level of granularity.

The complexity of lexical items can vary along different dimensions, namely syntactic, semantic and morphological (Cutler 1983: p. 43). This in turn gives rise to different types of complexity. For example, ambiguous words always activate all possible interpretations when encountered (Foss and Jenkins 1973).

Thus, ambiguous words can be seen as complex because they activate more than one semantic representation. Another example are idiomatic expressions. Idiomatic expressions are stored and accessed as lexical items (Swinney and Cutler 1979), as the meaning of their constituent parts does not allow for the derivation of the meaning of the whole expression.[7] Thus, idiomatic expressions can be seen as complex because of the many-to-one mapping from representation to meaning. On the other hand, words can be considered as complex if they have complex morphological structures such as is the case in synthetic and polysynthetic languages.

Often, frequency is taken as a proxy for lexical complexity (Rayner and Duffy 1986), i.e. the more frequent a word is, the less complex it is. This is especially prominent in the use of frequency lists, as elaborated in section 2.5. However, there are some attempts to move away from frequency as a measure of complexity, and instead use different indicators such as word co-occurrence (e.g. Li and Fang 2011; Brooke et al. 2012).

Different features might have seemingly opposing effects on complexity. If one considers for example frequency and polysemy, it might be argued that more frequent words would be less complex than less frequent words. Further, it might be argued that more polysemous words would be more complex than less polysemous words. However, more frequent words also tend to be more polysemous, leading to an apparent contradiction. It can further be argued that more frequent words might be seen as more complex *because* they tend to be more polysemous; this is partly corroborated by the findings of François and Watrin (2011). Each feature might contribute to complexity in different ways, and while for example frequency or polysemy could be used to approximate complexity on their own, we surmise that a combination of different features is able to give a more detailed picture; highly frequent non-polysemous words would be expected to have lower complexity than highly frequent polysemous words.

The importance of lexical complexity becomes apparent if one considers that lexical features have repeatedly been shown to be among the strongest predictors in text-level complexity assessment across several languages, both for L1 and L2 text-level complexity (Heilman et al. 2007; Brooke et al. 2012; François and Fairon 2012; Pilán, Vajjala and Volodina 2015; Reynolds 2016; Del Río Gayo 2019).

[7] This has been questioned and newer studies show that this may not be how idioms are encoded and accessed in the brain (Cieślicka 2015)

2.3.3 Single-word lexical complexity

Single-word lexical complexity is concerned with identifying and classifying the complexity of single words. This area has recently gained attention in the NLP community through two shared tasks on the topic of *complex word identification* (Paetzold and Specia 2016; Yimam et al. 2018). Complex word identification aims at classifying single words (and MWEs in the 2018 shared task) into simple and complex words (or expressions). The continued interest in the topic is made clear by the fact that there has been another CWI task for Spanish at the first Lexical Analyis Workshp at SEPLN (ALexS; Ortiz-Zambranoa and Montejo-Ráezb 2020) and that there will be another shared task on lexical complexity prediction for single- and multi-word expressions in 2021.[8]

In the 2016 shared task, participants were asked to predict *complex* words in English for downstream tasks such as lexical simplification, i.e. replace complex words and expressions with simpler alternatives (e.g. Specia, Jauhar and Mihalcea 2012; De Belder, Deschacht and Moens 2010; Shardlow 2014). Participants were given a sentence and a target word within the sentence and had to predict whether or not a non-native English speaker would be able to understand the meaning of the target word. The dataset was annotated by 400 non-native English speakers. Each instance was annotated by 20 annotators, and a word was considered complex if at least one of the 20 annotators marked it as complex. In total, 42 systems were submitted by 21 teams.

The 2018 shared task extends on the previous shared task by including three different genre datasets for English (News, WikiNews and Wikipedia) as well as datasets for German, Spanish and French. For French, no training data was released in order to see whether it was possible to construct language-agnostic complex word identification systems. Each data point for the English dataset was annotated by 10 native and 10 non-native English speakers, while for the German, Spanish and French data, each entry was annotated by 10 annotators (native and non-native speakers; exact ratios not disclosed). In total, 12 teams participated, including our own system described in chapter 11.

In the 2020 shared task, the organizers used transcriptions of academic video lectures in Spanish. However, no training data was provided, making the task an unsupervised one. The data consists of 55 transcriptions that were manually annotated for complex words by 430 students. In total, 3 teams participated.

In the 2021 shared task, participants can partake in two tasks: lexical complexity prediction for single words and lexical complexity predictions for multi-word expressions. In contrast to the 2018 shared task, the 2021 shared task only

[8] `https://sites.google.com/view/lcpsharedtask2021`

includes English data. Another difference to the two previous shared tasks is that the 2021 shared task predictions are to be done on a five-point scale (instead of a binary classification).[9] This, in turn, brings this task much closer to the work done in this thesis.

On the topic of second language learner focused complex word identification, which is the focus of the present thesis, several approaches have been taken. One approach taken by multiple researchers is to classify words into known and unknown words. In order to gather data, words are typically annotated manually for complexity, either in isolation (i.e. without contextual information) (e.g. Avdiu et al. 2019; Ehara et al. 2012, 2018; Lee and Yeung 2018) , or within a text (e.g. Tack et al. 2016a, b; Yancey and Lepage 2018). The resulting data is then used to train classifiers able to predict whether a word is complex or not. All of these works additionally use personalized models for each learner, i.e. the prediction as to whether a word is complex is dependent on the learner; a word might be complex for one learner but not for another learner. Palmero Aprosio, Menini and Tonelli (2020) also include the L1 of learners in order to detect false friends between different languages.

Another approach is to classify words into different proficiency levels, i.e. levels at which these words are introduced or otherwise targeted as a learning goal. Proficiency levels are typically derived from graded textbooks, as is done in the CEFRLex project,[10] or based on graded learner essays as has been done for example in SweLLex (Volodina et al. 2016b). Further, proficiency estimations can be based on expert knowledge such as the Global Scale of English (GSE), or based on a hybrid approach such as in the English Vocabulary Profile (EVP) which combines graded learner essays and expert knowledge. Similarly, Pintard and François (2020) combine expert knowledge in the form of the French reference descriptors (Beacco, Bouquet and Porquier 2004; Beacco and Porquier 2007; Beacco et al. 2008, 2011) with the French CEFRLex list. Section 2.5 elaborates further on data that has been used for lexical complexity research.

2.3.4 Multi-word lexical complexity

In contrast to single word expressions, which we characterize in linguistic terms, we focus on perceived complexity for multi-word expressions; in other words, we look at complexity of MWEs through the proxy of CEFR levels. The aim is the same for both approaches, namely to identify target proficiency

[9] The scale for the shared task ranges from 1 to 5 and is worded as “very easy”, “easy”, “neutral”, “difficult” and “very difficult”.

[10] http://cental.uclouvain.be/cefrlex/

levels for expressions, although the methodology to arrive at said target levels is different.

To the best of our knowledge, very few studies exist on the topic of linking MWEs to CEFR levels, and most of the work done on the topic falls under *resource creation* (see section 2.5) rather than research.

López-Jiménez (2013) investigates MWEs in L2 textbooks. In the study, the author looks at lexical collocations, compounds, and idioms, and how they are represented and practiced in 12 English textbooks and 12 Spanish textbooks covering three proficiency levels (beginner, intermediate, advanced). Results show that the amount of lexical collocations and idioms is practically identical in English and Spanish textbooks, whereas English textbooks contain about 25% more compounds than Spanish textbooks. Given the Germanic nature of English, it has a propensity for compounding in contrast to Spanish (and other Romance languages) which tend to use affixation and derivation rather than compounding (Renner and Fernández-Domínguez 2011: p. 3). The study also shows that there is an increase in the number of MWEs from beginner level to intermediate to advanced levels.

Chen and Baker (2016) investigate lexical bundles in CEFR-graded learner essays. Lexical bundles are recurring continuous word sequences, and due to their purely statistical description, often fall into the category of pragmatic and discourse markers. The authors extracted four-word sequences such as "there are a lot", "on the other hand", "is one of the". The aim is to find *criterial* features, i.e. features that demarcate one proficiency from another, using lexical bundles in a learner corpus of Chinese learners of English ranging from B1 to C1.

The main research problem with MWEs lies in the automatic identification of MWEs from text and the quantification of their degree of compositionality (sometimes called *idiomaticity*), i.e. how much the meaning of the parts making up an MWE contribute to the meaning of the MWE. As MWEs encompass a multitude of different constructs, most studies focus only on a specific type of MWE in a specific language, such as noun-verb expressions in Basque (Gurrutxaga and Alegria 2013), German noun-noun compounds (Im Walde, Müller and Roller 2013), compositionality in verb-particle constructions/phrasal verbs (Bhatia, Teng and Allen 2017; McCarthy, Venkatapathy and Joshi 2007) or verb-noun expressions (Taslimipoor et al. 2017; Venkatapathy and Joshi 2005).

2.4 Second language learner proficiency and linguistic complexity

In Second Language Acquisition (SLA), the notion of *proficiency* is a key concept. It describes the language “knowledge, competence, or ability” of a learner (Bachman 1990: p. 16, as cited in Carlsen 2012: p. 163). Conventionalized scales of proficiency levels are used in educational and assessing contexts, e.g. which group to place a student into (Bachman and Palmer 2010). However, a straightforward division into levels is a tricky endeavor, since there is no consensus how to define a level and its corresponding competence(s) in concrete terms. SLA research views proficiency as a “coarse-grained, externally motivated” construct (Ortega 2012: p. 134), where levels are always somewhat arbitrary (Council of Europe 2001: p. 17); further, one should distinguish between proficiency and *L2 development* which is “an internally motivated trajectory of linguistic acquisition” (Ortega 2012: p. 134). This parallels the notions of performance and competence in Chomskyan terms (Chomsky and Halle 1965); indeed, only performance can actually be assessed, while performance itself stems from internal competence. However, as we found in chapter 12, these two concepts seem rather strongly correlated.

Proficiency and L2 development fall on a continuum rather than into discrete and distinct categories and current research advocates that proficiency be seen as a continuum (Ortega 2012; Paquot, Naets and Gries 2020), as such views yield more realistic and nuanced results. Yet, for practical reasons, proficiency is often regarded as a set of related but distinct levels (Council of Europe 2018: p. 34).

Proficiency and complexity are related in the sense that as one becomes more proficient, one is confronted with texts of higher *complexity*, and one is also expected to produce more *sophisticated* writing (Council of Europe 2018: p. 110). Thus, it can be expected that proficiency and complexity evolve in tandem.

As with complexity, proficiency is a multi-faceted phenomenon. The Common European Framework of Reference (CEFR) for Languages (Council of Europe 2001) used in this thesis subdivides proficiency into three main categories: understanding, speaking and writing (Council of Europe 2001: p. 25); at the same time, they propose another subdivision, namely reception, production and interaction (Council of Europe 2001: p. 222), each of which is further subdivided into speaking and writing skills. In this work, we adopt the latter distinction between reception and production. We leave out *interaction*, as this notion covers both productive and receptive aspects with the only difference being the direct involvement of at least one other person in an interactive communicative setting. As we work exclusively with text-based material in the form of written textbooks and learner essays, the notions of *receptive* and *pro-*

ductive knowledge are taken to mean that a learner can *understand* respectively *produce* a certain word or expression in writing.

2.5 Lexical complexity research and resources

When selecting vocabulary items of an appropriate level, one possible approach is to use human judgments, e.g. teachers insights. However, this is a very subjective approach, as people's intuitions can vary greatly (Richards 1974: p. 70). A more objective way of selecting vocabulary comes in the form of word lists. Notable examples of – and despite their age still widely used – word lists for English are the *General service list of English words* (GSL) (West 1953) for general English and the *Academic word list* (AWL) (Coxhead 1998) for academic English.

For Swedish, one can name the *Nusvensk frekvensordbok baserad på tidningstext* 'Frequency dictionary of contemporary Swedish based on newspaper texts' (Allén 1971), *Tiotusen i top* 'Top ten thousand' (Allén 1972), *Frekvensordbok över svenska elevtexter* 'Frequency dictionary of Swedish learner essays' (Larsson, Rosén and Anderson 1985), *Talspråksfrekvenser* 'Colloquial language frequencies' (Allwood 1999), *Base Vocabulary Pool* (Forsbom 2006), or *Akademisk ordlista* 'Academic word list' (Sköldberg and Johansson Kokkinakis 2012).

However, word lists are criticized for being very different based on the source material they are compiled from; similar word lists rarely overlap to a large degree (Bongers 1947). In addition, word lists derived from L1 material rarely cover the type of vocabulary that is used by and useful for learners and teachers (Richards 1974: p. 72; François et al. 2014: p. 3767). Moreover, word lists quickly become obsolete, covering only the language used up to the point of its compilation. Efforts have been made to recompile existing lists such as the GSL resulting in **two** *New General Service Lists* (called new-GSL and NGSL respectively) (Brezina and Gablasova 2015; Browne 2014), and the Academic word list, resulting in the *New Academic Word List* (Coxhead 2000). Although it is quite dated, the original AWL is still being actively used in for example vocabulary testing (Coxhead 2011: p. 358).

Jackendoff (1997: p. 156) states that the amount of MWEs is about equal to the amount of single-word units in the mental lexicon. However, neither the GSL nor the AWL contain MWEs. In order to address this issue, Martinez and Schmitt (2012) created a list of phrasal expressions to complement these lists.

Another project focusing on producing word lists was the KELLY project (KEywords for Language Learning for Young and adults alike). The project aimed at creating lists of about 9000 words in nine different languages, Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian, and Swedish. The Swedish KELLY list is based on a large web-crawled corpus (Swedish Web as a Corpus, SweWaC). In order to compile the list, words are ordered by frequency, but dispersion is also taken into account, i.e. if a word only occurs in one sub-corpus, it is considered less central to the core vocabulary (Volodina and Kokkinakis 2012). The resulting ordered list is then subdivided into equal-sized slices and each slice is assigned a CEFR level, starting with A1 for the first slice and ending with C2 for the last slice.

Another resource that can be used for lexical complexity research is the MRC psycholinguistic database (Coltheart 1981). The MRC database lists entries along with several psycholinguistic dimensions such as age of acquisition, familiarity, imageability, and concreteness among others. While not directly related to lexical complexity, such features can be used to enhance lexical complexity predictions. We have used the machine readable version of the MRC psycholinguistic database (Wilson 1988) as a source of features in lexical complexity predictions as described in chapter 10.

Another project that works on single-word (and multi-word) complexity is the CEFRLex project[11]. CEFRLex is a family of related resources, each derived from a CEFR-graded textbook corpus (with the exception of SweLLex, which is derived from a CEFR-graded learner essay corpus). Each resource lists lemmata, part-of-speech and the normalized frequency across different levels of textbooks as illustrated in table 2.1. The currently available resources are FLELex (François et al. 2014) for French, EFLLex (Dürlich and François 2018) for English, SVALex (François et al. 2016) and SweLLex (Volodina et al. 2016b) for Swedish, ELELex (François and De Cock 2018) for Spanish and NT2Lex (Tack et al. 2018) for Dutch.

Another resource for English that covers productive knowledge – albeit not exclusively – is the English Vocabulary Profile[12] (EVP) (Hawkins and Buttery 2010; Hawkins and Filipović 2012). EVP lists single words, phrasal verbs, phrases, and idioms along with a CEFR level, definitions, examples from learner essays. They also differentiate between different senses, thus *bark* as a verb has two entries, one pertaining to the sound a dog makes (B2) and one to the sense of shouting harshly (C2). (Of course there is also *bark* as a noun). EVP is based on the Cambridge Learner Corpus, different coursebooks, word lists and vocabulary lists, the Cambridge English Corpus, and the Cambridge English Lexicon by Roland Hindmarsh. Outside experts were also consulted for validation of the resource. In the same line, the French Vocabulary Profile (FVP) – also known as *Référentiels* – (Beacco, Bouquet and Porquier 2004;

[11] http://cental.uclouvain.be/cefrlex/

[12] https://www.englishprofile.org/wordlists

Expression	PoS	A1	A2	B1	B2	C1	Total
video	noun	2.47	0.56	34.83	23.80	13.25	18.43
write	verb	934.71	378.34	760.73	536.38	713.33	549.91
empty	adjective	86.49	150.89	65.95	194.80	123.41	156.02
shopping center	noun	0	0	15.58	0	0.82	1.75
dream up	verb	0	0	0	0	0.82	0.23

(a) Example entries from EFLLex

Expression	PoS	A1	A2	B1	B2	C1	Total
bil ‘car’	noun	430.21	1234.20	728.98	422.283	363.54	618.85
kilo ‘kilogram’	noun	0	302.08	145.12	65.06	13.21	89.89
resa ‘travel’	verb	166.30	375.25	450.35	298.49	330.42	356.36
låg ‘low’	adjective	0	49.315	125.922	217.31	252.13	156.12
så klart ‘of course’	adverbial MWE	0	16.26	81.60	45.50	13.21	38.17

(b) Example entries from SVALex

Expression	PoS	A1	A2	B1	B2	C1	Total
bil ‘car’	noun	0	395.72	441.49	38.60	88.38	219.59
rättvisa ‘justice’	noun	0	0	0	0	0.14	0.02
kilo ‘kilogram’	noun	0	0	0	0.26	0	0.01
resa ‘travel’	verb	0	134.27	1297.91	183.22	43.66	363.79
låg ‘low’	adjective	0	0	0	37.45	76.58	41.09

(c) Example entries from SweLLex

Table 2.1: Example entries from EFLLex, SVALex and SweLLex (relative frequencies normalized to per-million-word and adjusted using dispersion)

Beacco and Porquier 2007; Beacco et al. 2008, 2011) lists single- and multi-word expressions with associated CEFR levels. In contrast to the EVP, the FVP is based more on introspection and synthesized experience; while acknowledging the value of the EVP's method, "their lack of adequate financial resources prevented them from using this methodology" (Kusseling and Lonsdale 2013: p. 439).

Similarly to the EVP, the Global Scale of English (GSE) Teacher Toolkit[13] contains, inter alia, English words and phrases linked to the GSE scale which is a further subdivision of the CEFR scale ranging from 10 to 90 in increments of 1. Entries are organized by topics and contain British English and American English pronunciations, definitions, example sentences and collocations. As for the EVP, entries are based on senses, thus *bark* as a verb also has two entries; the scores are 63 ($\approx$ B2) for the sound a dog makes and 81 ($\approx$ C1) for speaking in a loud angry voice. In contrast to EVP, GSE is based on expert knowledge.

As argued by Paquot et al. (2007), word lists generally target *receptive* learner knowledge, i.e. their understanding of words and phrases; however, more research should be put into targeting *productive* knowledge, e.g. word lists that cover learner *productions* such as extracted from learner essays. The SweLLex list for Swedish answers to this concern as it is calculated on the basis of the CEFR-graded learner essay corpus SweLL-pilot (Volodina et al. 2018) and thus represents what learners have *produced* at different levels of proficiency.

[13] `https://www.pearson.com/english/about/gse/`

3 Data and resources

In this chapter, we investigate in more detail the data and resources used in this thesis, discuss shortcomings of the data and elaborate on how to address these issues.

3.1 The Swedish associative thesaurus Saldo

The Swedish associative thesaurus Saldo (Borin, Forsberg and Lönngren 2013) plays a pivotal role in this work. It builds on the Swedish Associative Thesaurus SAL (Svenskt associationslexikon) (Lönngren 1988), a semantic lexicon for Swedish. Saldo is an electronic lexicon covering the modern written Swedish language and is designed for language technology applications, as opposed to for example dictionaries which are designed for humans. Thus, the data format may not necessarily be readily accessible to humans. Further, there are no examples sentences or similar information. The basic principle behind Saldo (and SAL) is semantics. Thus, the basic unit is the word *sense*. Saldo has been semi-automatically expanded and merged with other resources (Borin 2005), and is still under active development. In contrast to other lexico-semantic networks such as Princeton WordNet (Miller 1995; Miller and Fellbaum 1998; Fellbaum 1998), the semantic links in Saldo are less specific.

Saldo is one of the main resources behind Sparv, the automatic tagging and annotation tool.[14] For example, lemmatization and word sense assignment is done on the basis of Saldo (Borin et al. 2016). This should be borne in mind, as we base our research on the output of this automatic annotation, and thus inherently rely on Saldo. It can also be noted that Saldo not only plays an important role in this work but in Språkbanken Text's resources in general as more than 25 different lexical resources are linked together via Saldo.

[14] https://spraakbanken.gu.se/sparv/

3.2 From corpus to word list

The main data sources for this research are two graded corpora, the textbook corpus COCTAILL (Volodina et al. 2014a) and the learner essay corpus SweLL-pilot (Volodina et al. 2016a). The COCTAILL corpus consists of 18 textbooks aimed at learners of Swedish in Sweden and covering different proficiency levels from beginner levels (A1) to advanced levels (C1). The initial textbook collection also contained 3 textbooks of level C2. However, these textbooks were excluded during the compilation of the corpus, as the language represented therein "corresponds to regular NS [native speaker] language" (Volodina et al. 2014a: p. 132). The SweLL-pilot corpus consists of 339 learner essays covering the CEFR levels A1 to C1.[15] All essays have been manually graded by CEFR experts. Both corpora have been enriched with manual metadata annotation and automatically annotated for linguistic features using Sparv (Borin et al. 2016), Språkbanken Text's annotation pipeline. These steps add various information, of which *lemma* and *part-of-speech* (pos) are the most important for the creation of the word lists; we would want to be able to distinguish for example '[a] meeting' and '[to] meet' as in examples (1-a) and (1-b) (different lemma and part-of-speech) or '[a] *bear*' from '[to] *bear*' (different part-of-speech).

(1) a. We have a *meeting* at 12.
    b. We are *meeting* at 12.

From these two corpora, two word lists we created, namely SVALex (François et al. 2016) and SweLLex (Volodina et al. 2016b). In order to create the word lists from the annotated corpus, one counts how often each unique lemma-pos tuple occurs in the textbook texts/learner essays of different levels in order to arrive at a raw count of items across different CEFR levels. The raw counts are then normalized to per-million-word counts and adjusted according to *dispersion*, i.e. if an item occurs only in a single text or a single essay, it will be penalized; if an item occurs across all textbook texts or essays of a certain level, it will be assumed to be more representative of that level than instances occurring only in few texts.

3.3 Shortcomings of the word lists

The first problem with the data is that it has been automatically processed in different ways; automatic processing is not always accurate. For the COC-

[15] The SweLL-pilot corpus has since been extended to about 500 essays.

TAILL corpus, textbooks have been digitally scanned and automatically transcribed using optical character recognition (OCR). The resolution of the scan, along with other factors such as lighting conditions, can negatively impact on OCR. Moreover, OCR itself is not 100% accurate. The resulting digital corpus has then been manually annotated for pedagogical and textual features such as text topics, genre, skills and competences targeted (e.g. listening, speaking, writing, grammar, vocabulary), and text structure such as identifying and explicitly marking for example lessons, tasks, and reading comprehension texts. The corpus was then automatically processed and annotated for a variety of morphological, lexical, syntactic and semantic features, each of which may be wrong. The problems are the same for the SweLL-pilot corpus, which additionally may contain spelling errors and other infelicities, which may lead to further errors during the automatic annotation step. Also, for the SweLL-pilot corpus, manual transcription and annotation of learner data can be problematic, for example if the transcriber cannot decipher certain letters or words. In addition, humans may make mistakes, thus they might introduce errors in the process of transcription and annotation.

Another point of critique of the original SVALex and SweLLex lists is that they are lemma-based, thus conflating polysemous entries into a single entry. In this context, it should be noted that not all senses of a polysemous word are learned at once. Furthermore, if all senses are conflated, they can only be assigned a single level, which may be problematic when deciding when a learner should be able to understand or produce a certain word *sense* (cf. chapter 5).

In an L2 study on French, it was found that unigram language models showed a correlation between frequency and complexity, although not in the expected direction, meaning that more frequent words were associated with more complex texts (François and Watrin 2011). The authors expected more frequent words to occur more often in easy texts. This could well be due to the conflation of word senses. Indeed, more frequent words also tend to have more senses than less frequent words (Crossley, Salsbury and McNamara 2010). Thus, certain *senses* of frequent words could have been used more often in complex texts. In addition, more complex texts possibly make use of more different senses of frequent words.

3.4 Towards a sense-based graded vocabulary list

In order to address the aforementioned problems, we have re-created SVALex with sense distinctions, and further manually corrected and enriched the list. In order to reflect the distinction between the original and re-created sense-disambiguated word lists, we re-brand it as SenSVALex.

At the time of creation of the original lists, the annotation pipeline did not include a word-sense disambiguation module, thus the original lists could not have been built on a sense-level. However, word-sense disambiguation was later added to the pipeline, allowing for a sense-based version of the lists. It should be noted that the initial experiments described in chapters 9 and 10 were done on the original lists while subsequent experiments were done on the sense-based lists.

The SweLL-pilot corpus is currently undergoing active change, which entails the addition of 600 essays; essays are normalized, pseudonymized and annotated for corrections (Volodina et al. 2019), hence we focus on COCTAILL and SenSVALex. The methodology would remain the same for both corpora. It is hoped that in the future, a sense-based version of SweLLex will be created under the name SenSweLLex. However, the creation of that list is out of scope for this thesis and left for future work.

The original COCTAILL corpus was re-tagged by the new annotation pipeline to include information about senses. To illustrate the output of the annotation pipeline, figure 3.1 shows an excerpt of the annotation output in XML format. The tagged sentence is *Ni har tre hundar.* 'You have three dogs.'. In this XML fragment, one can see for example that the word *hundar* has been assigned two senses, *hund..1* and *hund..2*, of which *hund..1* has been assigned a probability of 69% of being the correct sense in this context.

The information of importance are *lemma*, the dictionary form of the word, *pos*, the part-of-speech, *lex*, the *lemgram*, which is a concatenation of the lemma, part-of-speech and a numerical identifier, and *sense*, which indicates the sense(s) the word could have along with the probabilities for each sense being the correct one in the current context. Note that the part-of-speech given by the *pos* tag and the part-of-speech given in the *lex* tag can be different since they are based on different tag sets.

We then calculate how often each lemma-pos-sense tuple (as opposed to the previously used lemma-pos tuple) occurs at each CEFR level. This yields a list with 22654 unique entries, of which 14539 entries are 'non-problematic', 5984 have not been assigned a sense, 1262 entries have one sense but multiple lemgrams, 807 entries have multiple senses and one lemgram, and 62 entries have multiple senses and multiple lemgrams. Problematic entries are checked manually and their treatment depends on the type of problem. Table 3.1 shows the distribution of new items per level over CEFR levels in comparison to the original SVALex.

Finally, the list is enriched with additional lexicographic information, both automatically and manually as described in chapter 13, as part of the L2 pro-

```
<w pos="PN" msd="PN.UTR.PLU.DEF.SUB" lemma="|Ni|ni|"
 lex="|Ni..pn.1|ni..pn.1|" sense="|ni..1:-1.000|"
 complemgram="|" compwf="|" ref="1" dephead="2"
 deprel="SS">Ni</w>
<w pos="VB" msd="VB.PRS.AKT" lemma="|ha|"
 lex="|ha..vb.1|" sense="|ha..1:0.618|ha..3:0.382|"
 complemgram="|" compwf="|" ref="2" dephead=""
 deprel="ROOT">har</w>
<w pos="RG" msd="RG.NOM" lemma="|tre|"
 lex="|tre..nl.1|" sense="|tre..1:-1.000|"
 complemgram="|" compwf="|" ref="3" dephead="4"
 deprel="DT">tre</w>
<w pos="NN" msd="NN.UTR.PLU.IND.NOM" lemma="|hund|"
 lex="|hund..nn.1|"
 sense="|hund..1:0.690|hund..2:0.310|"
 complemgram="|" compwf="|" ref="4" dephead="2"
 deprel="OO">hundar</w>
<w pos="MAD" msd="MAD" lemma="|" lex="|" sense="|"
 complemgram="|" compwf="|" ref="5" dephead="2"
 deprel="IP">.</w>
```

Figure 3.1: Example of annotated data

Level	# new items SVALex	# new items SenSVALex
A1	1157	1595
A2	2432	3452
B1	4332	6236
B2	4552	6794
C1	3160	4577
Total	15681	22654

Table 3.1: Distribution of entries per CEFR level in SVALex and SenSVALex

files project[16] financed by Riksbankens Jubileumsfond, grant number P17-0716:1. The L2 profiles project aims at identifying *central* (i.e. core) and *peripheral* (i.e. “good-to-know”) grammar and vocabulary items at each of the CEFR levels in order to support teachers, test makers, assessors and learn-

[16] https://spraakbanken.gu.se/en/projects/l2profiles

ers. The manual annotation for various lexicographic linguistic aspects such as morphological analysis is ongoing work. Further, problematic entries from SenSVALex are being manually corrected by assistants at the University of Gothenburg and the University of Helsinki.

4 Methods

In this chapter, we give an overview of different methods and techniques used in this thesis. For machine learning, the choice of algorithms is based on a literature review of previous work. Language models are widely used across a multitude of applications in NLP. Deep learning has reached state-of-the-art in many use cases, thus we include it in our experiments. The use of custom embedding models and crowdsourcing to rank expressions are largely experimental in nature.

4.1 Word embeddings

Word embeddings are based on the assumption that similar words occur in similar contexts, or in Firth's words: "You shall know a word by the company it keeps". In word embeddings, words are represented as points in a high-dimensional space (although the dimension is comparatively small [typically 100 to 300] in comparison to a space where each word would represent a separate dimension), with words that often occur together being closer in this space. Word embedding models have been shown to capture not only relations such as synonymy and antonymy but also other semantic characteristics, allowing for "mathematical operations" at the word level (Mikolov and Dean 2013). Thus, a well-trained embedding space will respond to the query "France is to Paris as Germany is to *what*?" with "Berlin" ($Paris - France + Germany = Berlin$). Similarly, the embedding space may learn that $coldest - cold + big = bigger$.

Recently, word embeddings have moved from word *forms* to word *senses*, either by directly building a sense-based embedding space (Nieto Piña 2019) or by including contextual information in models such as BERT (Devlin et al. 2018) or ELMo (Peters et al. 2018).

In this thesis, we used the word2vec model (Mikolov and Dean 2013) from the gensim package (Řehůřek and Sojka 2010), implemented in Python. A customized word embedding model is used in chapter 9 in order to evaluate automatic CEFR level assignments and to build an embedding space in which

CEFR labels are co-present with words; thus, we can calculate the distance not only between words but also between words and CEFR labels.

One hypothesis, although it remains to be tested, would be that custom word embeddings can be used to find easier or more difficult synonyms or otherwise related words using “mathematical operations”, for example $cold - A1 + B1$.

4.2 Machine learning

Machine learning is a broad umbrella term under which one finds a large number of different algorithms. In essence, machine learning is concerned with teaching computers to find solutions to problems based on example data or past experience (Alpaydin 2020: p. 3). One defines a *model* with *parameters* and then executes a program to learn which set of parameters best satisfy a given criterion (Alpaydin 2020: p. 3). Further, machine learning can be broadly subdivided into *supervised* and *unsupervised* methods. In supervised machine learning, one has ground truth data, i.e. one knows the solution to the problem at hand, at least for a certain amount of data, so that the machine learning algorithm can compare its own output to the true solution and approximate the true solution as best as possible. In unsupervised machine learning, one does not have any information about the solution and the algorithm tries to detect commonalities, patterns, or other kinds of regularities (Alpaydin 2020: p. 11).

In this thesis, we use the following supervised classification algorithms: Support Vector Machine (SVM) (Hearst 1998), MultiLayer Perceptron (MLP) (Vapnik and Vapnik 1998), Random Forests (RF) (Breiman 2001) and Extremely Randomized Trees (ExtraTrees, ET) (Geurts, Ernst and Wehenkel 2006). Supervised classification algorithms take as input labeled data and learn how to classify new data instances into one of the predefined (i.e. learned) classes based on a list of features extracted from the data. We use these algorithms for classification in chapters 10 and 11 in order to classify expressions into CEFR levels.

There are multiple implementations of machine learning algorithms for different programming languages available. In this thesis, we used the Waikato Environment for Knowledge Analysis (WEKA) toolkit (Hall et al. 2009), implemented in Java, and the scikit-learn package (Pedregosa et al. 2011) for Python. The WEKA toolkit is a standalone application that can be run without programming knowledge. The scikit-learn package requires Python programming in order to be used.

As the amount of data available is rather limited, “traditional” machine learning algorithms are a logical choice, as deep learning models generally require vast amounts of training data. However, we experiment with deep learning in the shared task, as the amount of data available for training is more substantial.

4.3 Deep learning

Deep learning is a subset of machine learning that aims at learning through a succession of *layers*, the number of which determines the *depth* of the network (Chollet 2018: paragraph 1.1.4). Each layer learns certain kinds of information and subsequent layers built upon earlier layers, learning increasingly more complex concepts, thus forming a hierarchy going from simple concepts in early layers to more complex concepts in later layers (Bengio, Goodfellow and Courville 2017: p. 5). Each layer in the network is made up of nodes, in analogy to neurons.

In this thesis, we experiment with two different deep learning architectures, namely convolutional neural networks (CNN) and recurrent convolutional neural networks (RCNN).

These networks are named after the mathematical operation of *convolution*. A convolution takes two functions and returns a third function (Hirschman and Widder 2012: p. 3). The exact description of the convolution operation is highly mathematical and not conducive to the aims of this thesis. It is thus left open to the interested reader to look up further information on the topic. For convolutional neural networks, convolutions consist of moving multiple “kernels” (fixed-size windows) over the input and calculating a function between the kernel and the input.

Figure 4.1 illustrates a simple CNN for image input. The input image is first transformed into a different form, most often by taking the raw image data and splitting it into the three color channels RGB (red, green, blue). The resulting input has the same width and height as the input, and a depth of 3, thus representing a 3-dimensional input.

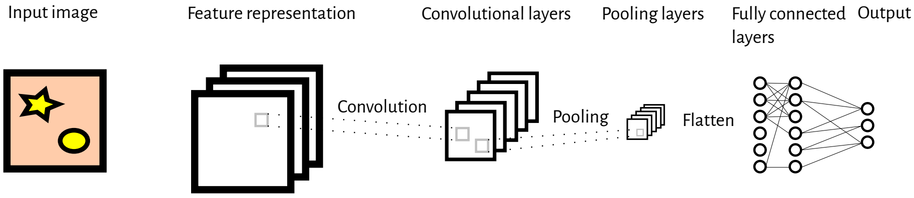
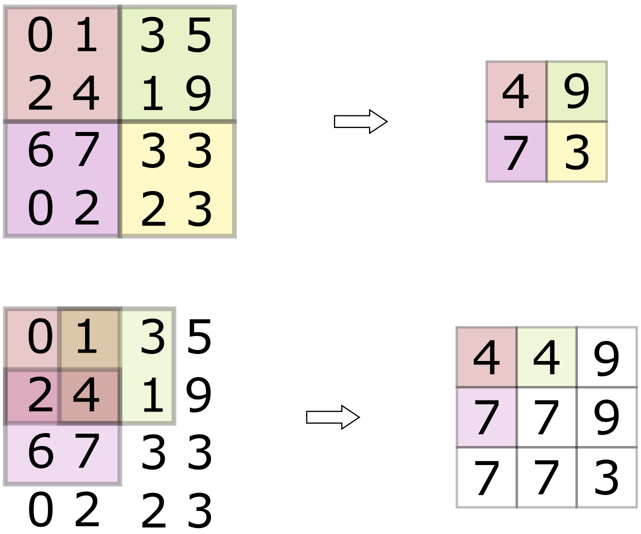


Figure 4.1: Schematic illustration of a simple CNN network for image classification

The figure shows one position of a kernel on the input (“Feature representation” in the figure) which is mapped onto a new layer via the convolution operation. The size of the resulting layers is further compressed by an operation called *pooling*. Pooling takes a kernel (fixed-size window) and maps its input into a fixed-size window of smaller dimension. There are multiple pooling options such as max pooling, min pooling or mean pooling. For max pooling, the pooling operator takes the highest value in its window and outputs that value. Figure 4.2 illustrates the concept of max pooling with a 2x2 kernel and strides of 2 and 1 respectively. With stride 2, one moves two spaces after each operation. Each color-coded block stands for one position of the kernel on the left side and its output on the right side. With stride 1, one moves one space after each operation, resulting in the second output in the figure; only three of the kernel positions are indicated as color-coded blocks as point-of-reference.



0	1	3	5
2	4	1	9
6	7	3	3
0	2	2	3

4	9
7	3

0	1	3	5
2	4	1	9
6	7	3	3
0	2	2	3

4	4	9
7	7	9
7	7	3

Figure 4.2: 2x2 max-pooling illustration with strides 2 and 1

Finally, the layers are flattened into a single row of nodes. In the final part of the network, there often is a succession of *fully connected layers*, i.e. layers in which every node is connected to every node in the subsequent layer. Figure 4.1 shows two fully connected layers and the output layer; only some of the connections are shown in the latter part of the graph.

When switching from image input to text input, the overall architecture of the network does not change significantly. The main difference lies in how text is converted into feature representations. Generally, each word is transformed into a fixed-length vector, for example by calculating and extracting a set of features (as done in chapter 10 and 11) or by taking as features word embedding vectors. The resulting feature representation tends to be 2-dimensional, although it can have a different dimensionality depending on how the input is transformed. Figure 4.3 illustrates a simple CNN for textual input.

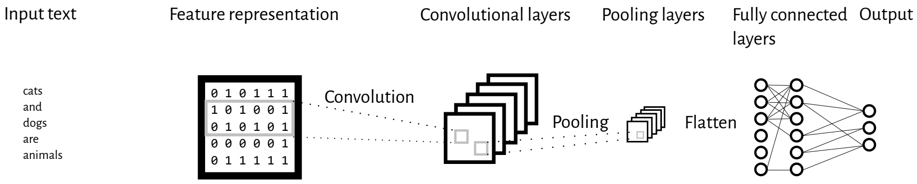


Figure 4.3: Schematic illustration of a simple CNN network for text classification

Recurrent convolutional neural networks (RCNNs) are convolutional neural networks that additionally contain *recurrent layers*. For such layers, nodes are not only connected to previous or subsequent layers but also to nodes of the same layer. The most well-known types of recurrent nodes are Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber 1997) and Gated Recurrent Unit (GRU, Cho et al. 2014) cells.

CNNs and RCNNs have been extensively used for visual tasks such as object recognition in images and video (Gu et al. 2018). However, they have also shown to be successful on textual data (e.g. Dos Santos and Gatti 2014; Cai and Xia 2015; Severyn and Moschitti 2015 for CNNs and Kalchbrenner and Blunsom 2013; Lai et al. 2015; Dong, Zhang and Yang 2017 for RCNNs).

In this thesis, we used Keras (Chollet et al. 2015) on top of TensorFlow (Abadi et al. 2015), as well as pyTorch (Paszke et al. 2017), both implemented in Python. We use these neural network models in chapter 11 for the classification of expressions into complex and non-complex classes.

As the shared task organizers made available much more data for training and evaluation than we have for Swedish, we try using neural networks, as they reach state-of-the-art performance in many tasks, superseding traditional machine learning techniques. The choice of using convolutional neural networks over other neural network architectures such as recurrent neural networks (RNNs) is grounded in the fact that they are much faster to train. In a course project, we also found that CNNs gave superior results to RNNs in the task of predicting whether a phrasal verb such as *pick up* was used literally (They *picked up* the clothes from the floor.) or figuratively (I *picked up* a little French during my holiday in Paris.).

4.4 Language models

In this thesis, we use contiguous character-based *n*-gram language models. In *n*-gram language models, an *n*-gram is a subset of the original string with length *n* (Cavnar et al. 1994). In the case of character-based *n*-gram language models, the original strings are words, and *n*-grams are sequences of *n* characters (letters). In *contiguous* character-based n-gram language models, n-grams are made up of adjacent characters (as opposed to *skip-grams*, which also include non-adjacent characters (Guthrie et al. 2006)). As an example, let us consider the word *language*. Its (contiguous) 1-grams (called *unigrams*) would be l,a,n,g,u,a,g, and e, while its 2-grams (called *bigrams*) would be la, an, ng, gu, ua, ag, and ge. The parameter n can be arbitrarily large, but generally takes values between 1 and 5. In order to capture specific sequences only occurring at the beginning or end of a word, special padding characters are often added (Cavnar et al. 1994).

In this thesis, we used both a custom language model implementation written in Java based on Alfter (2015) as well as the language model implementation from the natural language toolkit (NLTK) (Bird and Loper 2004) written in Python.

Language models are used to predict the probability of a sequence, or in other words, and for character-level n-gram models, how probable it is that a certain character follows a given sequence of length $n-1$. The probability $p(c_i|c_{i-n+1}^{i-1})$ expresses the probability that the character c_i occurs next, given the sequence c_{i-n+1}^{i-1}, with the notation c_a^b denoting the sequence of characters from c_a to c_b. Naive models simply estimate this probability as the maximum likelihood estimate p_{ML}:

$$p_{ML}(c_i|c_{i-n+1}^{i-1}) = \frac{count(c_{i-n+1}^{i})}{count(c_{i-n+1}^{i-1})} \tag{2}$$

The problem with this approach is that sequences that have never been observed in the training data will have probabilities of zero. In order to address this problem, language models typically use certain techniques to either reserve a certain amount of probability mass for unseen sequences, called smoothing, or backing down to lower-order models, or both.

The custom language models implemented in Java are implemented as trigram models and use a backup strategy called Katz's back-off. The estimate the probability p_{BO} as:

$$p_{BO}(c_i|c_{i-n+1}^{i-1}) = \begin{cases} d_{c_{i-n+1}^{i}} \dfrac{count(c_{i-n+1}^{i})}{count(c_{i-n+1}^{i-1})}, & \text{if } count(c_{i-n+1}^{i}) > 0 \\ \alpha_{c_{i-n+1}^{i-1}} p_{BO}(c_i|c_{i-n+2}^{i-1}), & \text{otherwise} \end{cases} \tag{3}$$

with $d_{c_{i-n+1}^{i}}$ being operationalized as Good-Turing discount and $\alpha_{c_{i-n+1}^{i-1}}$ the back-off weight.

The Python language models are instances of Witten-Bell interpolated models. They estimate the probability p_{WB} as:

$$p_{WB}(c_i|c_{i-n+1}^{i-1}) = \lambda p_{ML}(c_i|c_{i-n+1}^{i-1}) + (1-\lambda) p_{WB}(c_i|c_{i-n+2}^{i-1}) \tag{4}$$

with $0 \leq \lambda \leq 1$.

Language models are used in section 5.5.4 in order to distinguish C1 and C2 levels, in chapter 10 to predict the (5-class) complexity of expressions in Swedish, and in chapter 11 to predict the (binary) complexity of expressions in German, Spanish and French.

4.5 Crowdsourcing

Crowdsourcing is a technique in which a (normally unspecified) crowd is used to solve certain tasks (Brabham 2013: p. xix). Crowdsourcing is related to citizen science (Kullenberg and Kasperowski 2016). Citizen science posits that a sufficient number of non-expert judgments can be seen as equivalent to one expert judgment. This assumption is also made in crowdsourcing. In contrast to crowdsourcing, the focus in citizen science is more on the non-expert aspect; while crowdsourcing generally also makes use of a non-expert crowd, it remains more vague in its target crowd and that crowd may also include experts.

While crowdsourcing has rarely been combined with language learning, recent efforts investigate the use of crowdsourcing using language learners to create (or imitate) expert knowledge (e.g. Nicolas et al. 2020 and chapter 12 of this thesis).

In this thesis, we used pyBossa[17] by SciFabric[18] hosted at Språkbanken Text[19] in order to implement the crowdsourcing experiment. Crowdsourcing is used as main methodology in chapter 12 in order to gather judgements as to the difficulty of MWEs.

[17] `https://pybossa.com/`

[18] `https://scifabric.com/`

[19] `https://ws.spraakbanken.gu.se/ws/tools/crowd-tasking/`

5 Single-word lexical complexity

In this chapter, we investigate how the complexity of words can be characterized in terms of linguistic factors using natural language processing techniques.

5.1 Aims

The overarching research question in this thesis is concerned with how vocabulary items can be characterized in terms of lexical complexity using natural language processing techniques. In relation to this research question, we further ask when learners should be able to understand or produce an expression. The *when* is projected onto the proficiency levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001). CEFR defines six levels of proficiency, namely A1, A2, B1, B2, C1 and C2, where A1 is the beginner level and C2 is a near-native level. The question thus becomes: at what CEFR levels can learners be expected to understand or produce a certain expression?

As stated previously, the main data sources used are the word lists SVALex for receptive knowledge and SweLLex for productive vocabulary knowledge. In contrast to purely frequency-based lists, these lists contain distributions over CEFR levels, in other words they list for each word how often it was observed at different CEFR levels. Frequencies are normalized to per-million-words and further corrected for dispersion, i.e. how many different sources used the expression. If fewer sources used an expression, its frequency will be penalized rather than if more sources had used it, thus correcting for the subjective bias and idiosyncratic vocabulary use of individual textbooks or learners.

Consider for example the word *hund* ‘dog’. In the COCTAILL corpus, it has a total relative frequency of 143.[20] In SVALex, which is derived from

[20] In the examples given, the frequency values are truncated for readability purposes, meaning that digits after the decimal point are omitted.

COCTAILL, it has a total relative frequency of 140[21], and additionally has the following relative frequencies: A1 251, A2 81, B1 250, B2 74, C1 98. The per-level frequencies do not add up to the total frequency, as they are normalized on a per-level basis. In contrast, *öka* ‘to increase’ has a total relative frequency of 257 in COCTAILL and 290 in SVALex. If one were to follow the frequency-based approach, *öka* would be more frequent than *hund*, and thus *at least* not be placed at a more difficult level. However, if one looks at the per-level frequencies, one can see that *öka* was never observed at level A1, and has the following relative frequencies: A2 10, B1 416, B2 264, C1 503. Figure 5.1 shows the distributions over levels of these two words graphically using the CEFRLex search engine.[22,23] One can observe that *hund* is used much more often at early levels, while *öka* is used more at higher levels. This information is not apparent if one were to take into account only the total frequency.

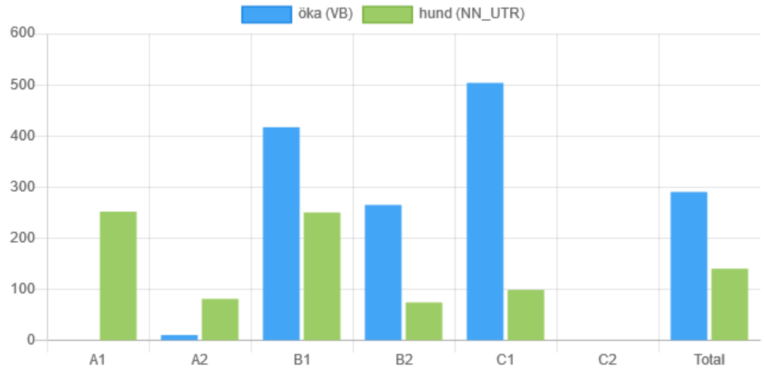


Figure 5.1: Distribution of frequencies for *hund* ‘dog’ and *öka* ‘to increase’

The first question thus is: how can this information be used to benefit learners? Ideally, one would want to link expressions to proficiency levels as has been done for English by the English Vocabulary Project or by Pearson’s Global Scale of English. However, moving from distributions to *target* proficiency levels – that is to say the level at which one expects a learner to be able to understand (if targeting receptive knowledge) or produce (if targeting productive knowledge) an expression – is not a trivial task.

[21] The numbers differ because SVALex only takes into account *reading comprehension* texts, while the version of COCTAILL available through Korp (`https://spraakbanken.gu.se/korp`) includes the whole corpus

[22] `https://cental.uclouvain.be/cefrlex/`

[23] No materials of level C2 were used in the compilation of the Swedish CEFRLex lists, thus this level can be ignored in the figure. The French CEFRLex lists contain C2 items, thus it is included in the graph.

Under the assumption that one has found a way to link expressions to CEFR levels, the next question is: is there a representation in terms of lexical complexity that allows one to derive the CEFR level from the representation? In other words, can we characterize expressions in terms of lexical complexity such that the complexity coincides with proficiency? While it is implicitly assumed that complexity and proficiency go hand in hand (e.g. Ortega 2003: p. 492; Bulté and Housen 2012: p. 21), finding a plausible and valid representation of an expression in terms of lexical complexity is – again – not a straightforward task. In continuation of the previous question, one is bound to ask next: if one *can* represent words in terms of complexity such that the CEFR level is deducible from the representation, can one predict CEFR levels for words one has not observed in the labeled corpus data?

Finally, one must ask the question of how one can check whether the assigned CEFR labels are plausible. To summarize, we will investigate the following research questions.

(1) How can CEFRLex lists be used to derive target levels for each expression?

(2) How can vocabulary be characterized in terms of complexity using NLP techniques?

(3) How can lexical complexity be predicted automatically, for unseen words?

(4) How can we check the validity of the assigned CEFR levels?

This chapter will elaborate on how we addressed research questions 1 to 4. For answering the first research question, we look at learner production, i.e. essays written by learners of Swedish at different proficiency levels. To this aim, we use the SweLL-pilot corpus (Volodina et al. 2016a) derived list SweLLex (Volodina et al. 2016b). For the second and third research questions, we take into account receptive knowledge by adding the COCTAILL corpus (Volodina et al. 2014a) derived list SVALex (François et al. 2016). Then, for answering the fourth research question, we investigate how the quality and accuracy of the projected CEFR labels can be tested using different techniques and resources.

5.2 From distributions to labels

The starting point for the work on single-word lexical complexity was the learner essay corpus derived list SweLLex (Volodina et al. 2016b). SweLLex only gives the overall frequencies per level but does not indicate at which level a word ought to be actively produced by learners. Thus, we first looked into how these frequency distributions could be mapped to single CEFR levels.

If we look at the original SweLLex frequency distribution of the lemma *heta* 'to be called', we see a peak in frequency at level B1, while the frequency at level A1 appears much lower, as illustrated in figure 5.2. However, most textbooks start with self-introductions such as *Jag **heter** David. Vad **heter** du?* 'My name is David. What is your name?'. Thus, it is to be expected that *heta* would be encountered much more often at the earlier levels in learner essays. This is, however, not reflected in the essay corpus frequency distribution.

The question thus is how we can derive target CEFR levels for SweLLex entries based on their frequency distribution across CEFR levels. Different methods are conceivable for deriving a single level from such distributions. One could take the peak (i.e. maximum) value, the first level where a word appears, the median value, the first level where a word occurs at least a given amount of times, etc.

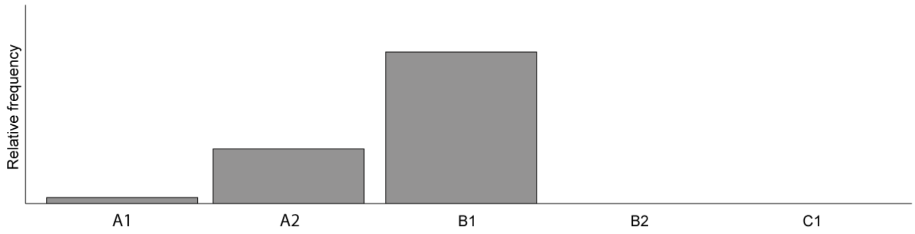


Figure 5.2: Frequency distribution over levels of *heta* 'to be called'

To this end, we calculate the *dispersion* $d(w,L)$ measure for each word w for each CEFR level L (we call this measure *diversity* in the referenced publication). Please note that the formula given in the publication is erroneous. The formula given here is corrected, and replaces d in $count(d,w,L)$ by y in order to avoid confusion between the dispersion d and the number of different learners who used w. The measure is defined as

$$d(w,L) = \frac{count(y,w,L)}{count(w,L)} \tag{5}$$

Thus, for a word w at level L, we count how often the word occurs at level L, and additionally, how many *different* learners y used the word w at level L. The intuition is that if only one learner used a certain word v x times at level L, it would be less representative of level L than if more learners used the word, even with frequency lower than x. After calculating dispersion, we normalize the SweLLex frequencies to the interval $[0-1]$, the same interval as the dispersion covers by definition. We then average both measures to arrive at a modified distribution, as shown in figure 5.3. The original CEFRLex lists already include dispersion as a weighting factor, penalizing entries that only occur in few texts per level, as opposed to entries that occur across multiple texts per level. However, we found that adding another layer of dispersion improves results.

The original SweLLex list was normalized by calculating dispersion D of a word w over K levels of difficulty, following Carroll, Davies and Richman (1971) as cited in François et al. (2016), as

$$D_{w,K} = \frac{log(\sum(p_i) - \frac{\sum(p_i)log(p_i)}{\sum(p_i)}}{log(I)} \tag{6}$$

with p_i being the probability that word w appears in subcorpus i and I being the number of subcorpora at level $k \in K$. The normalized frequency U is then defined as

$$U = \frac{1000000}{N_k} * (RFL * D + (1 - D) * f_{min}) \tag{7}$$

with N_k being the number of tokens at level k, RFL being the raw frequency by level, D the above dispersion measure and f_{min} being $1/N$ times the sum of the products of f_i and s_i with f_i being the frequency of a word in subcorpus i and s_i the number of tokens in the subcorpus.

Thus, the adjusted frequency per level is given as

$$f(w,L) = \frac{\frac{U(w,L)}{max(U(w))} + d(w,L)}{2} \tag{8}$$

with $U(w,L)$ being the normalized frequency U for word w at level L, $max(U(w))$ being the highest value of U across all levels for word w, and $d(w,L)$ being the dispersion value for w at L.

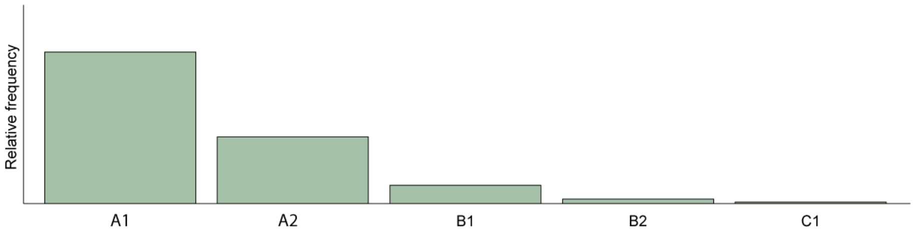


Figure 5.3: Shifted frequency distribution over levels of *heta* ‘to be called’

The modified distribution shows quite a different picture. Indeed, we see a peak at A1 with decreasing frequency over the other CEFR levels.

After having calculated the modified distribution, we apply a threshold method which we call *significant onset of use* (SOoU). The underlying assumption of this mapping technique is that if a word occurs *significantly* more frequently at a certain level than at previous levels, it can be considered to be known by a general learner of that level. Indeed, if only one learner produced a certain word form (at any level), it indicates, at best, that this specific learner knows the word, although there could be misjudgments based on spelling mistakes resulting in existing but unintended words. Learners can also, based on work or special interests, master vocabulary of a specific domain which is not known by other learners of the same level. Similarly, individual learner differences mean that in a class of B1 learners, some learners will be *more* B1, i.e. more advanced and leaning towards B2 while other learners might be leaning more towards A2. It is imaginable that more advanced learners can produce vocabulary above their currently assumed level, but this should not be taken as indicative that such productions of higher-level vocabulary constitute vocabulary of current overall class level. Thus, if a word is produced sufficiently often by different learners at a given level, it can be assumed that the word constitutes target vocabulary knowledge for that level.

On a subset of 100 manually selected items equally distributed across the CEFR levels, we find that a threshold value between 0.3 and 0.4 produces good results, meaning that a word should be considered to be included at a certain level if it occurs at least 30%-40% more often on that level than on the preceding level. Setting the threshold value to below 0.3 leads to the inclusion of higher-level items in lower levels while increasing the threshold value above 0.4 leads to the exclusion of lower-level items in the lower levels.

This method is similar to the 10-to-1 method used by the English Vocabulary Profile (Capel 2015) where they consider a word to be *criterial* of a certain level if it does not occur at least ten times as often on the following level (Hawkins and Filipović 2012: p. 38). For example, if a word *A* occurs once at level A1 and 12 times at level A2, it will be considered an A2 word, as $1 * 10 < 12$. If a word *B* occurs twice at level C1 and 17 times at level C2, the word will be considered a C1 word, as $2 * 10 > 17$.

It should be noted that the threshold values for our method and the EVP method are quite different. This is because the EVP method works with raw frequencies while our method uses normalized frequencies that fall in the range [0-1].

In later works, we drop the diversity calculation and instead we adopt the *first occurrence* (FO) approach, as has been done in previous work on French (Gala, François and Fairon 2013; Gala et al. 2014). This approach takes the first level at which a word occurs as the target level for that word. This approach is both easier to calculate and produces the same result in most cases.

However, this method is problematic if there is only little evidence at a certain level, and it does not allow to distinguish between central and peripheral vocabulary.

5.3 Features for complexity

After having derived target levels for each entry in the resource, we can devise a characterization scheme for lemmata which would allow us to characterize lemmata in terms of complexity.

In contrast to text-based complexity measures, i.e. complexity measures based on whole texts, single words offer a rather small set of directly observable and derivable measures. There is for instance no syntactic information, nor context information if one looks at words in isolation. In line with other research on word complexity, we used a variety of features including count-based features such as word length, number of syllables, suffix length, compound count, grammatical features such as part-of-speech and gender, semantic features such as polysemy and homonymy, pragmatic/topical features, psycholinguistic features from the MRC psycholinguistic database (Wilson 1988) (for example age of acquisition, concreteness), and presence/absence of most frequent n-grams based on Wikipedia data. Besides the explicitly named resources, information mainly comes from Saldo (Borin, Forsberg and Lönngren 2013), Swesaurus (Borin and Forsberg 2014) and the COCTAILL corpus (Volodina et al. 2014a). We set up a pipeline that automatically extracts and/or calculates these measures for any input word.

Table 5.1 lists all the features that we extracted. Of the count features, length is the length of the expression in characters, syllable count is the number of syllables, the next three features are binary flags indicating whether the expression contains non-alphanumeric characters (characters outside of the alphabet and numbers), and whether the expression consists of more than one word. For character bigrams, we calculate all bigrams and retain only the 53 most predictive bigrams. This feature indicates whether the expression contains any of these bigrams. n-gram probabilities are calculated using uni-, bi- and trigram language models trained on Wikipedia. For morphological features, part-of-speech is the part-of-speech of the expression as given by the parser, suffix length indicates the length of the expression after removing the stem, compound count counts the number of possible compounding analyses. For compounds, we calculated all compounding elements and retained only the 12 most predictive elements. This feature indicates whether the expression contains any of these elements. Gender is the grammatical gender of the word. For semantic features, the degrees of polysemy and homonymy are calculated as the number of sub-entries respectively the number of distinct entries in Lexin for a given word. Finally, topic distributions is a vector indicating all topics under which a word occurred as calculated from the topic-annotated COCTAILL corpus. Using recursive feature ablation (Guyon et al. 2002) we found that three of the features did not contribute to the classification, namely character bigrams, compound count and compounds.

Count features
Length (number of characters)
Syllable count
Contains non-alphanumeric character
Contains number
Is MWE
Character bigrams
n-gram probabilities
Morphological features
Part-of-speech
Suffix length
Compound count
Compounds
Gender
Semantic features
Degree of polysemy
Degree of homonymy
Context features
Topic distributions

Table 5.1: Overview of features for single-word lexical complexity

5.4 Automatic prediction of lexical complexity

Based on the feature representation of entries, along with the derived target CEFR level, we train different machine-learning algorithms in order to find a system that can learn which features are important in the classification of words into CEFR levels and that can, based on a feature representation of a word in the chosen feature space, predict the level for said word. To this aim, we use the WEKA toolkit (Hall et al. 2009) and the scikit-learn Python library (Pedregosa et al. 2011). We use different algorithms, namely Support Vector Machines (SVM), MultiLayer Perceptron (MLP), Random Forests (RF) and ExtraTrees (ET). Our best classifier (ET) achieves an accuracy of 59% in a 10-fold cross-validation setting. These results are in line with previous research on word-level complexity prediction for L2 French (Gala et al. 2014).

However, in the experiment, word *senses* are conflated, i.e. there is no distinction between different word senses. This means that polysemous words are assigned a single level; this is not a realistic scenario, as in most cases, one learns different senses of a word at different levels instead of learning them all at one. As has been done in the EVP and GSE projects for English, we will have to address this issue and not only create a sense based vocabulary list but also build a classifier that is able to differentiate between senses. While the former is addressed through the creation of SenSVALex, the latter may prove problematic. Having the same input material as the original word lists but further distinguishing between senses leads to more sparsity in the data, as single entries will be split up into multiple entries, and so will their frequencies. Further, the correct sense of a polysemous word is often derived from its context; words in isolation provide little to no information as to their intended sense. Yet, the presented pipeline is set up to extract word-based features in isolation. We have experimented with including context in the form of word embeddings in the Complex Word Identification 2018 shared task as described in chapter 11. However, we found these features to be irrelevant to the classification. This finding was corroborated by other participants in the shared task (De Hertog and Tack 2018); the winning system did not include any context features (Gooding and Kochmar 2018).

In our experiments, we find that one of the most important predictors in automatic prediction of complexity is topic distribution; without this feature, classification is around baseline level (i.e. majority class classification). The reason behind this might be that we take CEFR textbooks and project the levels of texts onto words in order to derive our reference data which the algorithm is supposed to learn. As the CEFR is based on *can do* statements such as “Can produce simple phrases and sentences about themselves [. . .]” (Council of Europe 2020: p. 67) for level A1, and because these statements are topical in nature (i.e. family, work, personal interests, . . .), topic distributions may capture the progression of the CEFR. Further, topic annotation for textbooks was done manually. In chapter 11 we automatically extract topic lists from Wikipedia articles, albeit in a supervised manner by specifying broad topics such as “animals”, “nature”, “science”, ... We surmise that using unsupervised topic model induction might be useful in a scenario where one has no pre-annotated topics and no prior knowledge as to the topics present in the data.

5.5 Evaluation

In order to corroborate the results from our mapping techniques, we have compared the output from said technique with the outputs from other techniques, as described in the following.

5.5.1 Significant onset of use versus first occurrence

We explored how the “significant onset of use” method compares to the “first occurrence” approach using six different CEFRLex resources (SVALex and SweLLex for Swedish, FLELex TT and FLELex CRF for French, NT2Lex for Dutch and EFLLex for English). This evaluation is an extended version of the evaluation presented in chapter 10, where we only compared SVALex and SweLLex. We found that both “significant onset of use” and “first occurrence” agree to a large extent and that most disagreements lie within one CEFR level, except for EFLLex, which shows slightly more disagreements of more than one level, and the French resources, which show a marked difference. Table 5.2 shows the results for the resources we used, as well as how often both mapping methods agreed (column ‘same’, i.e. assigned the same level by both methods), how many of the disagreements were within one level of difference (e.g. one method saying B1 and the other B2; column ‘W1L’, i.e. within-one-level), and how many disagreements were more than one level apart (e.g. one method saying B1 and the other A1, column ‘M1L’, i.e. more than one level apart). The table also shows the percentage of same-level agreement (column ‘Same %’) and the percentage of agreement if one counts within-one-level of difference as being in agreement (column ‘W1L %’).

Since the CEFR levels are regarded as a continuous scale rather than strictly separate levels (Council of Europe 2018: p. 34) (i.e. a word *X* may be *more centrally* A1 while a word *Y* might be A1 but leaning towards A2), taking within-one-level differences into account is a sensible choice.

As can be seen from the table, both mapping methods agree to around 80% in exact level matching, while they agree around 90% of the time when allowing for differences of one CEFR level. We thus surmise that the choice of mapping method should have little influence on downstream tasks. This has also been found in Gala et al. (2014) where they tried different mapping techniques and found that FO performs as well if not better than more elaborate techniques.

Resource	Same	W1L	M1L	Same %	W1L %
SVALex	12775	1592	1255	81.78	91.97
SweLLex	5689	706	516	82.32	92.53
FLELex CRF	11464	1748	4653	64.17	73.96
FLELex TT	8540	1618	4078	60.00	71.35
NT2Lex	9523	1388	1171	78.82	90.31
EFLLex	11914	1556	1664	78.72	89.01

Table 5.2: Agreement between mapping methods on different resources

We have repeated the experiment taking only into account words that occurred at least five times.[24] While this leads to the exclusion of a large number of words (75% in EFLLex up to 90% in SweLLex, around 82% for the other resources), the trend remains the same; both SOoU and FO agree on the level for the majority of words, while disagreements tend to be within one level of difference. During this experiment, we also found that the number of hapax legomena is quite high across the tested CEFRLex resources, with 48% of all entries in EFLLex being hapax legomena, 53% in SVALex, and 55% in NT2Lex, and 73% in SweLLex.[25]

5.5.2 Semantic space

We also explored using custom word embedding models to test how well our mapping compares to the output of a different technique.

Word embedding models have been found to capture semantic similarities such as synonyms and antonyms by learning which items typically occur in similar contexts. The concept of word embeddings has been proposed by Bengio et al. (2003), although they became widely used only when Mikolov and Dean (2013) presented an efficient algorithm for word embedding calculation known as Word2Vec. In word embeddings, words are represented in a continuous semantic space which allows for certain calculations such as distance, i.e. how far is word X, semantically speaking, from word Y. Another often cited example is $V(king) - V(man) + V(woman) \approx V(queen)$, where $V(X)$ stands for the word embedding vector for word X, meaning that if one takes the word embedding for *king*, subtracts from it the word embedding for *man* and adds

[24] Since the French resources do not contain raw frequency counts, they were excluded from this experiment.

[25] Since SweLLex is based on learner essays, many hapax legomena may be due to unique misspellings.

the word embedding for woman, one gets a word embedding that is close to that of *queen*.

We explore whether we can train word embeddings with CEFR labels in such as way that the model learns which words tend to occur at which levels instead of learning which words occur in similar contexts; alternatively, one could regard the CEFR labels as context in this case. This deviates from the standard word embedding model, as a traditional word embedding model would put semantically related concepts close together regardless of CEFR levels. In order to circumvent this problem, we train a word embedding space with CEFR anchors, called the *indexed* method. For each text, we take the overall CEFR level of the text and add this level after each word in the text. Thus, a sample text *X Y Z* of level *A2* would become *X A2 Y A2 Z A2*. Our hypothesis is that with this method, words occurring frequently at A2 will be closer to the A2 label. In effect, the CEFR labels are treated as regular words with which other words co-occur.

We have found that the custom embedding model clearly outperforms a traditional word embedding model and that using a large window ($w = 60$) gives the best results. Additionally, we can identify core and peripheral vocabulary. The following paragraph is taken verbatim from the relevant publication. For example, both our frequency list and our best performing semantic space label *resa* 'to travel' as an A2 word. From the semantic space, we can also see that it is much closer to B1 than to A1 – we can suppose that it is a rather "advanced" word that tends to lie between A2 and B1. In the same way, *fredag* 'Friday', labeled as A2 by the frequency lists, clusters in our space both with A2 and (less closely) A1 lemmas, showing that it is likely to be a term on the "easy" spectrum of the A2 vocabulary.

A further logical step would be to use sub-word embedding models (Bojanowski et al. 2017). The main advantage of such models is that they can deal with out-of-vocabulary words.

5.5.3 Multilingual aligned comparison

In chapter 14 we look at the Swedish, English and French CEFRLex lists and align them via a parallel multilingual corpus. The aim of the study is two-fold: first, we explore how the automatically derived CEFR levels in the English CEFRLex compare to more established standards, namely the CEFR levels as given by the English Vocabulary Profile (EVP) and the Global Scale of English (GSE). Second, we investigate how the Swedish, French and English CEFRLex lists agree with regard to the automatically derived CEFR levels by aligning these three resources via a multilingual parallel corpus. We also use the Swedish and English Kelly lists as sources of information; there is no French Kelly component.

Concerning the first part of the study, we find a good correlation between the English CEFRLex levels and the combined score from EVP and GSE. This further validates the methodology of automatically deriving CEFR levels by taking the level of first occurrence. Outliers may be due to various reasons, some of which include data sparsity (i.e. if one had more data, results might be different), textbook writers' subjectivity (i.e. one might not encounter a certain word until later levels because it was not introduced earlier; this does not mean that the word couldn't have been understood at an earlier level), or simply because some of the words are more peripheral to a given level.

Concerning the second part of the study, we end up with a combined list of about 6000 entries, of which almost 3000 entries are available in all three languages. Of these, about 400 entries have the same CEFR level in all three languages, and about 2000 have the same CEFR level in two of three languages. Concerning the distinction between core and peripheral vocabulary, one could assume that entries that have the same level in all three languages are more *central* to that level than entries that have the same level in only two of the three languages or that have different levels altogether.

5.5.4 Dealing with C2 and above

The described model for predicting the complexity of single words has been trained on data from graded textbooks where the textbook levels ranged from A1 to C1. The model is thus only capable of producing predictions that fall within this range. However, there are words that are beyond the level of C1, and even words that are beyond C2, i.e. words that cannot be expected to be understood by any learner even at near-native proficiency. When given such words, the model would at most be able to predict level C1, the highest level it has been trained on.

In order to give more accurate predictions, it would be desirable to also be able to assign levels C2 and beyond. This problem can be tackled in various ways. First, one could retrain the model with additional data from C2 level if one's aim was to be able to distinguish between the two highest levels. Unfortunately, the amount of data available for level C2 is rather scarce. The Swedish KELLY list (Volodina and Kokkinakis 2012) contains 1405 entries that are assigned level C2, while the COCTAILL textbook corpus contains no textbooks of level C2. Furthermore, introducing more data, and especially another output class to be predicted, will most probably lead to the inclusion of more noise and by extension a decrease in accuracy.

Second, if one wishes to further distinguish levels C1, C2 and *beyond C2*, or simply C1 and *C2 or beyond*, another approach would be needed, as even retraining the model with data from C2 would not allow the model to predict *beyond C2*.

The experiment described hereafter is not part of any publication included in the thesis but extends the work described in the publications. In order to address the problem of dealing with levels beyond C1 we have chosen to train character-based n-gram language models on COCTAILL data. First, we used all textbook reading texts in order to obtain a general textbook language model. Then, we trained separate language models on each of the five represented levels A1, A2, B1, B2 and C1. For predictions, we test a word against all five models and take as prediction the level of the model with the highest probability.

We tested 1-, 2-, 3-, 4- and 5-gram models, but the first three performed at chance level, while the 4-gram model managed to increase accuracy to about 50% in a five-class classification. The 5-gram model was able to reach around 65% accuracy in 10-fold cross-validation, beating the highest score (59%) of the more elaborate feature-engineered models presented in chapter 10. Using feature selection, we found that the general COCTAILL language model was not able to learn to distinguish between proficiency levels in any way, while the separate level-specific language models were predictive in all test cases. Thus, one could say that if all level-specific language models output very low probabilities, the word in question might be above C1. Thus, while this model is only able to differentiate between the represented levels and a generic category *above C1*, it might yield interesting information nonetheless. This is a first exploratory study on how to deal with C2 and its implementation into an automatic level prediction framework remains future work.

5.6 The usefulness of n-gram language models

In chapter 11, we find that character-based n-gram language models perform nearly as well as fully feature-engineered systems in the classification of words into complex and non-complex words. In contrast to this, in chapter 10, we find that using n-gram language models yields poor results. This seems to be an apparent contradiction as to the usefulness of n-grams as features in complexity prediction. However, there are some key differences that explain this apparent contradiction.

First of all, the target output of the two systems is different. The system presented in chapter 10 predicts five classes (A1 to C1) while the system presented in chapter 11 predicts two classes (complex and non-complex).

Further, the system in chapter 10 uses other features besides n-grams; these features might introduce noise that lower the predictive power of n-grams.

More importantly, there is a difference in genre: the system in chapter 10 uses language models trained on Wikipedia to predict the complexity of words from textbook corpora. On the other hand, the system in chapter 11 uses language models trained on Wikipedia to predict the complexity of words from Wikipedia and WikiNews as two of the three English strands.

Finally, and most importantly, the system predictions are compared to different references: the system in chapter 10 is measured against automatically derived CEFR levels, while the system in chapter 11 is compared against human annotated complexity labels.

5.7 Notes on the complex word identification shared task

Results from the Complex Word Identification Shared Task 2018 show that our system is able to compete with other systems, despite the fact that we rely on ‘traditional’ machine learning. Our best ranks for each sub-task are summarized in table 5.3. Besides the task and our best rank in each sub-task (‘Best rank’), table 5.3 also lists the number of systems submitted for each task (‘# submitted systems’; each participating team could submit more than one system per task, and they could choose to only participate in certain of the sub-tasks, thus the numbers are not equal), our best F_1 score (‘Our F_1’) and the winning team’s F_1 score (‘Best F_1’). F_1 is a common form of the F_β score with $\beta = 1$. F_β coalesces and weights precision (P) and recall (R) according to:

$$F_\beta = (1+\beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R} \tag{9}$$

While the results for English seem modest, they are competitive when considering not only the rank but also the total number of competing systems and the difference in F1 scores. For the non-English tasks, our results seem more convincing, although the number of participating systems was also lower.

For French, no training data was provided, as the task organizers wanted to see how well cross-lingual complex word identification would work. We tested two different setups for French: (1) using English, Spanish and German n-gram models to predict the complexity and use majority voting, and (2) using only the Spanish n-gram language model to predict the complexity. We found that using setup (2) leads to better results. This may be due to the smaller phylogenetic distance between Spanish and French; in turn, this may indicate that one could use high-resource languages to predict the complexity of related low-resource languages.

Task	Best rank	# submitted systems	Our F_1	Best F_1
English News	16	39	0.8329	0.8736
English WikiNews	15	34	0.8031	0.8400
English Wikipedia	7	36	0.7832	0.8115
German	2	14	0.7427	0.7451
Spanish	7	17	0.7281	0.7699
French	3	9	0.6266	0.7595

Table 5.3: Best ranks in the CWI shared task

The winning system for English uses a setup similar to our own, using feature extraction and ‘traditional’ machine learning. One major difference is that they add CEFR values extracted from the Cambridge Advanced Learner Dictionary (CALD). They are the only team to have used this information. While our system uses CEFR levels from EFLLex, we surmise that CALD might be more accurate; however, access to CALD is restricted.

Further, while we experimented with context features, they disregarded context completely; while they acknowledge that this is a major source of errors (“88.94% of the mis-classified words in the NEWS test set have multiple labels in the data” (Gooding and Kochmar 2018: p. 190)), their system still manages to reach rank 1 in all English sub-tasks. This is in line with our experiments which showed that context features performed poorly.

Finally, we experimented with CNN and RCNN architectures and while the scores achieved by these neural networks were promising, they did not manage to surpass the scores reached by ‘traditional’ machine learning. Again, this is in line with the overall results from the shared task, which concludes that “traditional feature engineering-based approaches [. . .] perform better than neural network and word embedding-based approaches”, but also that “more systems employed deep learning approaches and the results are getting better for the CWI task” (Yimam et al. 2018: p. 10).

6 Multi-word lexical complexity

In this chapter, we look at multi-word expressions. First, we check the accuracy of the automatic MWE recognition. Then, we explore compositionality as annotated manually by two annotators. Last, we conduct a crowdsourcing experiment to see how learners, teachers and assessors agree with one another when it comes to the difficulty of MWEs.

6.1 Multi-word expressions versus single-word expressions

As most of the features extracted for single-word complexity rely on the lemma as pivot, the exact same technique is not directly applicable to MWEs. Chapter 11 describes a two-pass approach for English that first treats all single-word expressions, then treats multi-word expressions as a function of its constituents. Features such as word length were regarded as additive (the length of a multi-word expression would be the sum of the length of its constituents) while other features such as word embeddings were calculated as the average of its constituents. While this technique has been applied to the treatment of multi-word expressions for English, it is a rather crude technique and it presupposes a certain degree of compositionality; there is an inherent assumption that information about the constituents of an MWE allows one to know features about the MWE. While this may be true of a certain subset of MWEs, it cannot be accepted as a general truth. Besides idiomatic expressions which are by definition non-compositional, other multi-word expressions such as particle verbs also tend to have non-compositional readings. For example, *ge upp* 'give up' has little to do with the separate meanings of 'give' and 'up'.[26]

The word lists used to create the single-word lexical complexity estimator *did* include some MWEs, and they were treated just like single-word expressions in the pipeline (unlike the two-pass approach used for English). This is due to the fact that they were identified as MWEs *because* they were listed

[26] In addition, *ge upp* is polysemous.

as MWEs in the Saldo thesaurus. This, however, does not allow for the treatment of unseen MWEs which are not in Saldo or any other resource we have. SiWoCo, the automatic prediction pipeline for **Si**ngle **Wo**rd **Co**mplexity, can deal with any input, however nonsensical it might be. This means that it can theoretically deal with MWEs as well. However, SiWoCo consistently overestimates the level of MWEs, as shown in table 6.1. The first column lists the expression under scrutiny, followed by its English translation. The next two columns indicate the automatically predicted level by SiWoCo for receptive and productive knowledge respectively. The final two columns indicate the level as determined by the first-occurrence approach in the receptively oriented SVALex and the productive knowledge derived SweLLex, or *n/a* if this expression is not part of the list. Thus, the first line should be interpreted as: the expression *kort sagt* 'in brief', literally 'shortly said', is expected to be receptively known at level C1 and productively used at level C1; in SVALex it was encountered for the first time at level A2 while it was first used in an essay at level C1.

Expression	Receptive	Productive	SVALex	SweLLex
kort sagt 'in brief'	C1	C1	A2	C1
just det 'exactly'	B2	B2	A1	C1
ingen fara 'no worries'	C1	C1	C1	A2
god natt 'good night'	C1	C1	A1	n/a
god morgon 'good morning'	C1	C1	A1	n/a
pommes frites 'french fries'	B2	C1	B2	n/a
på pin kiv 'for spite'	C1	C1	C1	n/a

Table 6.1: Target level estimation of MWEs in SiWoCo

While the analysis looks valid when querying more complex MWEs such as shown in the last line for the expression *på pin kiv* 'for spite', the consistent overestimation becomes apparent when querying easier expressions such as *god morgon* 'good morning', *god natt* 'good night', *just det* 'exactly', literally 'just that', or *ingen fara* 'no worries'. It should be noted that prediction of levels is done separately for receptive and productive knowledge; the predictor for receptive knowledge is only trained on SVALex data while the predictor for productive knowledge is only trained on SweLLex data. Thus, one cannot deduce from table 6.1 that for example C1 under *receptive* for *kort sagt* is plausible, given that it was seen at C1 in SweLLex.

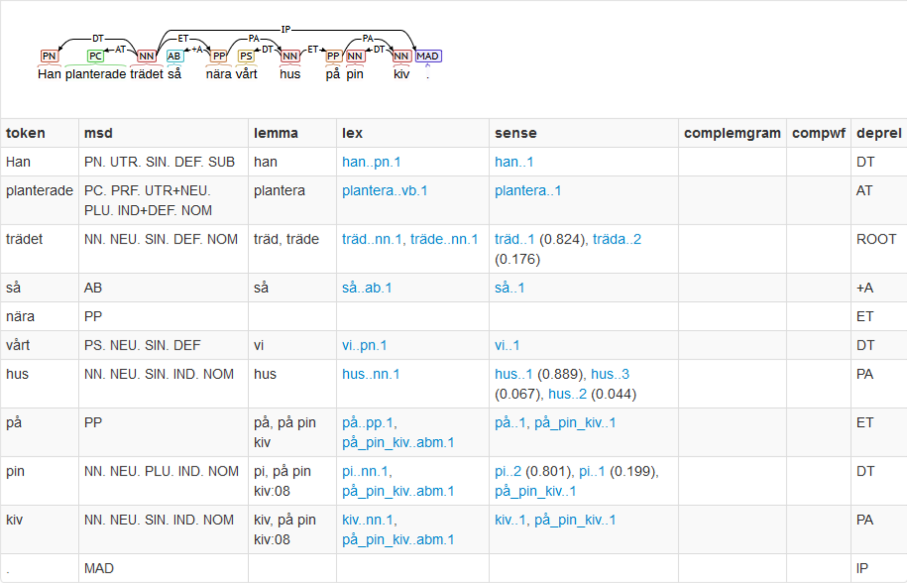
Training classifiers on such small amounts of data (1444 MWEs) is feasible, although the results would have to be taken with a grain of salt. Alternatively, one could train different models such as n-gram language models to predict the target level of MWEs.

6.2 Checking automatic multi-word expression recognition

MWEs are automatically identified by the Sparv pipeline used to annotate the texts. However, there are some issues related to the automatic identification. Sparv does not try to identify the correct analysis among multiple possible analyses ; it returns all possible analyses. For example, *på pin kiv* 'for spite' it analyzed as each of its constituents and as a whole expression. In case Sparv detects components of an MWE, it links each component to the first component detected. Figure 6.1 shows the Sparv analysis of an example sentence (Han planterade trädet så nära vårt hus på pin kiv. 'He planted the tree so close to our house for spite.') with the columns 'lemma' and 'lex' showing both the constituents and the MWE analysis. Further, the 'lemma' column indicates that 'pin' and 'kiv' are linked to word number 8 (på pin kiv:08), which is 'på', the first component of the recognized MWE. It can also be noted that the analysis for 'pin' is wrong; it is recognized as 'pi', having two possible senses (pi..1 and pi..2) and the word sense disambiguation assigned the sense pi..2 a high probability. In Saldo, the sense pi..1 corresponds to the mathematical constant π, while pi..2 corresponds to the Greek letter π; neither of these are correct in this case, as 'pin' in this MWE is a an adjective meaning 'very, to the highest degree' (Svenska Ordbok). Saldo does have an entry for the adjective 'pin' in this sense, but Sparv did not assign this sense in this case.

However, not every occurrence of constituents of an MWE occurring in a sentence means that they form the MWE. For example, in the sentence *Jag har bara ett fel.* 'I only have one shortcoming', *ha fel* 'to be wrong' is wrongly detected as MWE.

In order to check the automatic annotation reliability, we automatically selected 30 texts, three of each level, 15 each from COCTAILL and SweLL-pilot, for manual inspection. The essays selected from SweLL-pilot were manually normalized and analyzed both in original and normalized form, bringing the effective total of texts up to 45 (15 from COCTAILL, 15 from SweLL-pilot (original) and [the same] 15 from SweLL-pilot after normalization). The texts were selected to be representative of the level in question with regard to number of words in the sentence, average number of sentences in the text and average number of MWEs in the text. If the average number of MWEs for a level was below 1, it was set to 1, forcing each selected text to at least contain one MWE. Further, we tried to sample texts covering different topics by excluding topics that had already been encountered during the selection process. The presented results are currently not yet published.



token	msd	lemma	lex	sense	complemgram	compwf	deprel
Han	PN. UTR. SIN. DEF. SUB	han	han..pn.1	han..1			DT
planterade	PC. PRF. UTR+NEU. PLU. IND+DEF. NOM	plantera	plantera..vb.1	plantera..1			AT
trädet	NN. NEU. SIN. DEF. NOM	träd, träde	träd..nn.1, träde..nn.1	träd..1 (0.824), träda..2 (0.176)			ROOT
så	AB	så	så..ab.1	så..1			+A
nära	PP						ET
vårt	PS. NEU. SIN. DEF	vi	vi..pn.1	vi..1			DT
hus	NN. NEU. SIN. IND. NOM	hus	hus..nn.1	hus..1 (0.889), hus..3 (0.067), hus..2 (0.044)			PA
på	PP	på, på pin kiv	på..pp.1, på_pin_kiv..abm.1	på..1, på_pin_kiv..1			ET
pin	NN. NEU. PLU. IND. NOM	pi, på pin kiv:08	pi..nn.1, på_pin_kiv..abm.1	pi..2 (0.801), pi..1 (0.199), på_pin_kiv..1			DT
kiv	NN. NEU. SIN. IND. NOM	kiv, på pin kiv:08	kiv..nn.1, på_pin_kiv..abm.1	kiv..1, på_pin_kiv..1			PA
.	MAD						IP

Figure 6.1: Example parse

An assistant in the L2 profiles project[27] manually inspected each of the texts with regard to the automatic annotation, covering not only MWEs but also lemmatization, part-of-speech tags and word senses. Table 6.2 shows an overview of the findings, listing the total number of MWEs inspected per resource, as well as a breakdown over the different CEFR levels. It also indicates precision and recall. The columns in table 6.2 read as follows:

- Resource: lists the resource under scrutiny. The row with the name of the resource contains the total while the following rows break up the total into the different CEFR levels
- Identified: how many MWEs were identified in total
- Correct: how many MWEs that were identified were indeed MWEs
- Partial: how many MWEs that were identified were only partially identified
- Incorrect: how many MWEs that were identified were wrongly identified as MWE

[27] `https://spraakbanken.gu.se/en/projects/l2profiles`

- Missed: how many MWEs were missed by the automatic MWE recognition

- Precision: see below

- Recall: see below

Precision P is measured as

$$P = \frac{tp}{tp + fp} \tag{10}$$

where *tp* stands for *true positive* and corresponds to the *Correct* column in table 6.2, and *fp* stands for *false positive* and corresponds to the sum of columns *Partial* and *Incorrect*. We have opted to include partial MWE identifications under incorrectly identified MWEs, although this is open to debate. Precision thus expresses how many of all identified MWEs were indeed MWEs.

Similarly, recall R is defined as

$$R = \frac{tp}{tp + fn} \tag{11}$$

where *fn* stands for *false negative* and corresponds to column *Missed*. Recall expresses how many MWEs of all MWEs present in a text were recognized. Precision and recall have a range between 0 and 1, with 0 meaning that either none of the identified MWEs were actually MWEs (zero precision) or that none of the MWEs in a text were recognized as MWE (zero recall). A value of 1 corresponds to a perfect score, i.e. all identified MWEs were truly MWEs (perfect precision) although this does not mean that *all* MWEs in a text were identified; or that all MWEs in a text were identified (perfect recall) although the analysis possibly also includes non-MWEs that were recognized as MWEs. We also indicate the $F1$ score which is calculated as the harmonic mean between precision and recall:

$$F1 = 2 * \frac{P * R}{P + R} \tag{12}$$

As can be gathered from the table, on the overall level, precision and recall are in acceptable ranges (0.80 - 0.90 for precision and 0.71 - 0.78 for recall). If one breaks up the analysis into the different CEFR levels it becomes apparent that there is more variation. For COCTAILL, precision is high for the A and B levels while it drops quite some for C1. Recall on the other hand is moderately high for the A levels, slightly lower for the B levels and slightly lower again for C1. For SweLL-pilot original, precision is high across all levels, but recall is low for the A levels, moderately high for the B levels and acceptable for

Resource	Identified	Correct	Partial	Incorrect	Missed	Precision	Recall	F1
COCTAILL	60	48	3	9	20	0.80	0.71	0.75
A1	7	6	0	1	2	0.86	0.75	0.80
A2	7	6	0	1	1	0.86	0.86	0.86
B1	19	16	1	2	8	0.84	0.67	0.74
B2	16	14	1	4	6	0.88	0.70	0.79
C1	11	6	1	4	3	0.55	0.66	0.60
SweLL-pilot (original)	82	74	6	2	29	0.90	0.72	0.80
A1	6	6	0	0	5	1.00	0.55	0.71
A2	4	4	0	0	4	1.00	0.50	0.66
B1	14	13	0	1	6	0.93	0.68	0.78
B2	22	19	2	1	5	0.86	0.78	0.82
C1	36	32	4	0	9	0.89	0.72	0.80
SweLL-pilot (normalized)	99	84	13	2	24	0.85	0.78	0.81
A1	9	7	2	0	5	0.78	0.58	0.67
A2	7	6	1	0	3	0.86	0.67	0.75
B1	23	21	2	0	2	0.91	0.91	0.91
B2	34	29	4	1	6	0.85	0.83	0.84
C1	26	21	4	1	8	0.81	0.72	0.76

Table 6.2: Number of correctly identified MWEs including precision, recall and F1 score

C1. For SweLL-pilot normalized, one can observe a general drop in precision across all levels in comparison to the non-normalized version, while at the same time, recall increases. Indeed, one can observe that normalizing essays leads to a higher number of identified MWEs, but also a higher number of incorrectly identified MWEs, thus decreasing precision while increasing recall.

An analysis of the partially identified expressions from COCTAILL texts shows that of the 3 partially identified expressions, two are *få reda* 'find out' and *ha reda* '(actively) find out', which miss the particle *på* according to the assistant. However, in this case, *på* is a preposition and not a particle according to Saldo's principles. The third one is tricky because it consists of overlapping MWEs which are considered one MWE but could have different readings: *gå långt in i* 'go far into' consisting of *gå långt* 'go a long way', *gå in* 'go in' and *in i* 'into'.

6.3 Compositionality

As an experimental feature, we had two assistants from the L2 profiles project annotate a total of 1498 multi-word expressions[28] for *compositionality*, i.e. how transparently the meaning of an MWE can be derived from its constituents' meanings. For example, *samla ihop* 'to collect, gather', literally 'collect together' is rather compositional in its meaning, while *på håret* 'close call', literally 'on the hair' is rather non-compositional. Compositionality was indicated on a scale from 0 (totally compositional) to 100 (totally opaque) in increments of 1.[29] Of the 1498 MWEs, 710 were annotated by both annotators. Of these 710 MWEs, there were 662 disagreements in compositionality. This is to be expected with such a fine-grained scale. However, even when projecting the fine-grained compositionality scores onto a coarser 10-point scale, with the bins being 0-9, 10-19, and so forth, there are still 559 disagreements. Even on a binary scale, 184 disagreements remain. This shows how subjective compositionality judgments can be.

Another approach would be to calculate the relative span difference. Let c_1 and c_2 be two compositionality judgments for the same item by two different annotators. Let r be the relative span size. Let a be the adjusted agreement with $a = |c_1 - c_2| \leq r$. At $r = 0$, compositionality scores of 9 and 10 would

[28] The number of MWEs is different from the number mentioned in section 6.1 because the experiments in section 6.1 were done on the old version of SVALex while the experiments in this section were done on the new sense-based version of SVALex.

[29] After the initial trial run with compositionality annotation, the project team noticed that the concept of compositionality is not as straight-forward as initially thought and that there is a distinction between compositionality and transparency (Nunberg, Sag and Wasow 1994: pp. 495,498).

be counted as different, whereas at $r = 1$, they would be counted as equal. In contrast, a linear projection into 10 bins would result in those items landing in different bins (and thus count as disagreement).

Figure 6.2 shows the agreement of the MWEs where the annotators did not agree as a function of the relative span size r. If one allows a deviation of 16 points ($r = 16$) between the annotator values, agreement rises above 50%, meaning that for each item where they disagreed, over half of those items become “agreements”.

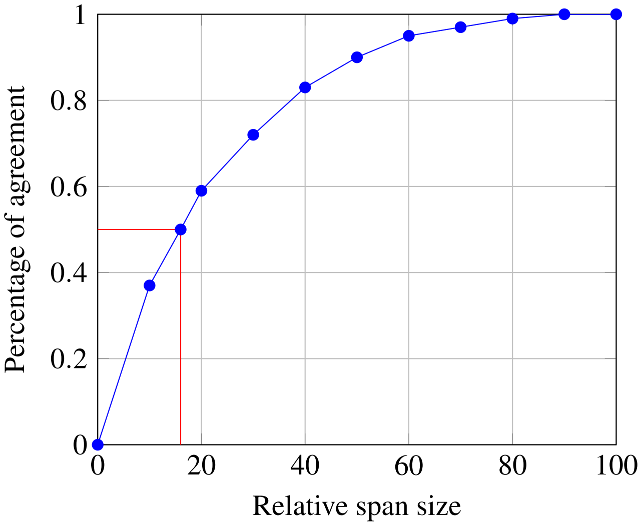


Figure 6.2: Agreement in disagreement

Due to the high subjectivity of these judgments, the low inter-rater agreement, and the conflation of transparency and compositionality in this measure, we decided not to use this feature in subsequent analyses.

6.4 Experts versus non-experts

In chapter 12 we explore whether learner intuitions about the difficulty of multi-word expressions agree with expert judgments. To test this, we set up crowdsourcing experiments.

In the experiment, we test three different groups of crowds, namely learners of Swedish (non-experts) teachers of Swedish as a second or foreign language, and CEFR assessors, i.e. teachers and assessors of Swedish as second or foreign language who actively use the CEFR. It should be noted that we later conflated the two latter groups into one group called “experts”. This choice was made in order not to mis-represent teachers as "non-experts", and to further emphasize the dichotomy between learners as non-experts and experts, i.e. teachers and assessors. For each of the three groups we set up three experiments with a selection of 60 MWEs per experiment. The three experiments are:

- Group 1: Interjections, fixed expressions and idioms
- Group 2: Verbal MWEs
- Group 3: Adverbial and adjectival MWEs

We found a very high correlation between all tested groups, suggesting that learner intuitions (internal development) correlates with expert judgments (external assessment), as illustrated in table 6.3. In this table, we indicate the Spearman rank correlation coefficient, which can range from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with a value of 0 indicating no correlation.

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
L2 speakers-L2 professionals	0.9509	0.9282	0.9203
L2 speakers-CEFR experts	0.9333	0.8115	0.8370
L2 professionals-CEFR experts	0.9386	0.8495	0.8579

Table 6.3: Agreement between voter groups in the crowdsourcing experiment

The study in chapter 12 focuses on the relationship between expert and non-expert results from the crowdsourcing experiment itself, and only tangentially mentions the coursebook-derived CEFR levels. However, we also analyzed the results with regard to the coursebook projected levels. These complementary results are presented hereafter.

CEFR levels projected from COCTAILL are nominal in nature, e.g. we have chosen 12 items of level A1, 12 items of level A2, etc., but there is no inherent ordering within the levels. Thus, one cannot say that item A of level A1 should come before item B of level A1. In contrast, linear scales derived from the crowdsourcing experiment are ordinal in nature, i.e. we can say that according to the crowdsourced annotations, item A should come before item B.

In order to make the annotations comparable, we have chosen to "upcast" COCTAILL levels by augmenting the data with relative frequencies from the COCTAILL corpus for each expression, as opposed to "downcasting" the crowdsourced data by discarding information about its ordinality. We thus say that item A occurred X times overall in COCTAILL (disregarding its distribution across CEFR levels), and that item B occurred Y times overall in COCTAILL. This means that we add to each expression its relative overall frequency from COCTAILL.

We then sort all expressions according to COCTAILL levels as primary sorting order and according to COCTAILL frequency as secondary sorting order. This means that the resulting list is ordered by CEFR levels (i.e. from A1 to C1), and that within each CEFR level (e.g. A1), items are sorted from most frequent in COCTAILL to least frequent. Based on this version of ordered list, we perform the comparison between coursebook projections and crowdsourcing rankings.

For the crowdsourcing experiment, we see that overall teachers' scores agree best with the coursebook-based rankings as illustrated in table 6.4. By comparing tables 6.3 and 6.4 we also see that the agreement is in fact much weaker between the individual groups and the automatic coursebook labeling than between the different voter groups.

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
L2 speakers-COCTAILL	0.5896	0.6701	0.6649
L2 professionals-COCTAILL	0.6349	0.6205	0.6826
CEFR experts-COCTAILL	0.5683	0.5074	0.5886

Table 6.4: Agreement between voter groups and COCTAILL rankings

In this comparison all correlations are only moderately positive, whereas between groups they are strong which seems to indicate that learners, teachers, and experts agree reasonably well between themselves, but not with the coursebook projections. This, of course, may depend on the way we project coursebook levels, namely, that we are using first occurrence in any of the coursebooks as an indicator and are not taking into account further information on an item's usage in the coursebooks for level projections.

A comparison between the explicit ranking done by the three experts and the levels assigned by projection from the coursebook texts shows that there is a rather moderate correlation for all groups and for some even a moderately high correlation.

Coursebook data (as well as L2 essay data) provide us with indications of which items to focus on initially, while crowdsourcing and expert annotation can help identify central items for learners at different levels. But there are other sources of evidence that need to be elaborated upon, such as access to speech data, frequency and dispersion in L1 corpora, typology of phraseological units, compositionality scores, etc. The size of the coursebook/essay data has its effects on the results as well.

Given the results of the study, one of the important questions is whether we have misplaced items in the coursebook projections, and in that case which methods to use to adjust the levels and to separate core vocabulary from peripheral (and even incidental) vocabulary at different levels.

As complementary experiment not included in the publications, we combine compositionality data with the crowdsourced data. We find that compositionality does correlate with complexity, although more strongly so with crowdsourced complexity judgments (0.6977) than with levels derived from coursebooks (0.3018). Furthermore, adding human compositionality scores to CEFR prediction models does result in a modest increase of 0.01 in F-score, accuracy and recall. In order to see whether lexical complexity predicts compositionality, we also ran the reverse experiment, trying to predict compositionality scores based on feature representations (as a proxy for complexity). However, our results were inconclusive.

7 Applications

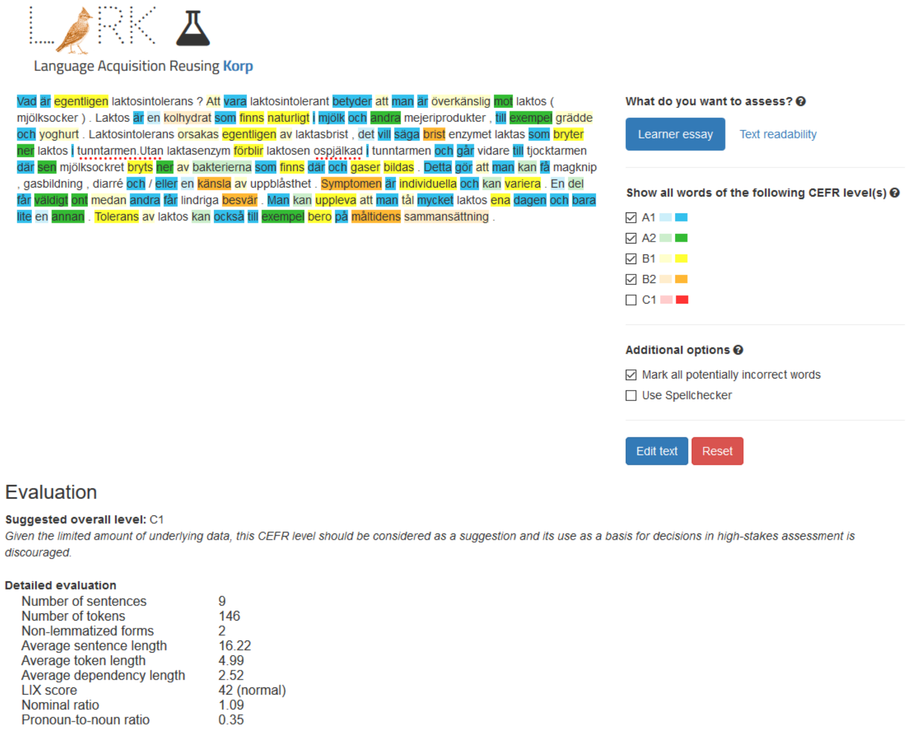
In this chapter, we present prototypes of applications that have been implemented (mostly within the Lärka platform) that make use of the results from this research. We also present further areas of application of the presented research. The main focus of applications lies in the applied linguistics practical domain, especially with regards to natural language processing and language learning.

7.1 Prototypes

7.1.1 Text evaluation

Graded word lists can be used in automatic assessment of writing; it has been shown that lexical features are among the best predictors for text-level complexity in numerous studies and for different languages, both L1 and L2 (Heilman et al. 2007; François and Fairon 2012; Pilán, Vajjala and Volodina 2015; Reynolds 2016; Del Río Gayo 2019). SVALex and SweLLex in their original form have been integrated into the automatic writing assessment module of Lärka, and it has been shown that substituting the previously used KELLY list by CEFRLex lists resulted in significantly improved predictions (Pilán, Volodina and Zesch 2016). On top of improving automatic text level predictions, the inclusion of graded word lists also allows for the visualization of words of different levels as shown in figure 7.1. Thus, one can visually see how many words of which CEFR levels are present in the text, in addition to getting an overall predicted level for the text and a more detailed analysis. Such highlighting of words of different levels has also been applied in the DuoLingo CEFR checker[30].

[30] `https://cefr.duolingo.com/`



Vad är egentligen laktosintolerans ? Att vara laktosintolerant betyder att man är överkänslig mot laktos (mjölksocker) . Laktos är en kolhydrat som finns naturligt i mjölk och andra mejeriprodukter , till exempel grädde och yoghurt . Laktosintolerans orsakas egentligen av laktasbrist , det vill säga brist enzymet laktas som bryter ner laktos i tunntarmen.Utan laktasenzym förblir laktosen ospjälkad i tunntarmen och går vidare till tjocktarmen där sen mjölksockret bryts ner av bakterierna som finns där och gaser bildas . Detta gör att man kan få magknip , gasbildning , diarré och / eller en känsla av uppblåsthet . Symptomen är individuella och kan variera . En del får väldigt ont medan andra får lindriga besvär . Man kan uppleva att man tål mycket laktos ena dagen och bara lite en annan . Tolerans av laktos kan också till exempel bero på måltidens sammansättning .

What do you want to assess?

Learner essay | Text readability

Show all words of the following CEFR level(s)

☑ A1
☑ A2
☑ B1
☑ B2
☐ C1

Additional options

☑ Mark all potentially incorrect words
☐ Use Spellchecker

Edit text | Reset

Evaluation

Suggested overall level: C1

Given the limited amount of underlying data, this CEFR level should be considered as a suggestion and its use as a basis for decisions in high-stakes assessment is discouraged.

Detailed evaluation

Number of sentences	9
Number of tokens	146
Non-lemmatized forms	2
Average sentence length	16.22
Average token length	4.99
Average dependency length	2.52
LIX score	42 (normal)
Nominal ratio	1.09
Pronoun-to-noun ratio	0.35

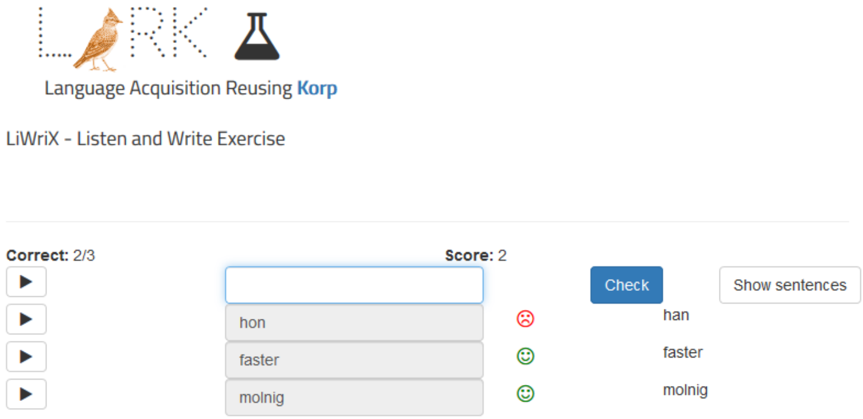
Figure 7.1: Texteval user interface

7.1.2 Automatic exercise generation

A major application of sense-disambiguated graded vocabulary lists is their use in language learning applications. Lärka implements a series of different exercise types such as gap filling, listening, and a hangman variant based on dictionary descriptions, described in chapter 15. For all of these exercises, it is possible to select a proficiency level in Swedish, which impacts the difficulty of the exercise by choosing from a different pool of words. The same is true of the particle verb exercise described in chapter 16. In addition to their use for language learning, these exercises can also be used obtain useful data for second language acquisition research by logging learner interactions. In the following, we present the exercises where graded word lists have been implemented.

7.1.2.1 Listening exercise

While it is not a new exercise in Lärka, the listening exercise has been overhauled from the old version to the current version. In the listening exercise, one hears a word pronounced by a text-to-speech (TTS) engine. One then has to write the word. The interface shows immediate feedback in the form of smileys as well as the correct solution. In addition, one can get hints in the form of sentences in which the target word occurs, and if needed the initial letter of the word in question. Figure 7.2 shows the user interface with two correctly answered items, one incorrectly answered item and the current item to be solved. The original version of this exercise was used to collect a database of L2 spelling errors.



LiWriX - Listen and Write Exercise

Correct: 2/3 **Score:** 2

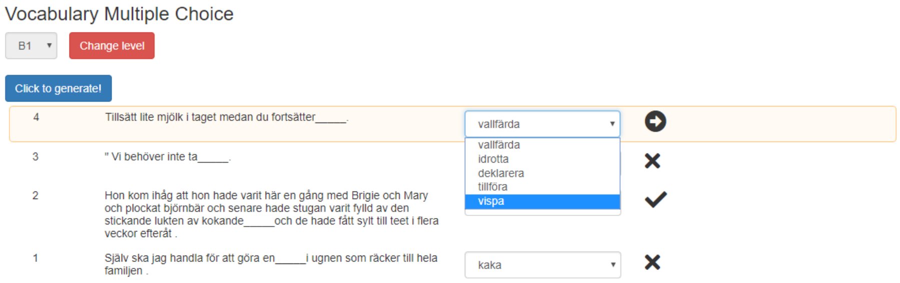
▶			
▶	hon	☹	han
▶	faster	☺	faster
▶	molnig	☺	molnig

Figure 7.2: Listening exercise

We acknowledge that the ability to spell words may be more reliant on the underlying (learned) phoneme-grapheme mapping than proficiency per se. However, we surmise that familiarity with words may be captured through the proxy of assigned target levels and that this factor may play into the success of recognizing words and recalling their spelling.

7.1.2.2 Vocabulary and inflection exercises

While also not new exercises in Lärka, the vocabulary exercise and inflection exercise have been overhauled and use the graded word lists to select a target pool of words of a given proficiency level. Figures 7.3 and 7.4 show the user interfaces. One is shown a sentence at a time with a gap and one has to select either which word or which word form belongs in the gap. Feedback is immediately given in the form of a cross (✖) for a wrong answer or a tick (✔) for a correct answer. Besides their pedagogical use, one could for example investigate which inflectional patterns are known (best) at different proficiency levels.


Vocabulary Multiple Choice

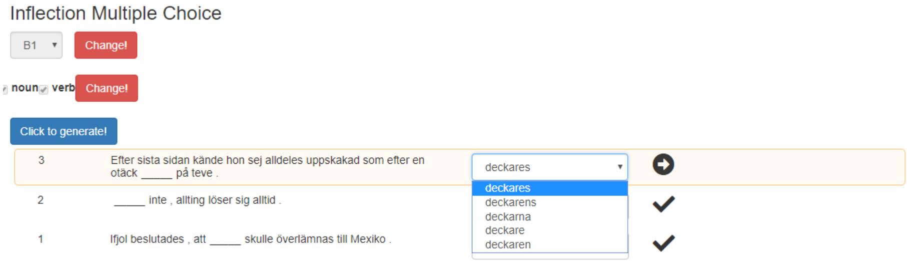
4 Tillsätt lite mjölk i taget medan du fortsätter_____.

3 " Vi behöver inte ta_____.

2 Hon kom ihåg att hon hade varit här en gång med Brigie och Mary och plockat björnbär och senare hade stugan varit fylld av den stickande lukten av kokande_____och de hade fått sylt till teet i flera veckor efteråt .

1 Själv ska jag handla för att göra en_____i ugnen som räcker till hela familjen .

Figure 7.3: Vocabulary multiple choice


Inflection Multiple Choice

3 Efter sista sidan kände hon sej alldeles uppskakad som efter en otäck _____ på teve .

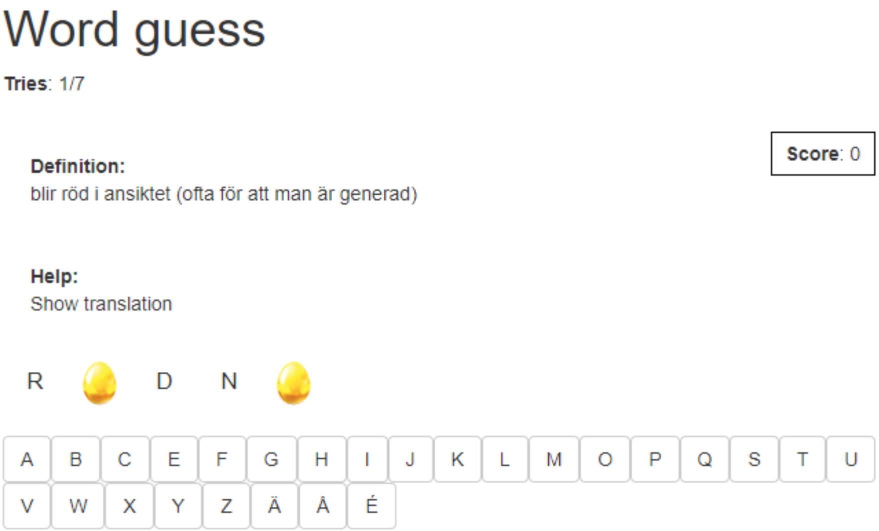
2 _____ inte , allting löser sig alltid .

1 Ifjol beslutades , att _____ skulle överlämnas till Mexiko .

Figure 7.4: Inflection multiple choice

7.1.2.3 Word guess

A new exercise that makes use of graded word lists is *Word guess*, a hangman-like gamified exercise. The exercise is based around dictionary definitions taken from Lexin (Hult, Malmgren and Sköldberg 2010), a multi-lingual dictionary aimed at immigrants in Sweden. Based on the dictionary definition, one has to find the corresponding word by clicking on letters. If the letter occurs in the word, it is shown in the respective positions. Otherwise, one loses a "try", of which one has seven per word. In addition, one can get a hint in the form of the translation of the word in one of the several languages included in Lexin. The use of the hint reduces the awarded point from 1 to 0.5. Figure 7.5 shows the user interface.



Word guess

Tries: 1/7

Score: 0

Definition:
blir röd i ansiktet (ofta för att man är generad)

Help:
Show translation

Figure 7.5: Word guess user interface

We also toyed with the idea of including images in addition to definitions. Figure 7.6 shows a demo version of the exercise for Estonian with images taken from the Basic Estonian Dictionary[31] which covers the CEFR levels A2 and B1 and contains images.[32]

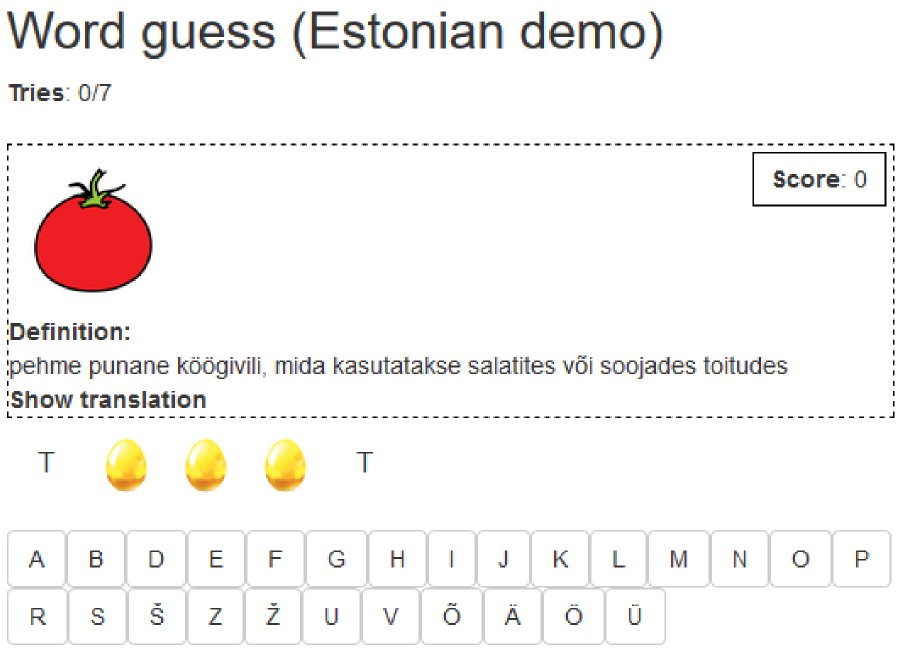
The idea is that after a learner has successfully guessed a word, it can be assumed that they have at least partial understanding of the word and its definition. The learner would then be shown four images and would have to select all images that depict the word in question, or none if none of the images show the depicted word, as shown in figure 7.7.

However, as there were no readily available resources that link Swedish word senses to images, we thought of using crowdsourcing to link words to images in a first step, and to later use the crowdsourced data to add images to the exercise in a second step. As source of images, we used the freely available ARASAAC color pictogram database. ARASAAC is the Aragonese Centre for Augmentative and Alternative Communication and the pictograms are meant to facilitate communication for people who have difficulties communicating.[33] The pictograms are available in black and white and in color. We chose the

[31] `http://www.eki.ee/dict/psv/`

[32] For the demo version, most images were, however, taken from the ARASAAC color pictograms.

[33] The pictograms used are property of Aragon Government and have been created by Sergio Palao to ARASAAC (`http://arasaac.org`) which distribute them under Creative Commons License (BY-NC-SA)



Word guess (Estonian demo)

Tries: 0/7

Score: 0

Definition:

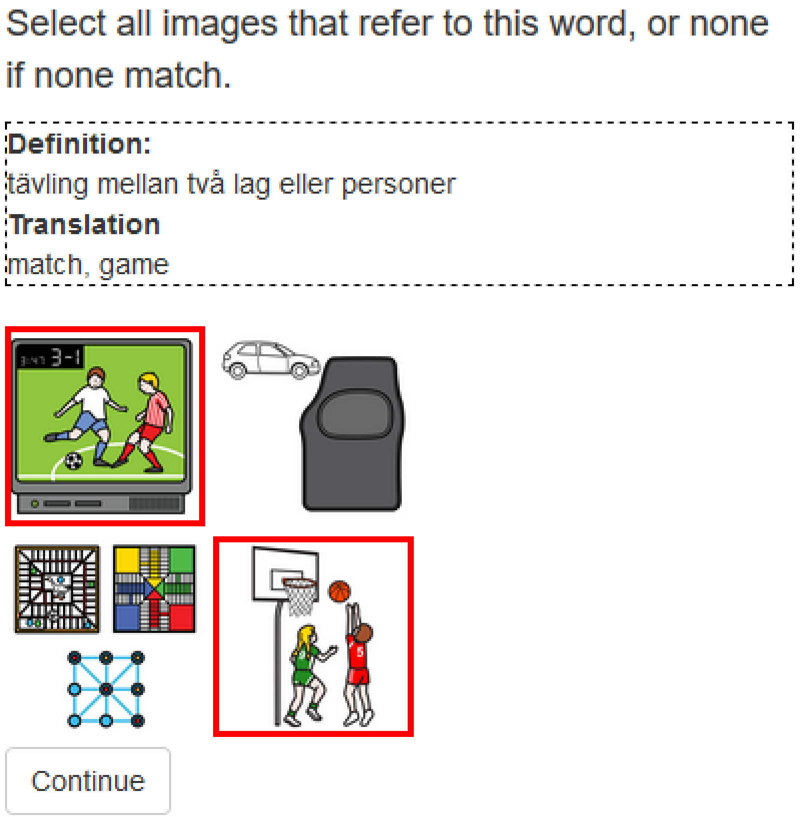
pehme punane köögivili, mida kasutatakse salatites või soojades toitudes

Figure 7.6: Word guess Estonian demo

color pictograms. Labels for the pictograms are available in four different languages, English, Spanish, German and French.

As there are no labels in Swedish, we use the multilingual FastText vectors (Grave et al. 2018) and translation matrices (Smith et al. 2017) to set up a multi-lingual embedding space. We then retrieve the nearest translation for each Swedish word in the other four languages (English, Spanish, German and French). For each of the translations, we then retrieve an image based on the similarity between the translation and the pictogram label. As the labels show certain non-standard features such as ‘table_6.png’ or ‘teacher (female).png’, we have chosen a fuzzy string matching approach using the FuzzyWuzzy Python library (Cohen 2011). The four retrieved images are then considered possible targets for the Swedish word that was used to retrieve them. Finally, if the raw image data of any two retrieved images match, we consider the probability of the pictogram being correct as rather high and retrieve another randomly chosen pictogram to include in the task so that each Swedish word has a set of four distinct images associated with it.

Figure 7.7 shows an example of four pictograms shown for the word *match* ‘match, game’. In this example, the first and fourth images have been selected. It can be argued whether the third image could also be considered a valid target.



Select all images that refer to this word, or none if none match.

Definition:
tävling mellan två lag eller personer
Translation
match, game

Figure 7.7: Pictograms for ‘match, game’

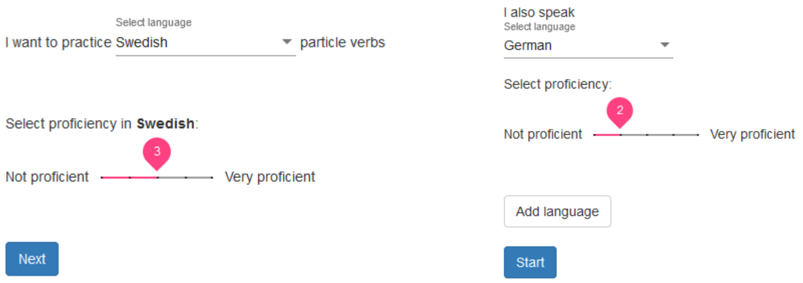
In theory, by using this approach, one can gather information about which images correspond to which words. Even for slightly unclear examples such as shown in figure 7.7, if one gathers a large enough number of judgments by different learners, one can see which image is considered more *prototypical* of the word. While a (separate) prototype of this application has been implemented as proof-of-concept,[34] the functionality is not currently included in the main *Word guess* exercise and does not save information on the selected images.

[34] `https://spraakbanken.gu.se/larkalabb/wordguess-image`

7.1.2.4 Particle verb exercise

The final exercise type implemented using graded word lists is a particle verb exercise.[35] This is the only exercise that was not deployed on the Lärka platform. The exercise targets learners of Swedish and learners of English and is based on particle verbs. The exercise uses multilingual aligned corpora to retrieve translations and the corresponding aligned sentences for all possible language pairs.[36]

The initial idea was that before starting the exercise, the learner is asked about their language knowledge background. As the exercise is based on translations, at least one language besides the target language has to be chosen. For each language, a self-proclaimed proficiency has to be indicated. Figure 7.8 shows the prototype for language selection. The current prototype, however, uses a static set of predefined values instead of the shown language selection. Nonetheless, one can investigate how the knowledge of different languages, both L1 and L2s beside the target L2, affect particle verb knowledge.



Select language
I want to practice Swedish particle verbs

Select proficiency in **Swedish**:
Not proficient 3 Very proficient

Next

I also speak
Select language
German

Select proficiency:
Not proficient 2 Very proficient

Add language

Start

(a) Step 1: target language selection *(b)* Step 2: other known languages

Figure 7.8: Language selection prototype

This exercise is strongly gamified; it features an in-game currency that can be used to "purchase" hints. The learner is also awarded a certain amount of currency for each correct answer. Using hints decreases the reward. The available hints are: translations in any of the selected known languages (chosen randomly), an aligned sentence containing the word pair in question, and the possibility to eliminate half of the answers ("50/50 joker"). The "cost" for translation hints is further weighted as a function of language proficiency; while the language for the hint is chosen randomly, a language in which one is

[35] `http://demo.spraakbanken.gu.se/johannes/paverx`

[36] The included languages are English, Swedish, Spanish, French, Italian, and Portuguese. English and Swedish are mutually exclusive; choosing one language as target language automatically adds the other language to the pool of potential translation candidates.

more proficient will be most expensive.

Figure 7.9 shows the user interface. In this example, the target language is English and the particle verb is shown on top, with the particle removed. The learner has to click on the correct particle from a list of particles with the help of translation hints and possibly the “50/50 joker”. The example shows the first translation variant in Spanish. It is possible to also see the words in context by clicking “Show example” below the translation. This reveals an aligned sentence in the two languages English and Spanish. The prototype is described in chapter 16.

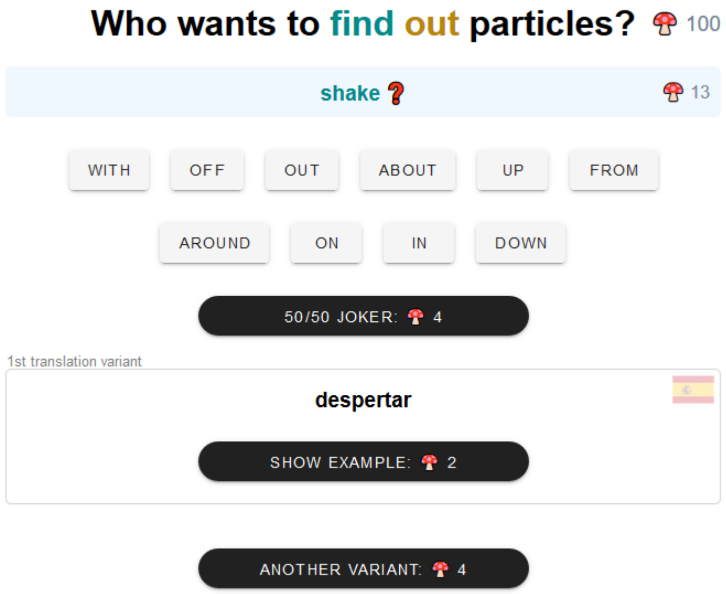


Figure 7.9: Particle verb exercise

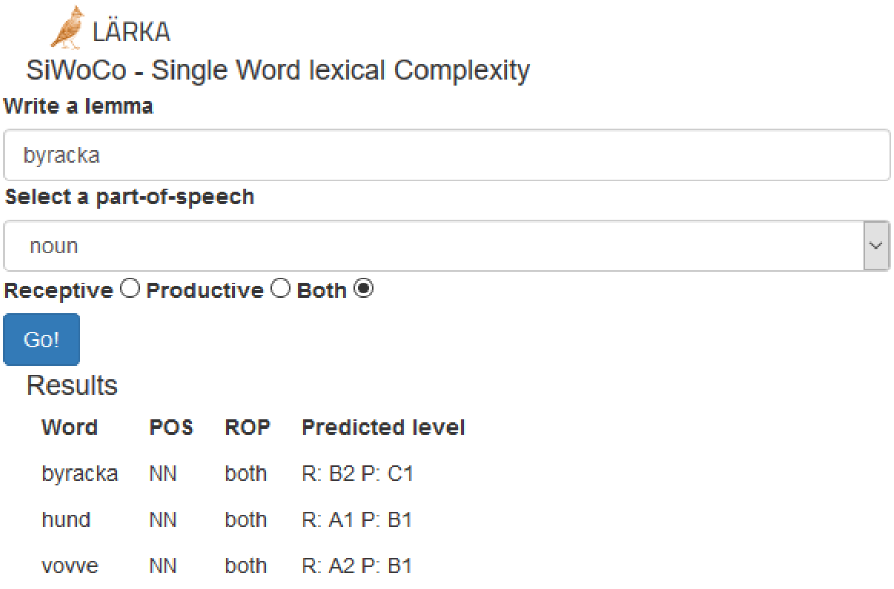
7.1.3 Single word lexical complexity prediction

From a more research-based perspective, one of the main outcomes of this work is the automatic proficiency level predictor for single words called SiWoCo. A user interface has been integrated into Lärka to allow one to submit queries for words.[37] Figure 7.10 shows the user interface. As input, one should specify a lemma and a part-of-speech. Next, one can select to have SiWoCo predict either receptive levels (i.e. at what level a word is expected to be understood at the earliest), productive levels (i.e. at what level a word is expected

[37] https://spraakbanken.gu.se/larkalabb/siwoco

to be produced by learners at the earliest), or both. The classifiers for productive and receptive predictions have been trained on different data and run independently from one another. Figure 7.10 shows the example of three words: *vovve* ‘doggie’, *hund* ‘dog’, and *byracka* ‘mutt’. The predictions for receptive knowledge are A2, A1 and B2, indicating that ‘dog’ and ‘doggie’ are expected to be understood at early levels of proficiency, while ‘mutt’ is expected to be understood at a much later stage of proficiency. The predictions for productive knowledge are B1, B1 and C1, indicating that ‘dog’ and ‘doggie’ are expected to be produced at intermediate levels while ‘mutt’ is expected to be produced at much higher levels. The research is explained in more detail in chapter 10.

SiWoCo, along with other CEFR prediction techniques presented such as custom embeddings and COCTAILL language models, is currently being used in order to identify core and peripheral vocabulary items. The idea is that if multiple methods agree on a CEFR level, it is more probable that the item in question is central to that level.



SiWoCo - Single Word lexical Complexity

Write a lemma

byracka

Select a part-of-speech

noun

Receptive ○ Productive ○ Both ●

Results

Word	POS	ROP	Predicted level
byracka	NN	both	R: B2 P: C1
hund	NN	both	R: A1 P: B1
vovve	NN	both	R: A2 P: B1

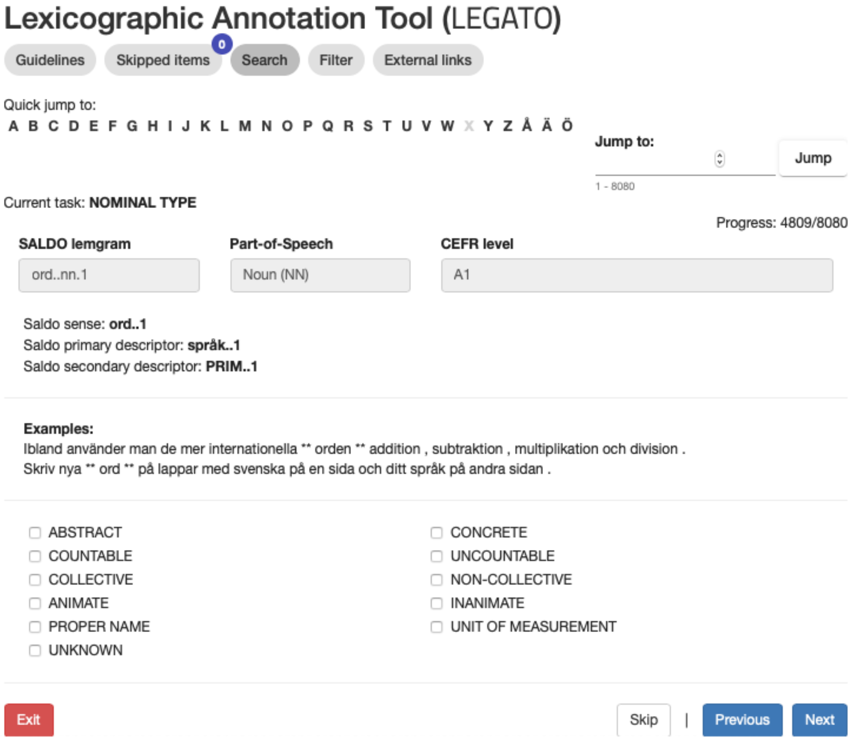
Figure 7.10: SiWoCo

7.1.4 Lexicographic annotation

In order to build more accurate lexical complexity prediction systems, it would be advantageous to have access to more linguistically-informed data such as morphological analysis or transitivity. Thus, in order to further enrich the graded word lists, Legato, a custom lexicographic annotation tool was developed. Figure 7.11 shows the user interface. Through Legato, a multitude of different aspects such as morphological analysis, multi-word classification, nominal type, or transitivity can be added.

Legato offers sophisticated methods of navigation such as an alphabetical quick jump possibility, a numerical index jump possibility, as well as search and filter functionalities to quickly find any entry. It also offers the possibility to *skip* an entry if one is not sure how to annotate it. Skipped entries are collected in a separate list that is available at any time so that the annotator may revisit previously skipped items quickly and easily. Further, a number of external links are provided to check information if needed. Legato is described in detail in chapter 13.

Legato plays a major role in expanding SenSVALex into a rich, multipurpose lexicographic resource that can be used by various people such as learners, teachers, researchers or lexicographers. The annotations also provide rich features for future experiments in word complexity prediction.



Current task: **NOMINAL TYPE**

Progress: 4809/8080

SALDO lemgram	**Part-of-Speech**	**CEFR level**
ord..nn.1	Noun (NN)	A1

Saldo sense: **ord..1**
Saldo primary descriptor: **språk..1**
Saldo secondary descriptor: **PRIM..1**

Examples:
Ibland använder man de mer internationella ** orden ** addition , subtraktion , multiplikation och division .
Skriv nya ** ord ** på lappar med svenska på en sida och ditt språk på andra sidan .

□ ABSTRACT
□ COUNTABLE
□ COLLECTIVE
□ ANIMATE
□ PROPER NAME
□ UNKNOWN

□ CONCRETE
□ UNCOUNTABLE
□ NON-COLLECTIVE
□ INANIMATE
□ UNIT OF MEASUREMENT

Figure 7.11: Legato user interface

7.2 Other areas of application

In the following, we delineate further areas of application that may benefit from the presented research but that have not been explored within the scope of this thesis.

7.2.1 Adaptive diagnostic testing

Graded word lists can be used in diagnostic testing (placement tests) to determine the proficiency of a language learner. In contrast to traditional diagnostic tests which often contain items testing the spectrum from beginner to advanced knowledge, graded word lists, alongside computer technology, can be used to implement *adaptive* diagnostic tests which adapt the difficulty according to the correctness of answers given. This significantly reduces the time required for testing, as the test can home in on the predicted level instead of extensively testing knowledge of each proficiency level.

7.2.2 Resource creation

Graded word lists can be useful for textbook writers in the compilation of dictionaries aimed at learners. They can also serve as tools for both language learners and teachers.

7.2.3 Lexical simplification

Graded word lists can be used in lexical simplification scenarios. For example, if one knows the target audience's proficiency level, all words of a text that are above said proficiency level might be replaced by easier synonyms.

7.2.4 Exposure and emergence in language acquisition

From a more theoretical perspective, it is also possible to study different aspects of language teaching and language acquisition, for example what compounding patterns learners are taught at different levels (nominal compounds, prefixes and suffixes, . . .). This is made possible through the rich lexicographic annotation, which includes morphological segmentation, among other features. Further, since we have two word lists, one based on receptive vocabulary and one based on productive vocabulary, one can also investigate the potential parallels between exposure and production, i.e. at what level learners are exposed to certain expressions compared to at what level learners start to use these expressions; this line of research is being followed within the L2 profiles project.

8 Discussion and conclusion

In this thesis, we look at lexical complexity of single- and multi-word expressions from a second language learner perspective using natural language processing methods.

8.1 Main findings

Looking solely at productive (e.g. EVP) or receptive knowledge (e.g. SVALex) only gives half the picture: one cannot draw conclusions about receptive knowledge based on a resource compiled from learner productions. Conflating both further blurs the picture. We take a dual approach by modeling receptive and productive knowledge as two separate processes based on two different word lists derived from two different types of corpora. In consulting both resources at once one can see similarities and differences between productive and receptive knowledge and their evolution across proficiency levels.

8.1.1 How can frequency distributions across proficiency levels be used to derive target levels?

For research question 1, we have presented a normalization by dispersion followed by a threshold approach based on inspection of the vocabulary list SweLLex that presents not only vocabulary items but also their frequency distribution across the different graded textbooks. The approach is similar to the 1-to-10 approach used by EVP. However, we later dropped this approach for the easier *first occurrence* approach. We surmise that the original threshold approach may have theoretical foundations and make sense in certain scenarios such as productive vocabulary knowledge; in essay writing, it is possible that learners may write existing words by mistake, originally meaning to use a different word (so called *real-word errors*, e.g. writing *colt* instead of *cold*). Only if one has observed a sufficient number of occurrences of a word can one be sure that the word is known to the learner. On the other hand, *first occurrence* may be more suitable in receptive knowledge scenarios where it can be more readily assumed that the language the learner is exposed to is standard error-free language. On our data, we found that both methods of assigning levels based on frequency distributions agree in the majority of cases. This is in line with what other researchers have found.

8.1.2 How can we assign target proficiency levels to words?

Concerning research question 2, we have presented a feature-based CEFR level classification system that uses projected coursebook text levels as target levels. We used a variety of features ranging from count-based features such as word length to semantic features such as degree of polysemy. Using feature selection algorithms, we confirmed that almost all of the selected features were predictive. In chapter 10, we remark on the fact that feature selection excluded 'most predictive bigrams', i.e. a list of bigrams that had the most discriminative power. When we investigated the automatically created list, we saw bigrams such as 'åä', 'xf' or 'xb', which led us to think that something went wrong during calculation. In hindsight, it may simply be that other bigrams did not have enough discriminative power, and that indeed only very few rare instances of bigrams such as the ones listed, did manage to discriminate between levels at all. On the other hand, due to their rarity, even though they would be predictive in a few cases, their overall use is diminished, as the majority of words would not include these bigrams. Indeed, we have a single instance of 'åä' in the word *småäta* 'to snack, to graze', two instances of 'xb' in *byxben* 'pant leg' and *läxbok* 'exercise book', and 16 occurrences of 'xf' in words such as *häxförföljese* 'witch hunt', *oxfilé* 'beef tenderloin' or *byxficka* '(pants) pocket'. The same line of thinking holds true for the other two discarded features, 'compound count' and 'most predictive compounds'; while there might be some compound features with discriminative power, their frequency might just be too low to affect the overall result.

8.1.3 How can we assign target proficiency levels to unseen words?

For research question 3, we have presented a machine learning architecture that uses the feature-based word representations and textbook-derived CEFR levels to learn how to predict CEFR levels for unseen words. We have also modified the system to work on English data in the 2018 shared task on complex word identification; our non-English systems were based on language model probabilities and word length. We also tested neural network architectures and although they do achieve good results, the results are slightly inferior to 'traditional' machine learning methods. This may be due to the amount of data available. However, even in the 2018 shared task on complex word identification where substantially more data was provided, those findings hold: our own neural network architectures did not surpass 'traditional' machine learning results, and the winning team for the English sub-task also used 'traditional' machine learning methods.

In connection to research questions 2 and 3, we found topic distributions to be highly predictive, possibly due to their alignment to the CEFR level descriptors. However, recent research corroborates the usefulness of topic information in non-CEFR-based complex word identification (Avdiu et al. 2019), and even in unsupervised settings (Zotova et al. 2020).

8.1.4 How can we check the validity of the assigned levels?

With regards to research question 4, we have conducted multiple evaluations that seem to confirm the validity of CEFR assignment by first occurrence. The study that least corroborated this method was on multi-word expressions; however, this study also contained the least amount of items under consideration. We are aware that using this methodology will lead to some erroneous assignments for a variety of reasons such as data sparseness. The results might have been skewed possibly due to the low number of items or due to the specific selection of items. Increasing the number of items under consideration might paint a different picture. Another reason might be that the study only looked at multi-word expressions and that multi-word expressions might not be assumed to be understandable or producible at the first level they are encountered.

Multi-word expressions are, from a computational perspective, quite different from single words and thus do not lend themselves readily to the same treatment. Some features such as grammatical gender are not available and other features such as suffixes might be questionable at best. While the automatic target level prediction algorithm presented in chapter 10 can handle any input, including multi-word expressions, the predicted levels are questionable. We thus opt to experiment with different features for multi-word expressions.

8.1.5 Does compositionality correlate with complexity?

For research question 5, preliminary experiments have shown that compositionality annotation of multi-word expressions is highly subjective. Comparing the compositionality judgments with textbook-derived CEFR levels, we find a weak positive correlation (0.30), while comparing compositionality judgments with results from the crowdsourcing experiment, we find that there is a strong positive correlation (0.69). This indicates that compositionality (or transparency) might be used as predictor of lexical complexity. Further preliminary experiments seem to indicate that the reverse does not hold true.

8.1.6 Can crowdsourcing techniques be used to create graded lists?

Concerning research question 6, we have found that learner intuitions as to the difficulty of multi-word expressions correlate to a high degree with intuitions as given by teachers and assessors of Swedish as a second or foreign language. This seems to suggest that internal development in second language learners of Swedish and external assessment methods as practiced by teachers and assessors also correlate. Thus, crowdsourcing techniques can be used to rank lists of words. While setting up and completing crowdsourcing experiments takes more time as compared to for example direct annotation by trained CEFR experts, our results have shown that the output generated in the crowdsourcing experiment by experts and non-experts is very similar, while the output of direct annotation by three CEFR experts is very different; in exchange for speed, crowdsourcing offers a more stable result across the board. In addition, our experiment has shown that, since the results from experts and non-experts are very similar, non-experts can be used for ranking vocabulary, either as complementary to using experts or in the absence of experts.

8.2 Limitations and future work

The major problem that we are facing is the absence of a *gold* standard, i.e. a resource that is trustworthy and ideally well-established, well tested and (manually) verified for correctness. In this case, it would be a resource listing lexical items and their target CEFR levels with regard to receptive and productive knowledge. In itself, the creation of such a resource will encounter numerous problems, one of which is that lexical complexity is subjective, i.e. it can vary from person to person. A certain person might perceive a word as complex, while another person would consider it as simple. This was seen in the data compilation for the Complex Word Identification shared tasks in 2016 and 2018; the organizers have chosen to consider a word complex if 1 out of 20 annotators annotated a word as complex. While this might be a prudent approach (it might be more beneficial to overestimate rather than underestimate complexity), it is not clear whether another approach would have been more sensible. For example, one could have considered a word complex if at least half of all annotators annotated it as complex. Furthermore, the same lexical item can be perceived differently in different contexts. This is further aggravated when taking into account polysemous items. We try to mitigate the effect of polysemy by re-creating the Swedish resources SVALex and SweLLex on a sense-disambiguated level.

One possibility to automatically adjust the predicted CEFR levels (or to find adequate levels even for unannotated data) would be to use exercises to gain insights into how learners perform with regards to their own proficiency and our predicted levels. As the target proficiency levels in the word lists have been automatically derived, they cannot be considered gold standard.

Let us posit the existence of a prototypical learner population L_x of proficiency level x. Let us further posit that there exists a method of unequivocally assigning proficiency levels to learner groups and that this method is perfectly accurate. Let W_x be the set of lexical items of proficiency level x. For any $w_x \in W_x$ and for any $l_x \in L_x$, if there is a marked difference in either correctness (i.e. a non-negligible amount of l_x answer incorrectly compared to other w_x) or time (e.g. it takes considerably longer to answer compared to other w_x), it can be assumed that the level automatically assigned to w_x is incorrect. In both cases, two different hypotheses can be formulated. If the degree of correctness or time for any $w_x \in W_x$ significantly deviates from the average over W_x, it can do so in two ways: positive and negative. In the first case, the degree of correctness is significantly higher, i.e. learners answer exercise items containing w_x correctly more often than other exercise items containing items from the same proficiency level; alternatively, learners answer exercise items containing w_x significantly more quickly compared to other exercise items containing items from the same proficiency level. In these cases, one can surmise that the assigned level is too low and should be corrected upwards. In the case of a negative deviation, degree of correctness will be lower while time taken to answer will be longer. In these cases, one can surmise that the proficiency level of w_x is too high and should be adjusted downwards.

One main aim of using graded word lists in language learning applications is to offer a pleasant learning experience that is neither too challenging nor too boring, i.e. an experience that is tailored to one's own proficiency. This aim is also expressed in the concept of *flow* (Csikszentmihalyi 2000) which describes a state of mental immersion in a task when skill level and challenge are balanced.If the task difficulty is matched with one's (self-perceived) proficiency, one can enter the flow state. If the task difficulty exceeds one's (self-perceived) proficiency, one will eventually become frustrated or anxious. If one's (self-perceived) proficiency exceeds the task difficulty, one will eventually become bored and cease the activity. Thus, if one can track learners' proficiency, one would also be able to adapt the exercise difficulty accordingly, maintaining flow.

The presented research, especially the curated graded word lists and the automatic proficiency level prediction algorithm, can be used to implement (1) automatically generated exercises that promote *flow* as well as (2) self-monitoring exercises that can be used to adapt the automatically predicted target proficiency levels of expressions based on implicitly collected learner feedback.

One aspect that is not taken into account in this work are learner-specific differences. Learners vary with respect to many variables such as their background, their mother tongue(s) and other languages known, interests, work, motivation, to name but a few. All these factors influence the vocabulary that might be known to any specific learner regardless of their level. However, textbooks target general audiences regardless of their background knowledge, i.e. a textbook geared towards A1 learners will presuppose the existence of a certain sub-population of learners that share certain characteristics with regard to their language abilities. Thus, we investigate at what level a word can *generally* be understood by an average learner of that level.

We have noted that there is a need for sense-based graded word lists such as EVP, especially in an L2 Swedish context. While one might consult resources (with sense distinctions) such as *Svenska ordbok*, the de facto Swedish dictionary, or Saldo, these resources do not specifically target learners and also do not contain information as to the complexity of items. With the resources and tools presented in this thesis and the ongoing annotation work using these tools, we hope that in the near future, we will be able to publish a sense-based richly annotated vocabulary list for Swedish L2.

Future work may look into sub-word features such as prefixes or suffixes for complexity research; this would be a logical next step given the step-wise decrease in the object of interest from text to sentence to word. With the ongoing lexicographic annotation of data through Legato, such data might soon be available.

While traditional word-embedding-based context features have been found not to be informative for complexity research on the word level, newer models such as the *BERT family (e.g. Devlin et al. 2018) might be worth investigating.

It might also be interesting to repeat the crowdsourcing experiment with single words or a mixture of single- and multi-word expressions to see whether the conclusions still hold.

Finally, further evaluating the correctness of our resources through automatically generated exercises might be of interest. One idea is to test the automatically assigned target proficiency levels by monitoring time and accuracy over different proficiency groups. If a certain group consistently answers incorrectly and/or takes longer on an expression predicted to be of their level, as compared to an average in precision and/or time over other items predicted to be of their level, one might assume that the true level of the expression might be higher; conversely, comparatively higher accuracy and/or quicker response times might indicate that the level of the expression should be adjusted downwards.

If the experiments were to be redone from scratch, it could be beneficial to more strongly confirm the assigned CEFR labels before training machine learning algorithms, either through manual annotation (preferably through crowdsourcing or similar techniques, as direct annotation tends to be highly subjective) or by calculating for example uncertainty through frequency (i.e. less frequent words would be "less" representative of a level).

8.3 Summary

In this thesis, we investigate how natural language processing (NLP) tools and techniques can be applied to vocabulary aimed at second language learners of Swedish in order to classify vocabulary items into different proficiency levels suitable for learners of different levels. In the first part, we use feature-engineering to represent words as vectors and feed these vectors into machine learning algorithms in order to (1) learn CEFR labels from the input data and (2) predict the CEFR level of unseen words. Our experiments corroborate the finding that feature-based classification models using 'traditional' machine learning still outperform deep learning architectures in the task of deciding how complex a word is. In the second part, we use crowdsourcing as a technique to generate ranked lists of multi-word expressions using both experts and non-experts (i.e. language learners). Our experiment shows that non-expert and expert rankings are highly correlated, suggesting that non-expert *intuition* can be seen as on-par with expert *knowledge*, at least in the chosen experimental configuration.

The main practical output of this research comes in two forms: prototypes and resources. We have implemented various prototype applications for (1) the automatic prediction of words based on the feature-engineering machine learning method, described in chapter 10, (2) practical implementations of language learning applications using graded word lists, described in chapters 15 and 16, and (3) an annotation tool for the manual annotation of expressions across a variety of linguistic factors, described in chapter 13. As for resources, we have started the creation of a sense-based graded vocabulary list based on the earlier SVALex list, but with sense distinctions and enriched with data linked from various other sources as well as manual linguistic annotation across multiple linguistic features and manual correction of problematic items (e.g. words that could not be lemmatized or words that have been assigned multiple senses).

While the question of what makes a word *complex* remains an open question, as evidenced by the ongoing line of complex word identification shared tasks, we have found that our research is comparable to research done at the word level for other languages such as French and English. Ongoing lexicographic annotations in Legato as well as various manual corrections will result in a curated and lexicographically rich resource that we hope will be of use to a broad audience.

Part II

Publications

9 FROM DISTRIBUTIONS TO LABELS

This publication is discussed in sections 5.2 and 5.5.2.

In addition to summarizing this paper, section 5.2 also explains the 10-to-1 method used by the English Vocabulary Profile.

Note
Please note that the formula for *diversity* given in this paper is erroneous. It should be

$$d(w,L) = \begin{cases} \frac{count(y,w,L)}{count(w,L)}, & \text{if } count(w,L) > 0 \\ 0, & \text{otherwise} \end{cases}$$

with y being the number of different learners who used a word w at level L. The original formula uses d both as name for the diversity function and for the variable of number of different learners, leading to confusion.
It should also be noted that this measure was not used in any subsequent experiments, as the *first occurrence* approach is easier to calculate and produces the same result in most cases.

This chapter is a postprint version of the following publication:

Alfter, David, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. SLTC, Umeå, 2016 (No. 130, pp. 1-7). Linköping University Electronic Press.

Abstract

This paper presents work on how we can link word lists derived from learner corpora to target proficiency levels for lexical complexity analysis. The word lists present frequency distributions over different proficiency levels. We present a mapping approach which takes these distributions and maps each word to a single proficiency level. We are also investigating how we can evaluate the mapping from distribution to proficiency level. We show that the distributional profile of words from the essays, informed with the essays' levels, consistently overlaps with our frequency-based method, in the sense that words holding the same level of proficiency as predicted by our mapping tend to cluster together in a semantic space. In the absence of a gold standard, this information can be useful to see how often a word is associated with the same level in two different models. Also, in this case we have a similarity measure that can show which words are more central to a given level and which words are more peripheral.

9.1 Introduction

In this work we look at how information from second language learner essay corpora can be used for the evaluation of unseen learner essays. Using a corpus of learner essays which have been graded by well-trained human assessors using the Common European Framework of Reference (CEFR) (Council of Europe 2001), we extract a list of word distributions over CEFR levels. For the analysis of unseen essays, we want to map each word to a so-called *target* CEFR level using this word list.

The aim of this project is two-fold: first, we want to create a list of words linked to target proficiency levels. Second, we want to apply this list for lexical complexity analysis of unseen learner essays.

Most vocabulary lists used for second language learner evaluation, such as estimation of vocabulary size, are often derived from native speaker (L1) materials and thus might be ill suited to the needs of second language (L2) learners (François et al. 2016). It is hypothesized that second language learners need to focus on aspects of a language which are not present in native speaker materials (François et al. 2016).

However, such word lists are important for example in essay classification or lexical complexity analysis (Pilán, Alfter and Volodina 2016; Volodina, Pilán and Alfter 2016). We thus base our approach on a learner corpus. From this corpus, we extract a list of words with their frequency distributions across proficiency levels. We then link each word to one single proficiency level. In contrast to traditional frequency based proficiency estimations, our approach includes information about learners. We look at “diversity” of a word, i.e. by how many different learners the word has been used at each level. We hypothesize that including diversity scores in the calculation of distribution-to-label mapping yields more reliable and plausible mappings.

The question that remains concerns evaluation. How can we measure the “accuracy” of our mapping in the absence of a gold standard? We address this problem by, on one hand, taking into account expert knowledge from teachers in order to refine the algorithms and, on the other hand, using a second separate approach to see to what extent both methods overlap.

The method we have chosen for evaluation is a semantic space approach. One of the advantages of the semantic space approach is that it gives us graded results; we can see to what *extent* words are similar to each other, possibly identifying core vocabulary and peripheral vocabulary at the different proficiency stages.

9.2 Related work

In the area of Swedish as a second language, several vocabulary lists have been created, such as SVALex (François et al. 2016), SweLL list (Llozhi 2016), Kelly list (Kilgarriff et al. 2014), the Base Vocabulary Pool (Forsbom 2006), SveVoc (Mühlenbock and Kokkinakis 2012) and Swedish Academic Wordlist (Jansson et al. 2012). Of those lists, only SVALex, SweLL list and Kelly list attempted to link vocabulary items to the different proficiency levels according to the Common European Framework of Reference (CEFR) (Council of Europe 2001), indicating at which level words should be introduced (François et al. 2016).

However, the Kelly list has been compiled from web texts intended for L1 speakers and the vocabulary used for first language (L1) speakers may differ from what beginner second language (L2) speakers need to concentrate on (François et al. 2016). Also, the division into the CEFR levels is based on frequency and the list lacks everyday words useful for learners of Swedish as a second language (François et al. 2016).

SVALex and SweLL list on the other hand have been derived from L2 Swedish material. SVALex has been compiled from the COCTAILL textbook corpus (Volodina et al. 2014a) and focuses on receptive vocabulary, while SweLL list has been derived from the SweLL corpus (Volodina et al. 2016a), a corpus of L2 Swedish learner essays, and focuses on productive vocabulary. Neither of these lists link vocabulary items to CEFR levels, but present frequency distributions of lexical items over CEFR levels (Volodina et al. 2014a, 2016a).

In this work we try to use such word lists with frequency distributions over CEFR levels to assign a single CEFR label to each word. This information can be used to analyze texts and visualize the information from a lexical complexity perspective.

9.3 The learner corpus: SweLL

Our experiments are based on SweLL (Volodina et al. 2016a), a corpus of essays written by Swedish as a second language (L2) learners. The data covers five of the six CEFR levels, namely A1-C1. Table 9.1 shows the distribution of essays, sentences and tokens per level. Each essay has been manually labeled for CEFR levels by at least two L2 Swedish teachers. The inter-annotator agreement in terms of Krippendorff's alpha (Krippendorff 1980) for assigning one of the five CEFR levels was 0.80 which reaches the threshold value specified in (Artstein and Poesio 2008) for assuring a good annotation quality. Furthermore, the texts have been automatically annotated across different linguistic dimensions including lemmatization, part-of-speech (POS) tagging and dependency parsing using the Sparv (previously knows as 'Korp') pipeline (Borin, Forsberg and Roxendal 2012). The essays encompass a variety of topics and genres and they are accompanied by meta-information on learners' mother tongue(s), age, gender, education level, the exam setting.

Level	Nr essays	Nr tokens
A1	16	2084
A2	83	18349
B1	76	30131
B2	74	32691
C1	90	60832
Total	**339**	**144 087**

Table 9.1: Number of items per CEFR level

9.4 Extracting the data

We extract a list of words and their frequency distributions over CEFR levels from the SweLL corpus. In contrast to the earlier SweLL list (Llozhi 2016), we calculate relative frequencies for each level and extract further information such as learner counts and topics over levels.

Table 9.2 exemplifies the resulting data. In the first column, we have the lemma of a word, in the second column the corresponding part of speech, followed by the distribution over the CEFR levels A1-C1. Then, we also have columns which indicate the learner IDs (indicated by LI A1, LI A2, etc.). These columns indicate which learner used the word at which level. This information is used when normalizing the data. Finally, we have columns which indicate the distribution of topics (T A1, T A2, etc.) for a given word over different levels. We plan on implementing topic modeling using this information at a later stage.

lemma	pos	A1	A2	B1	…	LI A1	LI A2	LI B1	…	T A1	T A2
göra	VB	0.12	0.23	0.61	…	x2, b1, a3, c7	x1, y1	z9	…	everyday life	…
…											
heta	VB	0.10	0.22	0.46	…	x1, b3, y6, z3	k2, l1, m1	n2, p1	…	personal info	…
…											

Table 9.2: Extracted data: Example

9.5 From distributions to labels

In order to link lexical items to CEFR levels, we have to define how we map from a frequency distribution over CEFR levels to a single level. The following sections describe the algorithm, the problem of why we can't directly map frequency distributions to labels, and word diversity normalization, which solves this problem.

9.5.1 Algorithm

In contrast to receptive vocabulary lists, the concept of 'target level', i.e. at which level a word should be understandable, is not applicable to word lists derived from productive vocabulary.

Instead we look at the *significant onset of use*, i.e. at which level a word is used significantly more often than at the preceding level.

In order to calculate the significant onset of use, for each word we calculate the score D_i at level i as the difference in frequencies between the current level i and the previous level $i-1$ as shown in equation 13. If $i = \text{A1}$, $f_{i-1} = 0$.

$$D_i = |f_i - f_{i-1}| \tag{13}$$

If D_i is higher than a certain threshold value, we take the level i as label for the word. Based on initial empirical investigations with L2 teachers that rate the overlap between teacher- and system-assigned levels, we have found that a threshold value of 0.4 works well; lower threshold values exclude relevant words from a certain level while higher threshold values include words which are deemed to be of a different level.

9.5.2 The problem

If we look at the data, we can see that mapping distributions to labels is not straightforward, e.g. figures 9.1 and 9.2 show the distributions of the words *heta* (verb) 'to be called' and *göra* (verb) 'to do'. Using the *significant onset of use* algorithm, we would predict B1 as label for these words.

However, those words will most probably be used earlier by learners, since CEFR, inter alia, defines CEFR proficiency levels through topics. For example, the CEFR document states that one should be able to "introduce him/herself and others and [...] ask and answer questions about personal details such as where they live, people he/she knows and things he/she has" (Council of Europe 2001: page 24).

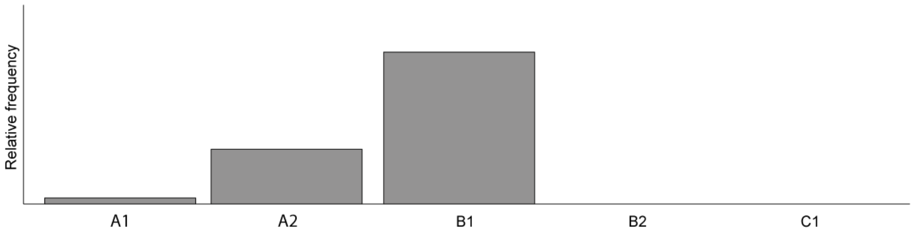


Figure 9.1: Frequency distribution of the word *heta* ‘to be called’

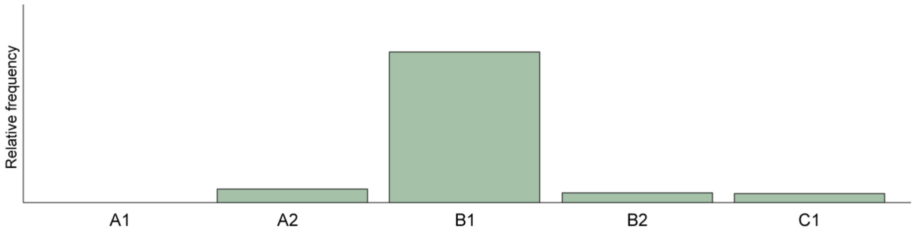


Figure 9.2: Frequency distribution of the word *göra* ‘to do’

The verbs *göra* and *heta* are encountered very often at the beginner level as beginners learn to introduce themselves (e.g. *Jag heter Peter.* ‘My name is Peter.’) and talk about things they do.

Thus, common sense dictates that we cannot simply use frequency distributions as indicators of when learners should be assumed to be able to start using certain words productively.

9.5.3 Word diversity

In contrast to directly mapping frequency distributions to labels, we have found that normalizing the frequencies using *word diversity* improved results significantly. We calculate word diversity for each word by looking at how often the word was used at each level and how many *different* learners used the word at each level. Word diversity of a word w at level L is calculated by dividing the number of occurrences of the word at level L by the number of distinct learners d that used the word at that level as shown in equation 14. The intuition is that if a word is used often at a certain level, but only by one learner, it is less representative of this level than if it is used by many different learners.

$$diversity(w,L) = \frac{count(w,L)}{count(d,w,L)} \tag{14}$$

After normalizing the original frequency distribution to fit into the interval 0-1, we average the word diversity distribution and the normalized frequency distribution to arrive at a new distribution. Figure 9.3 shows the new distribution for *heta*.

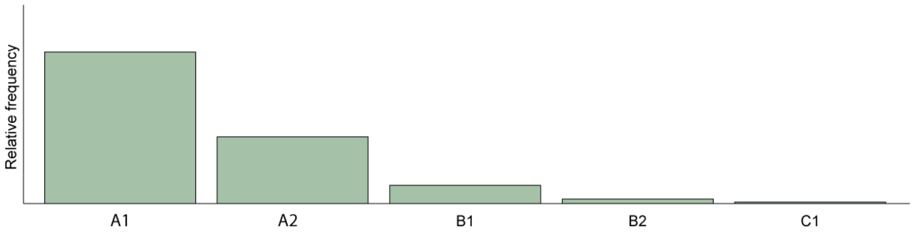


Figure 9.3: New distribution of the word *heta* 'to be called'

We can see that including word diversity shifts the original frequency distribution towards the left, with a peak at A1. Incidentally, the automatically predicted level for this word is also A1; however, it should be noted that the calculation of the significant onset of use differs from simply taking the peak. For example, figure 9.4 shows the recalculated distribution for the relatively common verb *göra* 'to do'. We can see maxima at A2 and B2, but the algorithm predicts the more plausible A1.

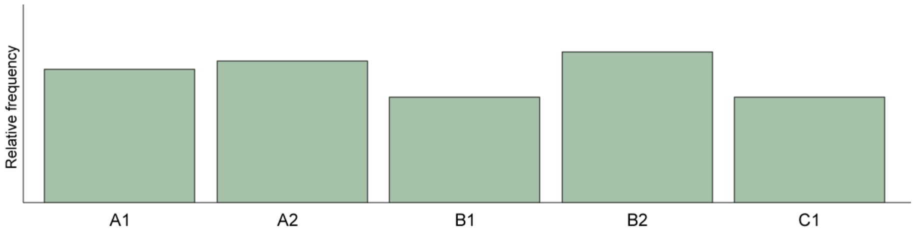


Figure 9.4: New distribution of the word *göra* 'to do'

9.6 Distributional semantics

We used the gensim implementation of Word2Vec (Mikolov and Dean 2013) to create a vector space model of our corpus of essays. Since we don't have a gold standard to validate our results, we wanted to see to what extent we might reproduce the same essay level labeling through a different method. We have 339 essays, each one labeled with a CEFR grade as assigned by a teacher. Given this data, we built two different kinds of semantic spaces: a simple context-based space taking into account a number of words at the left and right of the given lemma; and an "indexed" approach which, for each word in an essay, takes into account both its context and the proficiency level of the whole essay. In other terms, the proficiency level of an essay is treated as contextual information to build a word's distributional vector, in the same way as other words. We also tried a stricter approach where we constrained the system to take into consideration only the proficiency level to build the distributional vector of a lemma, under the assumption that words sharing the same proficiency-related distributional profile would tendentially cluster together in a semantic space, without need for further information.

It is important to understand what kind of spaces these approaches create. If we don't take proficiency levels directly into account, we generate a traditional semantic space where words that have similar contexts cluster together. The problem in creating consistent proficiency-related vocabularies with this approach is clear: if a C1 word happens to be a synonym of an A1 word (and thus used in similar contexts) it will be more similar to such A1 word than to other C1 words.

If we take into account both context and proficiency levels, proficiency level labels become themselves "points" in the multidimensional semantic space: thus, words that occur in the same level will tend to be near, but also a word will be nearer to the proficiency label it shares most context with. The advantage of this method is that we can directly compute the similarity between a lemma and a proficiency level; the disadvantage is that contextual information could actually work as noise. For example, if a complex word as *angelägenheter* 'concerns' (noun) co-occurs with a simple word as *tisdag* 'Tuesday', and *tisdag* mainly happens at level A1, then *angelägenheter* and the point 'A1' will become closer.

If, finally, we only take into account the proficiency level, words that occur in the same level will be similar in the semantic space. In this case we cannot meaningfully compute a word-level similarity but the risks of contextual noise are reduced. It can be interesting to note that since we are using a continuous semantic space we can try to predict the proficiency level (in a direct or indirect way) of full documents by averaging the individual vectors of their words.

We can use one of these models to compute the direct cosine similarity between a word and a level and that we could use to check whether the most similar words to a given level, e.g. B1, are the same we labeled as B1 in our frequency-based approach. On the other hand we can use the other model to see whether words cluster together consistently with our frequency-based lists.

	word-label test	word-word similarity test (*n*-nearest neighbours)		
		1st nearest	2nd nearest	3rd nearest
Indexed model (w=1)	33 (29)	35 (36)	27 (46)	34 (37)
Indexed model (w=60)	**51** (13)	**67** (31)	**44** (37)	**46** (37)
Non-indexed model (w=10)	18 (31)	38 (38)	24 (49)	28 (34)

Table 9.3: Results

9.7 Evaluation

The first reason we used a semantic space to model L2 essays vocabulary is to see whether, using a different approach, we might obtain results consistent with the frequency-based learner-augmented lists we described in the first part of the paper. As we explained, we don't expect simple distributional models to work very well on this task, but we tried to monitor the performance of a so-called "indexed method" to try to make words characteristic of specific proficiency level closer between them and to the level label itself in the semantic space. If a semantic space model trained as described above reproduces the predictions of our frequency-based lists (for example clustering together words that are in the same proficiency level in the lists) we could be a little more confident that our labeling is sensible. To test this we randomly selected 100 words from our frequency lists, equally distributed among the 5 proficiency classes A1-C1. On these 100 words we ran two tests: one based on the word-label cosine similarity, and one based on the word-word cosine similarity. The first test selects, in the semantic space, the nearest proficiency label to a given word. For example given the word *eftersom* 'because', we select the label holding the nearest cosine similarity with it, for example "A2"; if *eftersom* is mapped to the level A2 according to our mapping algorithm, we have an agreement among our models. We can then count how many "nearest labels" coincided with the frequency-based prediction and determine to what extent the two approaches are consistent in modeling the data.

The second test consists in simply retrieving, for every word, its *n*-nearest neighbours in the semantic space. We can then determine whether these neighbours belong to the same proficiency level of the given word in the frequency list. For example, we can retrieve the nearest neighbours of the word *tisdag* 'Tuesday' and find them to be *lördag* 'Saturday' and *trött* 'tired'. If these two words are of the same proficiency level as *tisdag* in our lists, we can suppose a certain consistency between the two approaches.

Table 9.3 shows the results for the different tests and different models. We tested two indexed models, with window size 1 and 60 respectively, and a non-indexed model with window size 10. The numbers indicate how many items were assigned the same proficiency levels in both the semantic space model and the frequency-based mapping, with the upper limit being 100. We are indicating counts, but as the upper limit is 100, the numbers can also be understood as percentages. For the word-word similarity test, we look at the first, second and third most similar words according to the cosine similarity and check whether their proficiency label is the same as the one assigned by the frequency-based mapping. The figures in parentheses indicate the number of *close mismatches* (off-by-one errors).

Apparently, an “indexed” semantic space with a large window shows the highest agreement with our model. Considering that we are predicting labels over five proficiency levels, accuracies of 51% and 67% are encouraging numbers. What is maybe even more interesting is the number of *close mismatches*. These cases are interesting because they could show that the models are setting different boundaries, but tendentially agree on the general progression of the vocabulary. If the number of close mismatches is high, it means that we have many cases where A1 words (in our frequency list) are “labeled” as, or cluster with, A2 words in the semantic space: it is easy to see that similar cases are qualitatively very different from cases where an A1 word clusters with C1 vocabulary. The large presence of similar cases in our results brings us the next reason that induced us to use semantic spaces: they can give nuanced results. If we use a distributional space to label a lemma, we’ll have not only the most probable level of such lemma, but also its distance to the next and previous level. For example, both our frequency list and our best performing semantic space label *resa* ‘to travel’ as an A2 word. From the semantic space, we can also see that it is much closer to B1 than to A1 – we can suppose that it is a rather “advanced” word that tends to lie between A2 and B1. In the same way, *fredag* ‘Friday’, labeled as A2 by the frequency lists, clusters in our space both with A2 and (less closely) A1 lemmas, showing that it is likely to be a term on the “easy” spectrum of the A2 vocabulary.

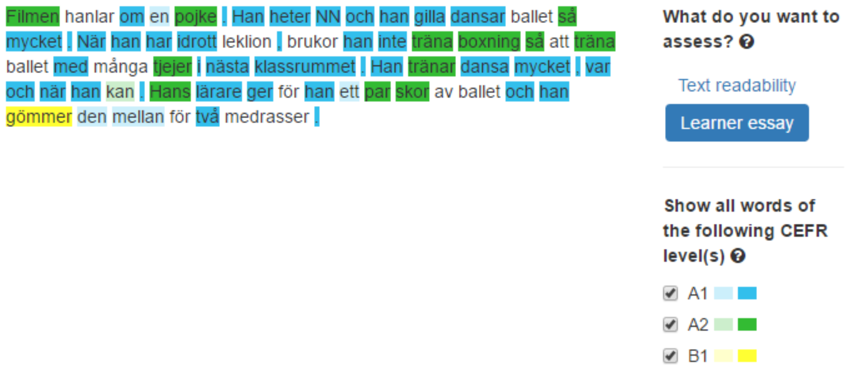
9.8 Lexical complexity analysis

In order to analyze an unseen learner essay, we annotate the essay using the Sparv pipeline (Borin, Forsberg and Roxendal 2012). This step results in a lemmatized and part-of-speech tagged text. Each lemma is then looked up in the previously calculated word list and marked as being of the level indicated in the word list.

We can then simply visualize this information using a graphical user interface[38] as shown in figure 9.5. After entering a text in the text box, it is possible to highlight words of certain CEFR levels. This kind of visualization can give a good impression of the distribution of word levels in a text.

We can also use the word list to predict the overall proficiency level of the essay. Rather than being used on its own, it is incorporated into larger systems. Recent research has shown that substituting traditional frequency based lists by distributionally mapped word lists in machine learning based automatic essay

[38] https://spraakbanken.gu.se/larkalabb/texteval



Filmen hanlar om en pojke . Han heter NN och han gilla dansar ballet så mycket . När han har idrott leklion , brukor han inte träna boxning så att träna ballet med många tjejer i nästa klassrummet . Han tränar dansa mycket , var och när han kan . Hans lärare ger för han ett par skor av ballet och han gömmer den mellan för två medrasser .

What do you want to assess?
Text readability
Learner essay
Show all words of the following CEFR level(s)
☑ A1
☑ A2
☑ B1

Figure 9.5: Text evaluation: Visualization

grading systems results in significantly better predictions (Pilán, Alfter and Volodina 2016).

9.9 Conclusion

In this paper we have shown how lists of frequency distributions of lexical items over CEFR levels can be used for lexical complexity analysis by linking each word to a single CEFR label. We have found that augmenting frequency based lists with learner counts yields more plausible mappings than taking into account only the frequency information. Using a semantic space approach we have shown that our results are consistent across different models. Finally, we have shown how this information can be visualized and used for essay grade prediction.

10 Single word lexical complexity

This publication is discussed in sections 5.3 and 5.4.

Section 5.5.1 presents an extended version of the evaluation of the mapping techniques presented in this article.

This chapter is a postprint version of the following publication:

Alfter, David and Elena Volodina. 2018. Towards single word lexical complexity prediction. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 79–88. New Orleans, USA.

Abstract

In this paper we present work-in-progress where we investigate the usefulness of previously created word lists to the task of single-word lexical complexity analysis and prediction of the complexity level for learners of Swedish as a second language. The word lists used map each word to a single CEFR level, and the task consists of predicting CEFR levels for unseen words. In contrast to previous work on word-level lexical complexity, we experiment with topics as additional features and show that linking words to topics significantly increases accuracy of classification.

10.1 Introduction

A way of addressing the second-language (L2) acquisition needs of the recent influx of new immigrants to Sweden would be to provide an extensive amount of digitally accessible self-study materials for practice. This could be achieved through the development of specific algorithms for exercise/material generation, but such algorithms generally heavily rely on linguistic resources, such as descriptions of vocabulary and grammar scopes per each stage of language development, so that automatic generation of learning materials would follow some order of increasing complexity.

Vocabulary scope can be described through graded vocabulary lists. These are lexical resources where each lexical item is linked to a level at which the item is appropriate for learners to study, one prominent example being the English Vocabulary Profile (Capel 2010, 2012). Graded lexical resources are useful, for example, for course book writers, language test designers, language teachers and language learners, since they can inform the users as to what knowledge is to be expected at which proficiency level, as well as which words to teach and test at which levels.

However, any graded list is a finite resource, as it would never be possible to list by levels all items that learners might encounter. We intend, therefore, to use previously compiled graded vocabulary lists to learn from them to predict levels of previously unseen, out-of-vocabulary (OOV), lexical items.

In practical terms, we look at three automatically created corpus-based vocabulary lists, namely Kelly list (Volodina and Kokkinakis 2012), a resource based on L1 web corpora that identifies frequent vocabulary to guide language learners in their acquisition of vocabulary[39], as well as SVALex (François et al. 2016) and SweLLex (Volodina et al. 2016b), two L2-targeted word lists covering receptive vocabulary and productive vocabulary respectively[40]. The aim of this work is, thus, to create a model that is able to predict the difficulty (i.e. appropriate CEFR[41] level) of any Swedish word with regard to productive and receptive aspects. These graded vocabulary lists are then intended for use in generation of exercises for learners of different levels, though other usage scenarios are also possible.

10.2 Related Work

There has been some work on the creation and evaluation of automatically graded vocabulary lists (Gala, François and Fairon 2013; Gala et al. 2014; Tack et al. 2016a).

[39] Swedish Kelly list is available with CC-BY license from *https://spraakbanken.gu.se/eng/resource/kelly*

[40] Both lists are a part of CEFRLex family of resources, and are available from *http://cental.uclouvain.be/cefrlex/*

[41] Common European Framework of Reference for Languages (Council of Europe 2001) describes six levels of proficiency, starting from A1 to C2

Gala, François and Fairon (2013) aim at identifying criteria that make words easy to understand, independently of the context in which they appear. Since it has been shown that the concept of difficulty depends on the target group (Blache 2011; François 2012), and thus different combinations of features might model certain groups better than others, they focus on speech productions by patients with Parkinson's disease. Gala, François and Fairon (2013) look at 27 intra-lexical and psycholinguistic variables. The intra-lexical variables include number of letters, number of phonemes, number of syllables, syllable structure (CV structure), consistency between graphemes and phonemes, and selected difficult spelling patterns such as double vowels and double consonants. Among psycholinguistic variables are orthographic neighborhood (i.e. words that only differ by one letter), lexical frequency and presence/absence from the Gougenheim list, a list of easy-to-understand vocabulary items.

They train a Support Vector Machine (SVM) classifier on the nine (out of initial 27) most predictive features to predict the difficulty level of unseen words. 5-fold cross-validation on the data shows an average accuracy of 62% in the three-way classification. They conclude that syllabic structures and spelling patterns are not very predictive of difficulty and that the most predictive features are the lexical frequency and presence/absence from the Gougenheim list.

Gala et al. (2014) focus on learners of French, both L1 learners and learners of French as a foreign language. They use Manulex (Lété, Sprenger-Charolles and Colé 2004) to model L1 learners' vocabulary and FLELex (François et al. 2014) to model L2 learners' vocabulary. In contrast to Gala, François and Fairon (2013), they use 49 features which can be grouped into orthographic features (e.g. number of letters, number of phonemes, number of syllables), morphological features (number of morphemes, affix frequency, compounding), semantic features (degree of polysemy) and statistical features such as frequency and presence/absence from the Gougenheim list. They train two SVM classifiers, one for L1 learners and one for learners of French as a foreign language. The first one is a three-way classification while the latter is a six-way classification. On the three-way classification, they reach 63% accuracy and on the six-way classification they reach 43% accuracy. As in Gala, François and Fairon (2013), they find the most predictive features to be lexical frequency and presence/absence from the Gougenheim list. However, they also find the binary poly-semous status, i.e. whether the word polysemous or not, as well as the degree of polysemy to correlate well with the complexity of words. This is an interesting finding, as the degree of polysemy is not directly correlated with frequency.

A related area of work is complex word identification for text simplification. For this task, it is important to identify target *difficult* words or phrases that need simplification (Shardlow 2013; Paetzold and Specia 2016; Yimam et al. 2018). However, in contrast to our work, complex word identification is a binary classification and the focus is slightly different, although there are significant overlaps. Tack et al. (2016b) and Tack et al. (2016a) for example aim at identifying and classifying words of a text into known and unknown ones either for an individual learner or for learners of a given proficiency level as a group. They compare different personalized models with a model based on the graded vocabulary list FLELex (François et al. 2014). Their personalized models also use frequency information, CEFR levels of single words as calculated in Gala et al. (2014), number of letters, and number of senses of a word. For the FLELex vocabulary based model and a learner of a given CEFR level, the model considers all words that are of the same or lower level as the learner's level as known and all words that are of higher level as unknown.

Our recent participation in the Complex Word Identification Task 2018 (Yimam et al. 2018) has yielded interesting findings that we hope will further improve the presented system (Alfter and Pilán 2018).

10.3 Data

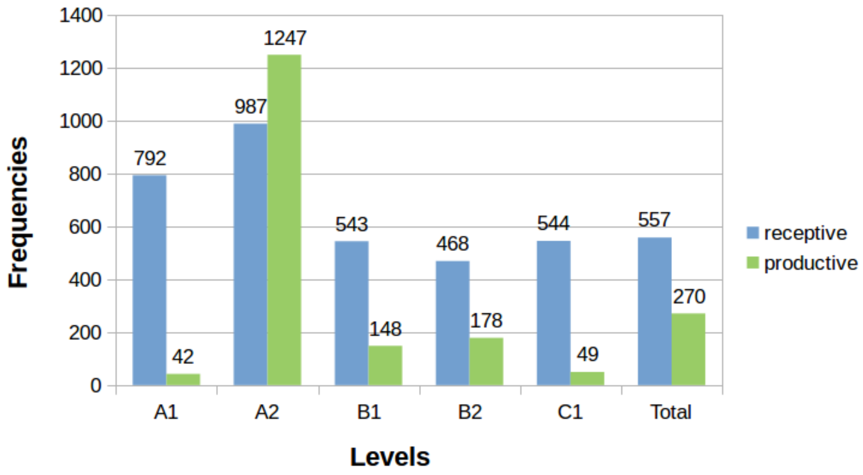


Figure 10.1: Distribution of the verb *arbeta* 'to work', in receptive and productive resources

Our data consists of three different word lists for Swedish, namely SVALex (François et al. 2016), SweLLex (Volodina et al. 2016b) and Kelly list (Volodina and Kokkinakis 2012).

	A1	A2	B1	B2	C1	Total
SVALex	968	1973	2761	6223	3697	15 681
SweLLex	602	1258	1317	1024	1248	6 965
Kelly list	1404	1404	1404	1404	2809	8 425

Table 10.1: Data distribution across lists. In SVALex and SweLLex vocabulary items partially overlap between levels, and hence the total number of items in the list does not equal the sum of items per level.

SVALex is compiled from the COCTAILL textbook corpus (Volodina et al. 2014a), comprised of reading comprehension texts marked for CEFR levels, and covers receptive vocabulary knowledge. SweLLex is derived from the pilot SweLL learner essay corpus (Volodina et al. 2016a) graded for CEFR levels and covers productive vocabulary knowledge. Kelly list is derived from the Swedish Web-as-Corpus (SweWaC) and contains the 8425 most frequent lemmas appearing in native speaker writing divided into CEFR level according to the frequency of the items and corpus coverage. See table 10.1 for the overview of the three resources.

While Kelly list already assigns each word to a target CEFR level, SVALex and SweLLex present distributions over CEFR levels, i.e. how often a word occurs at the different CEFR levels, as exemplified in table 10.2. Since SVALex and SweLLex cover 5 proficiency levels and Kelly list covers 6 proficiency levels, we assimilated the highest level in Kelly list (C2) to the previous level (C1).

To go from distributions to target levels in SweLLex and SVALex, we use the mapping procedures described in Gala, François and Fairon (2013), Gala et al. (2014) (first occurrence) and Alfter et al. (2016) (threshold). For *first-*

Lemma	Part-of-Speech	A1	A2	B1	B2	C1
beta ‘to graze’	VB	0.0	0.0	0.0	19.27	13.21
bo ‘to live’	VB	4978.93	2515.92	1252.19	718.53	497.75
hund ‘dog’	NN	251.89	81.26	250.26	74.29	98.87

Table 10.2: Example of word distributions over levels in SVALex

occurrence mapping, we assign each word to the level it first occurs at. For *threshold mapping*, we assign each word to the level where it occurs *significantly* more often than at the preceding level, with the level of significance set at 30%.

Figure 10.1 shows the distribution of frequencies for the word *arbeta* (Eng. “to work”) over the five CEFR levels in SVALex (receptive resource, 1st bar) and SweLLex (productive resource, 2nd bar). According to the *first occurrence* approach, the target level for both receptive and productive competence for the word *arbeta* would be A1, whereas the *threshold* approach suggests that A1 would be the target level for receptive knowledge, and A2 would be the target level for productive level.

We did a comparison of both mapping methods to find out to what degree they agree. Table 10.3 shows the levels assigned by both methods for the two resources SVALex and SweLLex. By comparing the output of these two mapping methods, we can see that both methods agree to a large extend. When both methods did not agree, they tended to still assign levels that were adjacent, e.g. if one method assigned level B1, the other would assign B2 or A2. This is not a surprise, as the border between different proficiency levels can be fluid. We call this type of disagreement *within one level*. We also see that a certain amount of words were classified as different levels but with the levels assigned being more than one level apart, e.g. one method assigns level A2 and the other method assigns level B2. We call this type of disagreement *more than one level*. Given this finding, and for comparability between studies, e.g. with Gala, François and Fairon (2013) and Gala et al. (2014), we have opted to use the first-occurrence approach in the remainder of the study.

Resource	Same level	Within one level	More than one level
SVALex	12775	1592	1255
SweLLex	5689	706	516

Table 10.3: Number of items that were assigned the same level, within one level and more than level by both mapping techniques

The SVALex and SweLLex data is noisy, because, for one, we cannot validate whether the automatically assigned (mapped) levels are accurate due to missing gold standard annotations, and secondly because of certain errors resulting from automatic corpus annotation. The data is also sparse, and since the mapping procedure for SVALex and SweLLex very much depends on the data available, this introduces further noise. These are the limitations we are aware of and plan to address in the future by collecting and annotating more data.

10.4 Features

From each word, including multi-word expressions such as *göra ont* 'to hurt' and *god morgon* 'good morning', we extract features, grouped into count-based features (i), morphological features (ii), semantic features (iii) and context-based features (iv). Table 10.4 gives an overview of the average values for some selected features per level and resource. As can be seen from this table, words at higher levels tend to be longer, have more syllables, longer suffixes, a higher number of compounds and lower degrees of polysemy and homonymy. Indeed, concerning polysemy, more common words, which are typically found at lower levels, tend to have more different senses than more specialized words found at higher levels.

(i) Count-based and surface form features

- *Length* is the length of the word in characters, our example word *arbeta* (Eng "to work") containing 6 characters. Word length has previously been used to assess linguistic complexity, among others in readability assessment formulas, for example in Smith (1961); Björnsson (1968); O'Regan and Jacobs (1992).
- *Syllable* count is the number of syllables in the word, where *arbeta* contains three syllables. Syllables are counted as number of vowels except for diphthongs ending in 'u' (e.g. 'eu', 'au') which are counted as one syllable. Syllable count has been applied in readability assessment as a measure of increasing text difficulty, e.g. in Flesch (1948); Kincaid et al. (1975), where multi-syllable words have been proven to increase the overall linguistic complexity of a text. By analogy, we assume that the same applies on a single word level.
- *Contains non-alphanumeric characters* is a boolean value that is true if the word contains non-alphanumeric characters, i.e. any character other than A-Z and digits 0-9, for example *13-åring* (Eng. 13-year old).
- *Contains number* is a boolean value that is true if the word contains digits or consists solely of digits.
- The *multi-word* feature indicates whether the lexical expression is made up of more than one single word.

	A1	A2	B1	B2	C1
Average word length					
SVALex	6.00	7.49	8.51	8.85	9.58
SweLLex	5.10	5.98	7.66	8.89	9.91
Kelly	5.74	7.00	7.54	7.86	7.80
Average syllable count					
SVALex	2.08	2.52	2.88	2.91	3.24
SweLLex	1.80	2.01	2.58	2.94	3.28
Kelly	2.04	2.44	2.62	2.78	2.76
Average suffix length					
SVALex	0.54	0.63	0.77	0.80	0.91
SweLLex	0.47	0.51	0.56	0.63	0.71
Kelly	0.70	0.80	0.86	0.88	0.87
Average number of compounds					
SVALex	0.014	0.037	0.052	0.062	0.067
SweLLex	0.038	0.058	0.112	0.125	0.162
Kelly	0.043	0.095	0.137	0.175	0.167
Average degree of polysemy					
SVALex	0.64	0.51	0.39	0.29	0.24
SweLLex	0.55	0.62	0.46	0.36	0.30
Kelly	0.84	0.73	0.67	0.56	0.56
Average degree of homonymy					
SVALex	1.25	1.11	1.06	1.05	1.02
SweLLex	1.35	1.18	1.10	1.08	1.04
Kelly	1.30	1.13	1.08	1.10	1.05

Table 10.4: (Selected) feature averages per level and resource

- For *bigrams*, we calculated all character-level bigrams from each word list and retained only the 53 most predictive ones. This feature is a vector indicating the presence or absence of these 53 bigrams in the target word.

- For *n-gram probabilities*, we calculate character-level unigram, bigram and trigram probabilities with a language model based on the Swedish Wikipedia dump from February 2018. We surmise this also implicitly captures information about grapheme-phoneme correspondence, frequency and suffixes.

(ii) Morphological features

- *Part-of-speech* corresponds to the part-of-speech of the word. For multiword expressions, the part-of-speech of the head noun is taken.

- For *suffix length*, we stem the word using the NLTK stemmer (Bird, Klein and Loper 2009) and subtract the length of the resulting stem from the length of the original word. In *arbeta*, the final -a is a suffix. Previous work on order of acquisition of inflectional versus derivational morphemes, e.g. Derwing (1976), argue that knowledge of derivational morphology is acquired gradually in the learning progress, thus motivating this feature for our experiments. This intuition also seems to hold when looking at average suffix length by level, as shown in table 10.4.

- For *compound count*, we run the word through the SPyRo/SALDO pipeline (Östling and Wirén 2013), which generates possible analyses of the word with regard to compounding. Compound count is the number of possible compounding alternatives. *Arbeta* can theoretically be analyzed as *ar* ‘are (unit of measurement)’ + *beta* ‘to graze’ and thus would have a compound count of 1. *Glasskål* on the other hand can be analyzed as *glas* ‘glass’ + *skål* ‘bowl’, *glass* ‘ice cream’ + *skål* ‘bowl’ and *glass* ‘ice cream’ + *kål* ‘cabbage’ and thus would have a compound count of 3. The cognitive load for processing a word, that potentially has several (compounding) interpretations, hypothetically also influences the word’s complexity, and hence the level at which it is acquired.

- For *compounds*, we calculate all compound elements, i.e. words that have been identified in compounds, in all lists and selected the 12 most predictive compounds. This feature is a vector indicating the presence or absence of these compounds in the target word.

- *Gender* for nouns is taken from Saldo’s morphology (Borin, Forsberg and Lönngren 2008) and encoded numerically as -1 (no information about gender or not applicable), 0 (common gender, aka “en-ord”), 1 (neuter, aka “ett-ord”) and 2 (variable gender). For *arbeta* the value would be -1 since gender only applies to nouns. The majority of nouns in Swedish are of common gender (e.g. in the Kelly-list there are 3465 nouns of common gender, while 1065 are neuter).

(iii) Semantic features

- *Degree of polysemy* is calculated by counting the sub-entries of a given dictionary entry in Lexin (Gellerstam 1999). The verb *arbeta* has only one sub-entry, and is thus non-polysemous. From empirical sources (e.g. various frequency lists), we can observe that non-polysemous words tend to be less used constituting a large bulk of non-frequent words, something that is quite logical given that most word lists are compiled based on lem-grams (e.g. a combination of base form of a word plus its part-of-speech), and not on senses. Usages of several senses of the same lem-gram are thus grouped together in one entry and push the word to the top of the frequency lists. Highly polysemous words, like *komma* 'to come' are thus often learned in the beginning. This seems to be a contradictory trend with regards to our example word, *arbeta* 'to work'. However, if we extend the search to phrasal verbs with *arbeta* in Saldo, there would be seven more entries, and in Lexin four more.

- *Degree of homonymy* is calculated by counting the number of dictionary entries in Lexin with the same orthographic form. An example of a homonym across word classes would be *gift*: it could either be the adjective meaning "married" or the noun meaning "poison". Homonymy within the same word class would be *vara* (Eng. "to last", "to be"). The example word *arbeta* has only one entry in Lexin. Studies on homonymy within second language learning (Mashhady, Lotfi and Noura 2012) show that honomymous words take longer to remember and differentiate between meanings than e.g. several synonyms relating to the same concept, demanding disambiguation of a homonym given the context, which makes homonymy an interesting feature to include into our experiments.

(iv) Context features

- For *topic distributions*, we indicate in which topic lists the target word occurs. Topic lists were extracted from the COCTAILL corpus, where each reading text is assigned one or more topics. We thus extracted all lemmata from reading texts, assigning them to the topics as given in the corpus. We then ran a TF-IDF algorithm over the lists to eliminate words that occurred across all topic lists. This yielded 33 topic lists, such as animals, arts, daily life, food and drink, nature, places, or technology.

Thus, for the verb *arbeta*, we can summarize the above features into the following (simplified) word complexity description: 6-letter 3-syllable non-polysemous non-homonymous verb with one possible suffix, one possible compound analysis, no gender information (since this only applies to nouns), not a multi-word expression and a word used in topics characteristic of presenting people (CEFR levels A1 and A2) which is - supposedly - the reason why the empiric data points out A1 level for receptive and productive knowledge according to *first-occurrence* approach; and A1 for receptive and A2 for productive knowledge if we follow the *threshold* mapping strategy.

	Svalex	Swellex	Kelly
Majority baseline	0.29 ± 0.00	0.29 ± 0.00	0.33 ± 0.00
SVM	0.32 ± 0.02	0.37 ± 0.05	0.39 ± 0.04
MLP	0.32 ± 0.03	0.37 ± 0.04	0.39 ± 0.04
ET	0.27 ± 0.02	0.33 ± 0.05	0.32 ± 0.04
SVM+T	0.44 ± 0.03	**0.41** ± 0.04	**0.45** ± 0.05
MLP+T	0.53 ± 0.04	0.38 ± 0.05	0.44 ± 0.05
ET+T	**0.55** ± 0.05	0.37 ± 0.06	0.43 ± 0.05
SVM+TL	0.48 ± 0.03	0.41 ± 0.05	0.45 ± 0.04
MLP+TL	0.53 ± 0.04	0.39 ± 0.06	0.44 ± 0.03
ET+TL	**0.59** ± 0.03	0.37 ± 0.06	0.42 ± 0.03

Table 10.5: Results: Accuracy and standard deviation using 10-fold cross-validation

10.5 Classification

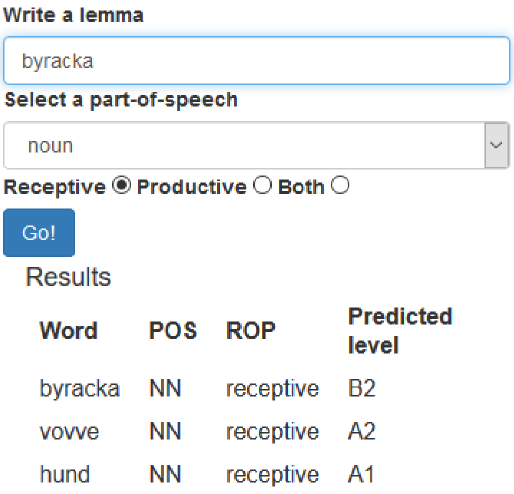
In order to check how well the features we have chosen model single word complexity, we use different classifiers and stratified 10-fold cross-validation on the different data sets.

For classification of unseen words, we train classifiers on the available data. We train one classifier for receptive predictions on SVALex and one classifier for productive predictions on SweLLex.

The classification task consists in assigning each word in our word lists a target CEFR level. For evaluation of the features, accuracy is calculated by comparing the predicted level with the level given by the graded word list. We cannot, at this moment, evaluate classifiers for unseen words, as we would have to have manually graded word lists against which to compare our predictions.

10.6 Results

Table 10.5 shows the results of 10-fold cross-validation classification using different algorithms. Majority baseline always predicts the majority class. Since our data is not balanced, this deviates from the expected chance baseline of 0.2 for five-class classification. SVM is a support vector machine with default parameters $C = 1$ and radial basis function (rbf) kernel. MLP is a multilayer perceptron with 100 hidden layers and a learning rate of 0.01. These parameters were chosen based on a randomized grid search over the parameter space. ET is an extra trees classifier, a classifier from the group of random tree classifiers. Preliminary experiments have shown an initial increase in accuracy with an increase in the number of estimators of the ET algorithm but which shows no further improvement after 100 estimators. We thus have fixed the number of estimators for the ET algorithm at 100. SVM+T, MLP+T and ET+T show the accuracies obtained by the same algorithms but with topic distributions added to the data. For comparability, since we have included all word classes in our experiments, we also tried classifying only lexical word classes (nouns, verbs, adjectives and adverbs) as in Gala et al. (2014). The results of these experiments are shown in the rows SVM+TL, MLP+TL and ET+TL.



Write a lemma

byracka

Select a part-of-speech

noun

Receptive ● Productive ○ Both ○

Results

Word	POS	ROP	Predicted level
byracka	NN	receptive	B2
vovve	NN	receptive	A2
hund	NN	receptive	A1

Figure 10.2: User interface for lexical complexity prediction

In addition, we have created a user interface[42], as shown in figure 10.2. This user interface can be used for getting predictions of any word, not only words present in the word lists . The input word is transformed into a feature vector as described above and then fed into the classifier, which predicts a label. Figure 10.2 shows the predictions for *hund* 'dog', *vovve* 'childish or endearing term for dog' and *byracka* 'derogatory term for dog'.

[42] `https://spraakbanken.gu.se/larkalabb/siwoco`

10.7 Discussion

We found that our features excluding topic distributions barely outperform the majority baseline, yielding even lower scores than the baseline in some cases. Adding topic distributions significantly improves accuracy.

In comparison to the results presented in Gala et al. (2014), we can see an expected trend. Indeed, on the L1 resource Manulex and Kelly (which is based on L1 data but intended for L2 audiences), they reach 63% accuracy in a three-way classification while we reach 45% accuracy in a five-way classification. On the L2 textbook corpus resources FLELex and SVALex, they reach 43% accuracy in a six-way classification while we reach 59% accuracy in a five-way classification.

If we are comparing our results without topic distributions, which are more similar to the results presented in Gala et al. (2014) due to the similarity of features, we see that our best system on L2 data performs worse in a five-way classification (0.32) than theirs in a six-way classification (0.43). This is probably due to the size of the corpus that was used to compile these lists. While FLELex was compiled from 28 textbooks and 29 readers, COCTAILL was compiled from 12 textbooks only. As such, their distributions are less sparse and hypotheses about the target level can be made with more certainty.

Another point is that, in contrast to previous work, we have not included information about lexical frequency explicitly. Including such information could possibly further improve accuracy. It can be argued that n-gram probabilities latently encode this information, but it would be interesting to see whether a more explicit approach would lead to better results.

We also ran cross-validated recursive feature elimination (Guyon et al. 2002) to get a ranking of features and discard useless features. This interestingly identified bigram features (presence/absence of most predictive bigrams; not to be confused with bigram frequency) and compound features as useless, but excluding those features does not lead to an increase in accuracy. However, looking at the most predictive bigram and compound files, it seems that something went wrong during calculation of these, since, for example in bigrams, there are only very rare combinations such as 'åä', 'åo', 'xf' and 'xb'. We would like to address this issue in future work. The final model uses 64 features.

One problem for the classifiers could be that representing words as vectors can lead to the same representation for different words with different levels, which leads to a decrease in learnability since it introduces contradictory data points. We have checked for this and found out that our data contains about 5% of contradictory data points. A possible approach could be to add more disambiguating features.

10.8 Conclusion and future work

We have presented insights from work-in-progress on single word lexical complexity. In contrast to previous work, we show that adding topic information significantly improves results on the classification task. However, the current topic lists can be further refined, for example by synonym expansion, in the hope of improving accuracy.

For future work, one concern that was also expressed in Gala et al. (2014) is that the current lists do not discriminate between different senses of a word. Thus, words like *glas*, meaning either 'glass' as substance or 'glass' as receptacle for drinks, would be assigned one single level while their different senses clearly should be assigned different levels. We are currently working on recalculating the resources SVALex and SweLLex on the sense level by including a word sense disambiguation component in the pipeline.

Another interesting experiment could be to include number of phonemes in our study, since Swedish has some non-transparent grapheme-to-phoneme correspondences.

There is currently ongoing work concerning the collection and annotation of learner essays, which we hope will alleviate the data sparseness problem that we face at the moment, especially with regard to the learner essay based word list.

We would also like to implicitly crowdsource learner knowledge by embedding words from these automatically mapped lists in automatically generated learner exercises. By monitoring how learners of a given level are dealing with words predicted to be of their level, we hope to be able to draw conclusions about the target level of words, i.e. if learners of intermediate B1 level consistently have problems with certain words that our mapping predicts to be of B1 level, we can assume that the prediction was incorrect.

In the future, we intend to evaluate these resources both with teachers of Swedish as a second language as well as language learners to estimate the validity of the automatic mapping. We would also like to create gold standard annotations, both based on these resources as well as new resources.

10.9 Acknowledgements

This work has in part been funded by an infrastructure grant from the Swedish Research Council to Swedish National Language Bank. We would also like to thank the anonymous reviewers for their constructive feedback.

11 Adapting the pipeline to other languages

This publication is discussed in sections 5.4, 5.6, and 6.1.

Section 5.7 additionally discusses conclusions drawn on the basis of this article.

This chapter is a postprint version of the following publication:

Alfter, David and Ildikó Pilán. 2018. SB@GU at the Complex Word Identification 2018 Shared Task. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 315–321. New Orleans, USA.

Abstract

In this paper, we describe our experiments for the Shared Task on Complex Word Identification (CWI) 2018 (Yimam et al. 2018), hosted by the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) at NAACL 2018. Our system for English builds on previous work for Swedish concerning the classification of words into proficiency levels. We investigate different features for English and compare their usefulness using feature selection methods. For the German, Spanish and French data we use simple systems based on character n-gram models and show that sometimes simple models achieve comparable results to fully feature-engineered systems.

11.1 Introduction

The task of identifying complex words consists of automatically detecting lexical items that might be hard to understand for a certain audience. Once identified, text simplification systems can substitute these complex words by simpler equivalents to increase the comprehensibility (*readability*) of a text. Readable texts can facilitate information processing for language learners and people with reading difficulties (Vajjala and Meurers 2014; Heimann Mühlenbock 2013; Yaneva, Temnikova and Mitkov 2016).

Building on previous work for classifying Swedish words into different language proficiency levels (Alfter and Volodina 2018), we extend our pipeline with English resources. We explore a large number of features for English based on, among others, length information, parts of speech, word embeddings and language model probabilities. In contrast to this feature-engineered approach, we use a word-length and n-gram probability based approach for the German, Spanish and French data.

Our interest for participation in this shared task is connected to the ongoing development of a complexity prediction system for Swedish (Alfter and Volodina 2018). In contrast to this shared task, we perform a five-way classification corresponding to the first five levels of the CEFR scale of language proficiency (Council of Europe 2001). We adapted the pipeline to English, and included some freely available English resources to see how well these would perform on the CWI 2018 task and to gain insights into how we could improve our own system.

11.2 Data

There were four different tracks at the shared task. Table 11.1 shows the number of annotated instances per language. For the French sub-task, no training data was provided. Each instance in the English dataset was annotated by 10 native speakers and 10 non-native speakers. For the other languages, 10 annotators (native and non-native speakers) annotated the data. An item is considered complex if at least one annotator annotates the item as complex.

In the dataset, information about the total number of native and non-native annotators and how many of each category considered a word complex is also available.

A surprising aspect of the 2018 dataset was the presence of multi-word expressions (MWE), which were not part of the 2016 shared task. For the 2018 task, the training data contains 14% of MWEs while the development data contains 13%.

Language	Training	Development
English	27299	3328
Spanish	13750	1622
German	6151	795
French	/	/

Table 11.1: Number of instances per language

11.3 Features

We extract a number of features from each target item, either a single word or a multi-word expression. The features can be grouped into: (i) count and word form based features, (ii) morphological features, (iii) semantic features and (iv) context features. In addition, we use psycholinguistic features extracted by N-Watch (Davis 2005). In Table 11.2, we list the complete set of features used for English.

Word length in terms of number of characters has been shown to correlate well with complexity in a number of studies (Smith 1961; Björnsson 1968; O'Regan and Jacobs 1992).

Besides the number of characters, we also consider the number of syllables (S1 and S2). As the calculation of syllables in English is not straight-forward, we use a lookup-based method for S1. In case the word is not present in the lookup list, we apply a heuristic approach as a fall-back. A high number of multi-syllabic words has been shown to increase the overall complexity of a text (Flesch 1948; Kincaid et al. 1975), so we assume it could also be helpful in predicting the complexity of smaller units.

The feature related to bigrams (B1) indicates which character bigrams occur in the target item. We calculate all character-level bigrams in the training data and only retain the 36 most predictive bigrams using Correlation-based Feature Subset Selection (Hall 1999).

N-gram probabilities are based on language models trained on the English Wikipedia dumps from June and July 2015[43]. We calculate character-level unigram, bigram and trigram probabilities.

The Ogden list contains 850 words from Basic English (Ogden 1944) and this feature indicates whether a word is part of this list.

AWL distribution considers the ten Academic Word List (AWL) sublists (Coxhead 1998) and indicates in which lists the word occurs. The AWL list

[43] We already had these pre-calculated language models from previous experiments. For simplicity and time reasons, we chose not to retrain them on more recent Wikipedia dumps.

contains word families which appear often in academic texts but excludes general English vocabulary, making it specific to the academic context. The ten sub-lists are ordered according to frequency, so that words from the first sub-list are more frequent than words from the second sub-list, and so forth.

CEFRLex distribution indicates the presence/absence in the 5th, 10th and 20th percentile English CEFRLex lists[44]. These lists are obtained by aligning and sorting four different vocabulary lists for English (EFLLex) (Dürlich and François 2018), French (FLELex) (François et al. 2014), Swedish (SVALex) (François et al. 2016) and Dutch (NT2Lex) (François and Fairon 2017) by frequency and only taking words which occur in the 5th, 10th and 20th percentile across all languages.

Morphological features include information about parts of speech and suffix length. Suffix length is calculated by stemming the word using the NLTK stemmer (Bird, Klein and Loper 2009) and substracting the length of the identified stem from the length of the original word.

Semantic features are: number of synsets, number of hyponyms, number of hypernyms and sense id. These features are calculated from WordNet (Miller and Fellbaum 1998). The first three are obtained by calculating how many items WordNet returns for the word in terms of synsets, hyponyms and hypernyms. Sense id is obtained by using the Lesk algorithm (Lesk 1986) on the sentence the target item occurs in.

Context features consist of topic distribution and word embeddings. For word embeddings, we use the pre-trained Google News dataset embeddings. We calculate the word context of a word w_i in a sentence $S \in w_1...w_n$ as the sum of word vectors from w_{i-5} to w_{i+5}, excluding the vector for w_i. In case there is not enough context, the available context is used instead. Topic distributions are calculated by first collecting Wikipedia texts about 26 different topics such as animals, arts, education or politics. These texts are tokenized and lemmatized. We then exclude words which occur across all topic lists. Topic distribution indicates in which of these topic lists the target item occurs.

Features from N-Watch include frequency information from the British National Corpus (BNC), the English part of CELEX, the Kučera and Francis list (KF), the Sydney Morning Herald (SMH); reaction times and bi- and trigram character frequencies (B2 and T2). While these features are redundant in some case, such as number of syllables (S1 and S2), their values can differ due to being calculated differently.

Since our pipeline was not designed to handle multi-word expressions, we address this by a two-pass approach. First, we extract all features for single words and store the resulting vector representations. Then, for each multi-word

[44] `http://cental.uclouvain.be/cefrlex/`

expression, if we have feature vectors for all constituents making up the MWE, we sum the vectors for count-based features such as length and number of syllables and average the vectors for frequency counts. We have experimented with adding all vectors and averaging all vectors, but found that summing some features and averaging other features not only yields higher scores but also is linguistically more plausible. Context vectors for MWEs are not added but calculated separately as described above with the difference that for a multi-word expression MWE $\in w_i, ..., w_{i+k}$ occurring in a sentence $S \in w_1, ..., w_n$ as the sum of vectors from w_{i-5} to w_{i-1} and w_{i+k+1} to w_{i+k+5}. In case not all constituents of a multi-word expression have corresponding vectors from phase 1, we set all feature values to zero and only use the context.

Count features
Length (number of characters)
Syllable count (S1)
Contains non-alphanumeric character
Is number
Is MWE
Character bigrams (B1)
N-gram probabilities (Wikipedia)
In Ogden list
AWL distribution
CEFRLex distribution
Morphological features
Part-of-speech
Suffix length
Semantic features
Number of synsets
Number of hypernyms
Number of hyponyms
Sense id
Context features
Topic distributions
Word embeddings
N-Watch features
British National Corpus frequency (BNC)
CELEX frequency (total, written, spoken)
In Kučera Francis (KF) list
Sydney Morning Herald frequency (SMH)
Reaction time
Bigram frequency (B2)
Trigram frequency (T2)
Syllable count (S2)

Table 11.2: Overview of features

11.4 Experiments on the English data

We tried three different configurations for the English data set, namely context-free, context-only and context-sensitive. For context-free, we use the features described above, excluding word embedding context. For context-only, we only use the word embedding context vectors. For context-sensitive, we concatenate the context-free and context-only features.

11.4.1 Classification

We tried different classifiers, among others Random Forest (Breiman 2001), Extra Trees (Geurts, Ernst and Wehenkel 2006), convolutional neural networks and recurrent convolutional neural networks implemented in Keras (Chollet et al. 2015) and PyTorch (Paszke et al. 2017). For Random Forest and Extra Trees, we tried different numbers of estimators in the interval $[10, 2000]$ and found that generally either 500 or 1000 estimators reached the best results on the development set. For neural networks, we tried different combinations of hyperparameters such as the type of layers, number of convolution filters, adding LSTM layers, varying the number of neurons in each layer. We tried two different architectures, one taking as input the features extracted as described below and convolving over these features, the other taking both the features and word embeddings as separate inputs and merging the separate layers before the final layer.

11.5 Experiments on other languages

11.5.1 Predicting the German and the Spanish test set

During testing, we noticed that using the character-level n-gram model trained on the English Wikipedia and using only unigram, bigram and trigram probabilities and word length as features yielded scores in the vicinity of our best-performing feature-engineered models at that time (0.81 F1 vs 0.82 F1).

Following this finding, we used character-level n-gram models trained on Wikipedia dumps[45] for Spanish, German and French and calculated unigram, bigram and trigram probabilities for these languages. In addition, we used target item length in characters as additional feature.

[45] See footnote 1

11.5.2 Predicting the French test set

As there was no training or development data for the French test set, we used the n-gram language model to convert each French entry into n-gram probabilities. We then used the n-gram classifiers for English, German and Spanish to predict labels for each word. We tested two configurations:

1. Predict with English, German and Spanish classifier and use majority vote to get the final label
2. Predict with Spanish classifier and use label as final label

The rationale behind the second configuration is that French and Spanish are both Romance languages. The single Spanish classifier might thus model French data better than incorporating also the English and the German classifiers, as German and English are both Germanic languages.

11.6 Results

Table 11.3 shows the results of the best classifiers on both the development data and the test data. For the English News and WikiNews, the best classifier is an Extra Trees classifier with 1000 estimators with the reduced feature set (see subsection 11.6.1) and trained on each genre separately, as opposed to the general English classifier trained on all three genres. For all other tasks, the best classifier is an Extra Trees classifier with 500 estimators with the reduced feature set.

	F1 (dev)	F1 (test)
EN News	0.8623	0.8325
EN WikiNews	0.8199	0.8031
EN Wikipedia	0.7666	0.7812
German	0.7668	0.7427
Spanish	0.7261	0.7281
French	/	0.6266

Table 11.3: Results of best classifiers

11.6.1 Feature selection for English

Out of the set of features proposed for a certain task, usually some features are more useful than others. Eliminating redundant features can result not only in simpler models, but it can also improve performance (Witten et al. 2011: 308). We therefore run feature selection experiments in order to identify the best performing subset of features. We use the SelectFromModel[46] feature selection method as implemented in scikit-learn (Pedregosa et al. 2011). This method selects features based on their importance weights learned by a certain estimator. We base our selection on the development data and the Extra Trees learning algorithm, since it performed best with the full set of features. We use the median of importances as threshold for retaining features. For the other parameters, the default values were maintained for the selection.

The feature selection method identified a subset of 64 informative features. We list these features in Table 11.4, indicating in parenthesis the amount of features per feature type where it is relevant.

Selected features	
Length	Sense id
Is adjective	# Syllable count S2
Is noun	BNC freq.
Is verb	CELEX freq. (3)
Syllable count S1	KF list
Suffix length	Reaction time
# synsets	SMH
# hypernyms	Bigram B2 freq (4)
# hyponyms	Trigram T2 freq (4)
Topic distr. (22)	Is MWE
Char. bigram B1 (8)	Unigram prob
In Ogden list	Bigram prob
CEFRLex distr. (3)	Trigram prob

Table 11.4: Selected subset of features

The best performing features included, among others, features based on word frequency, information based on words senses and topics as well as language model probabilities.

[46] We also tested other feature selection methods, namely an ANOVA-based univariate feature selection and recursive feature elimination, but we omit the results of these since they were inferior.

As only lexical classes were annotated for complexity, it is not surprising to see that, even though our pipeline considers all part-of-speech classes, the feature selection picked adjectives, nouns and verbs.

11.7 Additional experiments on English

11.7.1 Native vs non-native

Since we had information about how many native speakers and non-native speakers rated target items as complex, we experimented with training classifiers separately for these two categories of raters. We applied the native-only classifier on the native judgments of the development set, as well as on the non-native judgments, and similarly the non-native classifier on native judgments and non-native judgments. In all four configurations, we found accuracy to be the same, at about 75%.

11.7.2 2016 vs 2018

Before this shared task, we experimented with the 2016 CWI shared task data and trained classifiers on it. We tried applying the best-performing classifier trained on the 2016 data on the 2018 development data, but results were inferior to training on the 2018 training data and predicting 2018 development data. The same is true in the other direction; applying the best-performing 2018 classifier on the 2016 data yields inferior results. Table 11.5 shows the result of these experiments. This raises the question of how generalizable these complex word identification systems are and how dependent they are on the data, the annotation and the task at hand.

Configuration	Accuracy	Recall	F1
2016 on 2018	0.6499	0.7463	0.6948
2018 on 2018	0.7992	0.7269	0.7613
2018 on 2016	0.6610	0.6335	0.6470
2016 on 2016	0.8062	0.6511	0.7204

Table 11.5: Results of 2016/2018 comparison

11.7.3 Genre dependency

During the training phase, we concatenated the English training files for News, WikiNews and Wikipedia into one single training file. We did the same with the development data. We trained a single, genre-agnostic English classifier on this data. During the submission phase, we used the single classifier but also split the data into the three sub-genres News, WikiNews and Wikipedia again and retrained our systems, which improved performance. This hints at the genre-dependency of the concept of *complex* words.

11.7.4 Context

As the notion of complexity may be context-dependent, i.e. a word might be perceived as more complex in a certain context, we used word embedding context vectors as features. However, our feature selection methods show that these context vectors do not contribute much to the overall classification results. Indeed, of the 300-dimensional word embedding vectors representing word context, not a single dimension was selected by our feature selection.

However, if we only look at features which can be derived from isolated words, we also have a problem of contradictory annotations. This means that representing isolated words as vectors can lead to the same vector representation of different instances of a word with different target labels. We calculated the number of contradictions and found that representing each word as a vector leads to 5% of contradictory data points.

11.8 Discussion

One interesting aspect of the data is the separation of annotators into native and non-native speakers. However, while it can be interesting to try and train separate classifiers for modeling native and non-native perceptions of complexity, and this information can be exploited at training time, using features that rely on the number of native and non-native annotators could not be used on the test data, as the only information given at test time is the total number of native and non-native annotators, and these numbers do not vary for the English data.

Our best classifiers are all Extra Trees. All other classifiers that we tested, especially convolutional neural networks and recurrent convolutional neural networks, reached lower accuracies. This might be due to insufficient data to train neural networks, a suboptimal choice of hyperparameters or the type of features used.

While our systems did not reach high ranks on the English datasets (ranks 13, 13 and 6 on News, WikiNews and Wikipedia respectively), we reached place 2 on the German data set and place 3 on the French data set. Given the simplicity of the chosen approach, this is slightly surprising. However, we surmise that n-gram probabilities implicitly encode frequency among other things, and frequency-based approaches generally perform well.

Further, we found that using only the Spanish classifier on the French data lead to better scores than using all three classifiers and majority vote. This speaks in favor of the hypothesis that closely related languages model each other better. This can be interesting for low-resource languages if there is a related language with more resources.

11.9 Conclusion

We presented our systems and results of the 2018 shared task on complex word identification. We found that simple n-gram language models perform similarly well to fully-feature engineered systems for English. Our submission for the non-English tracks were based on this observation, circumventing the need for more language-specific feature engineering.

11.10 Acknowledgements

We would like to thank our anonymous reviewer for their helpful comments and the organizers of the shared task for the opportunity to work on this problem.

12 CROWDSOURCING MULTI-WORD EXPRESSION COMPLEXITY

This publication is discussed in section 6.4.

In addition to the information in this publication, section 6.2 contains an analysis of a manual annotation evaluation in order to determine the accuracy of the automatic MWE recognition system. Further, section 6.3 contains an evaluation of manually annotated compositionality values. Finally, the end of section 6.4 contains additional information on comparisons between coursebook-derived levels, manual compositionality annotations and crowdsourcing results.

This chapter is a pre-print version of the following article:

Alfter, David, Therese Lindström Tiedemann, Elena Volodina. In press. Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. Northern European Journal of Language Technology.

Abstract

In this study we investigate to which degree experts and non-experts agree on questions of linguistic complexity in a crowdsourcing experiment. We ask non-experts (second language learners of Swedish) and two groups of experts (teachers of Swedish as a second/foreign language and CEFR experts) to rank multi-word expressions in a crowdsourcing experiment. We find that the resulting rankings by all the three tested groups correlate to a very high degree, which suggests that judgments produced in a comparative setting are not influenced by professional insights into Swedish as a second language.

12.1 Introduction

Many of the challenges in automatically driven solutions for language learning boil down to the lack of data and resources based on which we can develop language learning materials or train models. Resources like the English Vocabulary Profile (Capel 2010, 2012; Cambridge University Press 2015) are a luxury that cost a lot of time and resources to create, and for most languages such resources do not exist. Crowdsourcing has been suggested as one of the potential methods to overcome these challenges. Recently, a European network enet-Collect[47] (Lyding et al. 2018) has been initiated to stimulate synergies between language learning research and practice on the one hand, and crowdsourcing on the other. New initiatives have arisen as a result, e.g. using implicitly crowdsourced learner knowledge for language resource creation (Nicolas et al. 2020), crowdsourcing corpus cleaning (Kuhn et al. 2019), development of the Learning and Reading Assistant LARA (Habibi 2019). However, there are many questions that need to be investigated and answered with regards to methodological issues arising from using crowdsourcing as a method in/for Language Learning.

In this article, we raise some methodological questions about crowdsourcing in the context of second language (L2) learning material creation. To go back to the example of the English Vocabulary Profile – could we generate something similar for other languages without involving lexicographers and experts? For example, given a set of some unordered vocabulary items (e.g. phrases), how can we order them by difficulty/complexity and split them into groups appropriate for teaching at different levels of linguistic proficiency? Could a crowd help us in this scenario? Who can be “the crowd” in that case? How many answers are enough? How many contributors are needed? Are the results reliable?

We focus on whether a crowd of non-expert crowdsourcers can be used to generate language learning materials and how the annotations by experts such as L2 Swedish teachers, assessors and researchers, i.e. people with formal training in teaching and assessing in Swedish, compare to the annotations by non-experts, by which we mean learners and speakers of L2 Swedish.[48] On a more general note, we investigate whether crowdsourcing as a method can be *reliably* applied to language learning resource building using a mixed crowd.

We use a selection of multi-word expressions (MWE) and ask experts (teachers, assessors etc) and non-experts (language learners) to arrange MWEs by complexity. The crowdsourcing part of the experiment is designed in such a

[47] https://enetcollect.eurac.edu/

[48] By L2 Swedish we mean Swedish as a second (third, fourth, . . .) language and as a foreign language

way that we test which *intuitions* that people have about the relative difficulty of *understanding* word combinations. In this design, we do not expect our participants to know anything explicitly about language learning theories, instead relying on their intuitive comparative judgments as intuitive comparative judgments – including ranking items against each other – has been proven to be easier than assigning items to a category (e.g. a level of proficiency) (Lesterhuis et al. 2017). We hypothesize that given an unordered list of expressions, using crowdsourcing, we can derive a list ordered by difficulty that can be used in language teaching. We surmise that difficulty and proficiency are correlated, thus one might expect more difficult expressions to be learned at later stages of language development.

The theoretical notion of *L2 proficiency* is of special importance in connection to this study. Proficiency is a key concept in Second Language Acquisition (SLA) research. It is used to describe the language "knowledge, competence, or ability" (Bachman 1990: p. 16, as cited in Carlsen 2012: p. 163) of a learner on a conventionalized scale, one example being the 6-level scale adopted by the Common European Framework of Reference (CEFR) (Council of Europe 2001, 2018). Conventionalized scales of proficiency levels are useful in educational and assessing contexts, e.g. which group to place a student into (Bachman and Palmer 2010) and in various social and political scenarios, e.g. whether an applicant can be granted citizenship (Forsberg Lundell 2020). However, a straightforward division into levels is a tricky endeavor, since there is no consensus how to define a level and its corresponding competence(s) in concrete terms. SLA research is specific about viewing proficiency as a "coarse-grained, externally motivated" construct (Ortega 2012: p. 134), where levels are always somewhat arbitrary (Council of Europe 2001: p. 17) and proficiency should be seen as different to *L2 development* which is "an internally motivated trajectory of linguistic acquisition" (Ortega 2012: p. 134).

For this reason, current approaches to proficiency advocate rather a scalar/interval approach as it is more powerful, realistic and nuanced (Ortega 2012; Council of Europe 2018; Paquot, Naets and Gries 2020). The current experiment is proof of the usefulness of such an approach where rather than stating that certain vocabulary belongs to a certain level, we can instead state that some vocabulary items are perceived as easier or more difficult in comparison to each other and form a growing scale of items which are likely to be learned in that approximate order.

This article is structured as follows: we introduce related work in Section 2 and describe the data used for the experiment is Section 3. Sections 4 and 5 introduce Methodology and Experimental design. In Sections 6 and 7 we present our main results, analyze and discuss them. Section 8 concludes the article.

12.2 Related work

Previously, several approaches have been used in identifying and ordering relevant vocabulary items for second language learning. A popular approach is to use reading material written by first language (L1) speakers such as newspapers to generate frequency-based word lists (e.g. the Kelly lists (Kilgarriff et al. 2014), the General Service List (West 1953), the New GSL (Brezina and Gablasova 2015)). Such word lists tend to use frequency of occurrence in a corpus as the only criterion for deciding which items that should be taught first, and which ones should be introduced later, following the hypothesis that more frequent words would be easier (cf e.g. Eskildsen (2009) regarding usage-based approaches to L2 acquisition) and more important to know and more rare words more difficult and less critical for communication in a target language. While such lists are useful, they also have drawbacks, especially in the context of second language learning. Indeed, L1 reading material is rarely adequate for language learner needs and lacks important vocabulary items (François et al. 2014: p. 3767).

In order to address the L2 learner needs, there has also been work on using L2 materials as a basis for word lists. One possible approach is to use graded textbooks as a starting point, as has been done in the CEFRLex project.[49] The motivation behind this approach is that textbooks generally target a specific proficiency group of language learners and have been carefully written with the needs of second language learners in mind. The project so far has resulted in the creation of six corpus-based language lists in six languages (FLELex for French (François et al. 2014), SVALex for Swedish (François et al. 2016), EFLLex for English (Dürlich and François 2018), NT2Lex for Dutch (Tack et al. 2018) and ELELex for Spanish (François and De Cock 2018)). Each of these word lists not only contains the overall frequency but also the distribution of frequencies over the different CEFR levels. These projects have assumed that, in theory, the level at which a text is used in a language learning scenario can be used as an indication of a level at which vocabulary of that text can be assumed to be understood by learners and thus can be qualified as a learning target. In practice, however, this relationship is not as straightforward.

Another approach based on L2 material is to use graded learner essays. This has been done in projects such as the English Vocabulary Profile (EVP)[50] (Capel 2010, 2012) and SweLLex (Volodina et al. 2016b). SweLLex belongs to the CEFRLex family, as it has been created by the same methodology behind CEFRLex, but in contrast to other resources in the family, it is based on learner

[49] https://cental.uclouvain.be/cefrlex/

[50] https://www.englishprofile.org/wordlists

essays, more specifically the SweLL pilot corpus (Volodina et al. 2016a) of graded essays written by learners of Swedish. Both of these resources have also experimented with a threshold approach to assigning levels (Hawkins and Filipović 2012; Alfter et al. 2016), i.e. taking as indicative level not simply the first occurrence but the first *significant* occurrence, i.e. the level at which a word or expression is used a certain number of times as defined by a threshold value. Deriving word lists from learner essays may prove more reliable as the amount of data increases (Pilán, Volodina and Zesch 2016), and when the non-standard learner language has been effectively standardized (i.e. corrected) to the target language forms since automatic annotation is almost always trained on standard L1 materials (cf. Stemle et al. 2019). Both aspects, however, are non-trivial and very few languages enjoy the luxury of extensive corrected collections of learner-produced data.

Finally, one can consult L2 experts to rely on their judgments as to the difficulty of items. Expert judgment as a method has been widely applied in general linguistics as well as in second language oriented experiments and L2 resource creation (e.g. Spinner and Gass 2019; Capel 2010, 2012), although not without criticism. One of the potential stumbling blocks is the *subjective intuitive* nature of judgments, something which is claimed to be a major obstacle to reliable scientific conclusions; observations, i.e. language *production*, are regarded as a more reliable and desirable source of data (Bloomfield 1935). However, Chomsky and Halle (1965) argue that judgments versus observations reflects the dichotomy between competence versus performance. In the end, expert judgments reflect experts' professional experience, and are based on evidence coming from *their* practices and theoretical assumptions about L2 teaching, and thus inevitably reflect personal interpretations of these. The challenge is, thus, to overcome the subjectivity of judgments without losing correctness of the final conclusions, so that the results can be used as a basis for assumptions about language learning paths and for scheduling learning materials in an optimal (although obviously never perfect) way. We call this method of annotation *direct labeling*[51] by experts. In Spinner and Gass (2019) it is also referred to as the *"Hey Sally"* method, indicating decision making based on consulting with other expert colleagues to either reduce or confirm the personal subjective bias.

Due to problems with the reliability of manual level assignment, some people have experimented with the number of experts and procedures that would be necessary to gain reliable objective results. Carlsen (2012) notes that the Norwegian L2 corpus project ASK (Tenfjord, Meurer and Hofland 2006) used 10 CEFR assessors for their essays, who for the most part worked in groups of

[51] alternatively called explicit labeling/annotation

5 so that each essay was marked by at least 5 assessors to get a reliable result. Whereas Leńko-Szymańska (2015) used 2-4 raters for the level assignment of her subset of the international corpus of learner English (ICLE) (Granger et al. 2009) to reach agreement between the raters and Díez-Bedmar (2012) reported very low inter-rater reliability when using only 2 raters to assign CEFR-levels to Spanish university entrance exams. Furthermore, previous research has shown that the background of the rater is of much importance (cf Díez-Bedmar (2012) for an overview), although the results have been mixed. Experienced raters have sometimes rated more strictly (Sweedler-Brown (1985) as cited in Díez-Bedmar (2012)) but in other studies they were more lenient (Weigle (1998) as cited in Díez-Bedmar (2012)). Whether a native speaker or not has also been seen to have an effect, and in addition gender, but once again the results were mixed. Díez-Bedmar (2012) also shows that how different rater backgrounds rate the proficiency have also depended on whether holistic or analytic scales were used.

The inherent order of teaching the items on the various lists above, however, is not always obvious. Frequencies can be misleading, insufficient or sometimes idiosyncratic Expert judgments might be perceived as less idiosyncratic but can be inaccessible due to the costs entailed in expert work. Crowdsourcing as a method of annotation could be worth exploring to address the above mentioned weaknesses.

To the best of our knowledge, crowdsourcing has not been extensively used for such ordering tasks. However, we surmise that it might be an alternative to the more heavily resource reliant methods. Crowdsourcing can take different forms. On the one hand, it can be quite explicit about the crowdsourcing aspect. In its original form, it would consist in the annotation of the same data by different annotators (Fort 2016) or the collaborative creation and curation of resources such as Wikipedia (Stegbauer et al. 2009). Such forms generally rely on intrinsic motivation. However, if there is a lack of intrinsic motivation for whatever reasons, two different approaches have been taken, the first of which is paying people, and the second of which is making the task more fun by adding game-like elements (Chamberlain et al. 2013). The monetary aspect is expressed in platforms such as Amazon Mechanical Turk which pays participants to answer questions and/or solve tasks (Buhrmester, Kwang and Gosling 2016). On the other hand, crowdsourcing can be more subtle, such as in Games With A Purpose (Lafourcade, Joubert and Le Brun 2015) (GWAPS). GWAPS are games or gamified platforms that serve a specific purpose which is not merely ludic.

There is research on creating language resources using crowdsourcing, some of which are: Zombilingo for syntactic annotation (Fort, Guillaume and Chastant 2014), Phrase detectives for co-reference annotation (Chamberlain, Poesio and Kruschwitz 2008) or JeuxDeMots for the creation of a lexico-semantic network (Lafourcade and Joubert 2008). However little work has been done on the combination of crowdsourcing and language *learning*. Probably the most well-known approach on combining crowdsourcing and language learning was done by Duolingo (Garcia 2013), although besides the stated goal of "translating the web while learning a language", it is not quite clear how the output is used. Recently, the use of implicit crowdsourcing techniques using language learners for the creation of language resources on par with expert-created content has also been explored (Nicolas et al. 2020).

A related field of work is crowdsourcing for education, of which the closest sub-aspect pertaining to this work is the creation of educational content. Initiatives include for example crowdsourced textbook generation (Solemon et al. 2013) or crowdsourcing video captioning correction by language learners to enhance learning (Culbertson et al. 2017). The interested reader is referred to Jiang, Schlagwein and Benatallah (2018) for an extensive review of current literature and practices.

12.3 Data

COCTAILL (Volodina et al. 2014a) is a corpus of coursebooks for Swedish as a second language that we used as the basis for identification of candidate MWEs for this experiment. COCTAILL contains texts and exercises aimed at adult learners of Swedish, and covers five CEFR levels: A1, A2, B1, B2, C1, where A1 is beginner level and C1 is advanced level (Council of Europe 2001), with several coursebooks at each level (see Table 12.1). In the corpus, each chapter (lesson) in a coursebook has been assigned a level at which it is known to be used in an L2 teaching context. For example, suppose a textbook T contains 9 chapters and that practicing teachers are using chapters $T1$-$T4$ when teaching students aiming for the A1 level, and chapters $T5$-$T9$ aiming for the A2 level. All texts that are used in chapters $T1$-$T4$ are surmised to target A1 level knowledge, while texts that are used in chapters $T5$-$T9$ are assumed to target A2 knowledge, and so on. Further, all words that are used in the texts in chapters $T1$-$T4$ are labeled as potential target receptive vocabulary for the A1 level. All new vocabulary items that are used in texts in chapters $T5$-$T9$ (and that have not been used at previous levels) are labeled as potential target vocabulary at the A2 level, and so on. This approach allows us to generate useful vocabulary lists for both pedagogical and assessment use, as well as for automatic classifications of various kinds. However, generalizations about the levels at which vocabulary items that should be targeted remains only an assumption that needs to be confirmed. Thus, the projected levels at the word level can serve as indications that certain items might be easier or harder, although we make no claims about the correctness of these projections.

Table 12.1 shows an overview of the corpus, detailing how many books targeting each CEFR level that are included, how many authors we rely on, as well as the number of chapters, texts, sentences and tokens.

CEFR level	#Textbooks	#Authors	#Chapters	#Texts	#Sentences	#Tokens
A1	4	10	37	101	1581	11132
A2	4	10	105	232	4217	37259
B1	4	12	83	345	6510	79402
B2	4	8	31	314	8527	101583
C1	2	2	22	115	5085	71991
Total	18	42*	278	1106	25920	301367

* 26 unique

Table 12.1: Statistics over COCTAILL per level

COCTAILL is annotated automatically with the Sparv-pipeline[52] (Borin et al. 2016) for base forms, word classes, syntactic relations, word senses, MWEs and some other linguistic aspects. MWEs are identified on the basis of Saldo (Borin, Forsberg and Lönngren 2013) entries, which means that only MWEs that are contained in Saldo will be recognized. As Saldo is under active development, the automatic pipeline will probably be able to identify more MWEs in the future. From the annotated version of COCTAILL, we have generated a new version of the SVALex list (François et al. 2016) based on senses, as Sparv has been updated to include a word sense disambiguation module since the creation of the original list. Word sense distinctions are based on Saldo senses.

An entry in the list consists of a combination of a base form with its word class (i.e. a lemgram), plus a word sense. Polysemous items have several distinct entries in the list and different frequency counts are associated with each of the sense entries. Each item contains its frequency distribution across different CEFR levels where it occurred and is associated with the lowest CEFR level of the texts in which it is observed. Starting from the list of 1351 MWEs in the list, two annotators classified them manually according to a custom typology.

For the experiment, we chose three different groups of MWEs based on this manual annotation, aiming to select a wide yet balanced variety of different types of expressions. This resulted in the selection of the following three

[52] https://spraakbanken.gu.se/sparv

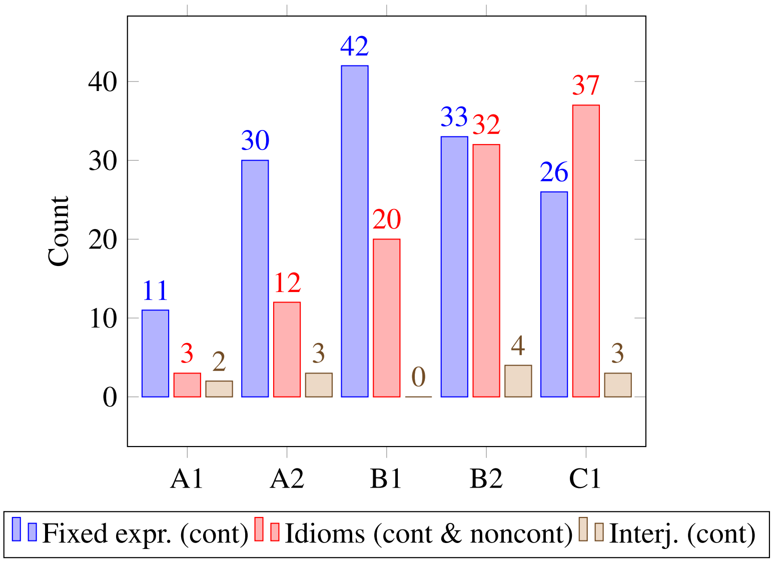


Figure 12.1: Group 1 in the crowdsourcing experiment

groups: (1) interjections, fixed expressions and idioms,[53] (2) verbal MWEs and (3) adverbial, adjectival and non-lexical MWEs. For the sake of conciseness and spatial limitations, we will refer to group 1 as "interjections", to group 2 as "verbs" and to group 3 as "adverbs". Figures 12.1, 12.2 and 12.3 show the number of occurrences per group per level in the resource based on the first round of annotation. From each of these three groups, we selected 60 expressions to be used in the experiment, with 12 items for each CEFR level, for a total of $3 * 60 = 180$ expressions. Expressions were de-contextualized in the sense that we did not provide any example sentences illustrating the use and context of the expression. While this decision may hinder the decision making process, it ensures that decisions are solely based on the expressions themselves, as opposed to syntactic complexity or other features that might be judged in a sentence.

Within each of the groups we prioritized items that had been classified and agreed upon by both annotators. We double-checked all items in the COCTAILL corpus to see that the *sense* we had listed was the one used in the corpus at the automatically assigned CEFR level; this step was necessary, as

[53] We are aware of the difficulty of such distinctions. We tried to give strict definitions of fixed expressions and idioms as well as providing illustrative examples of both. However, comparisons of the annotations of the two annotators have shown that what annotator 1 classified as one of the categories could sometimes be annotated as one of the other categories by the other annotator which is why we decided to have these as a joint group for the experiment.

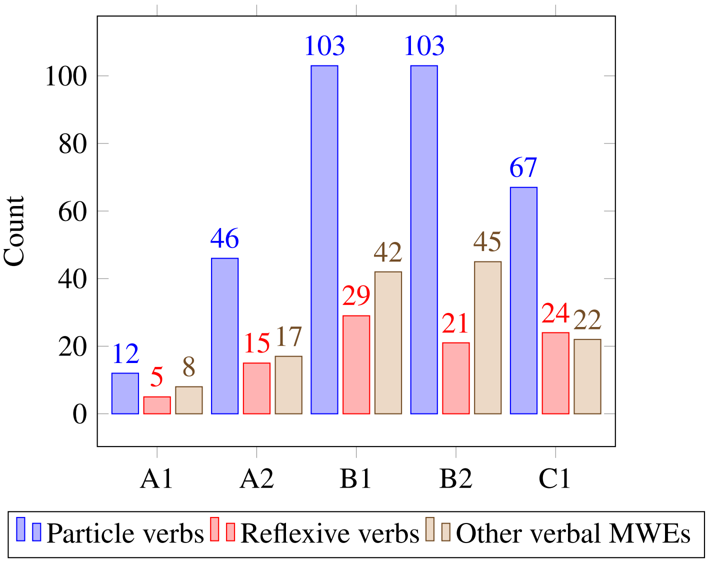


Figure 12.2: Group 2 in the crowdsourcing experiment

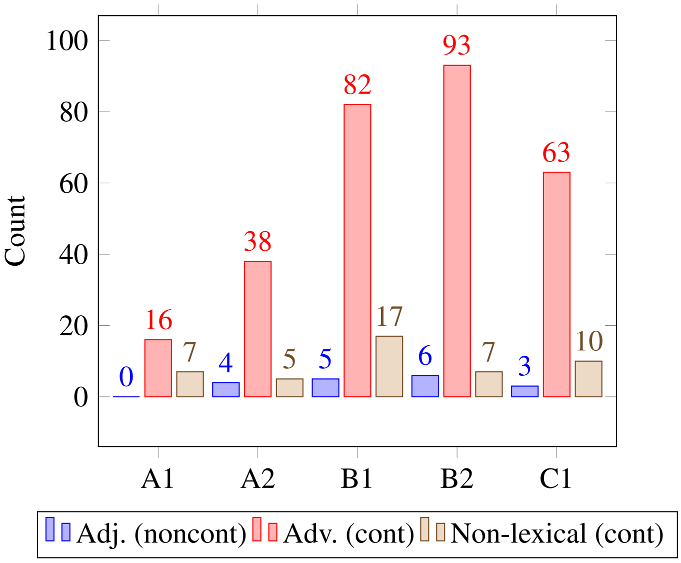


Figure 12.3: Group 3 in the crowdsourcing experiment

the automatic annotation of the corpus might not always identify the correct sense of a word or expression.

To make the experiment a learning experience and to make sure the level of difficulty was annotated in relation to a particular sense, we added definitions to all items. As far as possible we picked definitions from *Svensk ordbok* (svenska.se). When this was not possible, we used Saldo, Wiktionary, Lexin, or provided definitions of our own.

12.4 Methodology

Instead of having volunteers annotate each MWE with a target CEFR level (a task that requires in-depth knowledge of the CEFR), and following previous results showing that relative comparative judgments are easier than assigning items to a category (Lesterhuis et al. 2017), we opted to use best-worst scaling (Louviere, Flynn and Marley 2015) for the crowdsourcing task. The rationale is that language proficiency is a continuum rather than a set of discrete proficiency levels, although for practical reasons it is simplified to a set of discrete levels (Council of Europe 2018: p. 34) (cf Section 12.1). Thus, using a relative ranking method may be more fruitful than trying to classify expressions into discrete classes; in addition, operating on a continuous scale allows for more sophisticated statistical measures to be used (Ortega 2012: p. 131). Further, using best-worst scaling we get a maximum amount of information with a minimal amount of clicks from the crowdsourcers (Chrzan and Peitz 2019). Finally, such a set up requires no knowledge of the CEFR, as participants rely on their intuition when judging expressions against each other.

In best-worst scaling, one is presented with a group of items to rank and asked to rank one of the items as the “best” or the “easiest” and one of the items as the “worst” or the “hardest”. If one presents four items to the annotator to be ranked, having them choose both the easiest and the hardest out of the four expressions, it will result in 5 out of 6 possible relations.

To illustrate this further let us consider an example to show that four expressions give us six relations. Indeed, among four expressions, there exist six possible relations. Let us consider an example with expressions A, B, C and D, and let us assume that we want to know which of the expressions A, B, C and D that is the easiest and which is the hardest. This means that we thus have the following combinations of items between these expressions:

- *AB*
- *AC*
- *AD*
- *BC*
- *BD*
- *CD*

As the relations are symmetrical, we do not need to consider other combinations such as B A, as it is identical to A B; saying that B is easier than A implies that A is harder than B. With best-worst scaling, if one chooses B as the easiest expression and C as the hardest expression, we have knowledge of the following relations:

- $B < C$
- $B < A$
- $B < D$
- $C > A$
- $C > D$

The first point is self-explanatory: as we have stated that B is easiest and C is hardest, B must be easier than C. The other relations follow logically. As we have declared B as the easiest item, B must be easier than any of the other items (points 2 and 3). As we have declared C to be the hardest item, it must be harder than all other items (points 4 and 5). The only relation that we do not have information about is the relation between items A and D. However, this relation will be covered by subsequent tasks in which A and/or D occur.

In order to cover all possible combinations using best-worst scaling, we have chosen a redundancy-reducing combinatorial algorithm to calculate the minimum amount of combinations of four items needed to cover all relations in such a way as to minimize redundancy, i.e. repeating items that have already been encountered, based on Čibej et al. (In preparation).

With four items per task and 60 expressions there are 1,770 possible relations and 487,635 possible combinations. Using the redundancy-reducing combinatorial algorithm, this means that we need to have 326 tasks. Of the 1770 relations,

- 1362 (77%) are non-repetitive
- 33 with 1 relation known
- 50 with 2 relations known
- 12 with 3 relations known
- 3 with 4 relations known
- 1 with 5 relations known

Thus 77% of the relations are covered by non-repetitive combinations, while 23% of the relations are covered by partially repetitive combinations.

Finally, using best-worst scaling leads to a decrease in effort spent on the task. If one were to rank four items out of four in relation to each other, one would need at least four clicks, while best-worst scaling requires (a minimum of) two clicks, reducing the workload by half.

12.5 Experimental setup

The aim of this study is to test how one's background influences the outcome of a crowdsourcing experiment. To take a step towards that aim, we experiment with two different ways of ranking MWEs according to difficulty.

1. Intuition-based (implicit) labeling, i.e. crowdsourcing: We ask a heterogeneous group of L2 speakers of Swedish (non-experts) as well as experts (L2 Swedish professionals e.g. teachers, researchers) to rank items by taking part in a crowdsourcing experiment where we subdivide the expert group into a general L2 professional group and a group of CEFR-experts:

    - Non-experts: L2 speakers of Swedish at intermediate level (B1) or above (according to self-assessment)
    - Experts – L2 Professionals: Teachers, assessors and/or researchers of Swedish as a second language (referred to as L2 professionals)
    - Experts – CEFR experts: A separate subgroup of L2 professionals who use CEFR in their L2 Swedish practices

2. Expert judgment-based (explicit) labeling: We ask a small group of CEFR experts (teachers/researchers/assessors) to label MWE items manually for the levels at which they expect L2 learners to understand them. This annotation task is formulated in *levels* rather than *relative ordering* to resemble a real-life annotation scenario as much as possible where experts would be involved – which, however, entails some difficulties in comparison of the results.

12.5.1 Practicalities

Figure 12.4 illustrates the steps necessary to take part in the experiment. In the first step of the experiment, to comply with the GDPR (EU Commission 2016) we asked our participants for consent to use their background information for this research and to send out gift certificates.[54] At the same time, we collected information about the linguistic background as well as some other demographic variables as illustrated in Table 12.5.4.

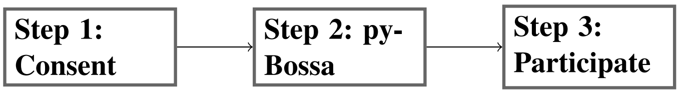


Figure 12.4: Practical steps for participants in the crowdsourcing experiment

After filling out the consent form, participants were provided with guidelines and links for the crowdsourcing part of the experiment in the form of an automated email sent to the email address specified in the consent form.[55] The guidelines were intentionally provided only in Swedish as a "selection" principle to exclude L2 speakers of lower proficiency levels.

In the second step, participants were asked to create an account on the crowdsourcing platform, with the explicit instruction to use the same email address as provided in the consent form so that we could link their background information to the crowdsourcing results. Email addresses were solely collected for this purpose and were discarded after this linking step was performed.

As a final step, participants were asked to participate in the projects proper. Each crowdsourcer was expected to complete at least 84 items out of 326 in each of the three projects, which amounted to a total time of about 30-45 minutes per project. Participants who completed at least 84 tasks per project were sent a gift (step 3 in Figure 12.4).[56]

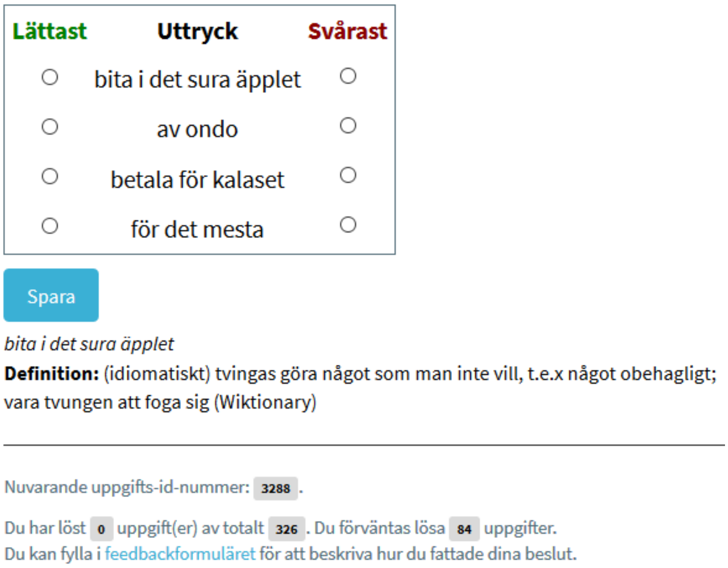
To reach the crowdsourcing population, we published announcements via e-mail, social networks, and through professional and private networks. For CEFR experts, we listed requirements with regards to their qualifications and recruited three experts on the basis of this.

We left a calendar month for the crowdsourcing experiment from the date of the first announcement, with periodic reminders to recruit broader participation. All crowdsourcers that met our requirements of the minimal contributions, were sent small gifts. Experts were paid by the hour.

[54] Expert form (Swedish only):
`https://spraakbanken.gu.se/larkalabb/mwe-cs-annotation-teacher`
Non-expert form (Swedish only):
`https://spraakbanken.gu.se/larkalabb/mwe-cs-annotation-crowd`

[55] Guidelines (in Swedish): `https://docs.google.com/document/d/1E700mnqaZ15cHr_3gXMvg0d4onm2ncMf36t-KUQuRrw/`

[56] In the later stages of the experiment when it was not possible to contribute 84 tasks in one or more projects, we relaxed the constraints for gift eligibility to ≈ 240 tasks in total.



Lättast	Uttryck	Svårast
○	bita i det sura äpplet	○
○	av ondo	○
○	betala för kalaset	○
○	för det mesta	○

bita i det sura äpplet

Definition: (idiomatiskt) tvingas göra något som man inte vill, t.e.x något obehagligt; vara tvungen att foga sig (Wiktionary)

Nuvarande uppgifts-id-nummer: 3288 .

Du har löst 0 uppgift(er) av totalt 326 . Du förväntas lösa 84 uppgifter.
Du kan fylla i feedbackformuläret för att beskriva hur du fattade dina beslut.

Figure 12.5: Example of an MWE ranking task in pyBossa (lättast = easiest, svårast = most difficult, uttryck = expression; spara = save)

12.5.2 Implementation

For the crowdsourcing experiment, we set up nine projects for the three different participant backgrounds. All projects were implemented in pyBossa, an open-source customizable framework for crowdsourcing tasks developed by SciFabric.[57] For each of our three target groups (Non-experts = L2 speakers, L2 Professionals = L2 teachers, researchers; CEFR Experts = L2 teachers, researchers, assessors with CEFR experience) we prepared three projects consisting of three sets of different MWE-types (3 participant groups x 3 projects = 9 crowdsourcing projects). In addition, we set up a tenth crowdsourcing experiment for people who did not conform to any of the three target groups or for people who wanted to see how the projects work.[58]

For each of the projects, we arranged the 60 selected items per MWE group in such a way that the crowd could vote on their relative difficulty. Figure 12.5 shows the graphical user interface used for this task, based on Čibej et al. (In preparation). In the user interface, crowdsourcers were shown four MWEs and were asked to indicate which expression they found the easiest and the hardest

[57] https://pybossa.com/

[58] Test-project (in Swedish): `https://ws.spraakbanken.gu.se/ws/tools/crowd-tasking/project/l2p_mwe_group2_other/`

to understand by using the buttons on the left and the right of the expressions, based on their own intuition. In addition, one could click on any of the four expressions to be shown a definition in case one was not sure about the meaning of an expression. The interface also showed a pyBossa-internal ID number, the number of tasks that had been completed by the crowdsourcer, the number of total tasks (326 for each project) and the expected number of tasks that each crowdsourcer should finish (84 for each project, except for the "CEFR experts" who were expected to complete all 326 tasks). Finally, we also included a link to a feedback form where crowdsourcers could indicate their reasoning about assigning the labels for easiest and hardest, or any other feedback they may wish to provide.

As additional safe-guards, we implemented checks for user errors for the following cases:

1. No value selected
2. Only one column is selected
3. Same value in both columns

As we wanted to collect the easiest and the hardest expression among a set of four expressions, it was disallowed not to provide any value (point 1), to only choose either an easiest expression or a hardest expression but not both (point 2) or to select the same expression as both the easiest and the hardest (point 3). Furthermore, as we wanted to maximize user interaction, we took care to make sure that the platform was functional and usable not only on desktop PCs but also on smaller screens such as smartphones. By doing so, people could use their smartphones wherever they were and whenever they had a minute to continue working on the tasks. Other considerations concerned the placement of the "easiest" and "hardest" columns, color schemes, and the ease of use on a smart phone. After registration in pyBossa participants could log in and continue from where they left off at any time suitable to them and on any platform (smartphone, tablet, computer). As to the number of votes per task, i.e. how many different answers were needed per task for a task to be considered complete, we set the number to 5 for L2 speakers and to 3 for L2 professionals and CEFR experts. These numbers were picked based on the estimated number of participants in the various groups. This meant that each single task in the project would have 5 respectively 3 answers (i.e. judgments about the easiest and the hardest expression) by different annotators.

We assigned the following scores to expressions: 1 for the expression that was rated as the easiest, 3 for the expression that was rated as the hardest and 2 for the two unrated expressions.

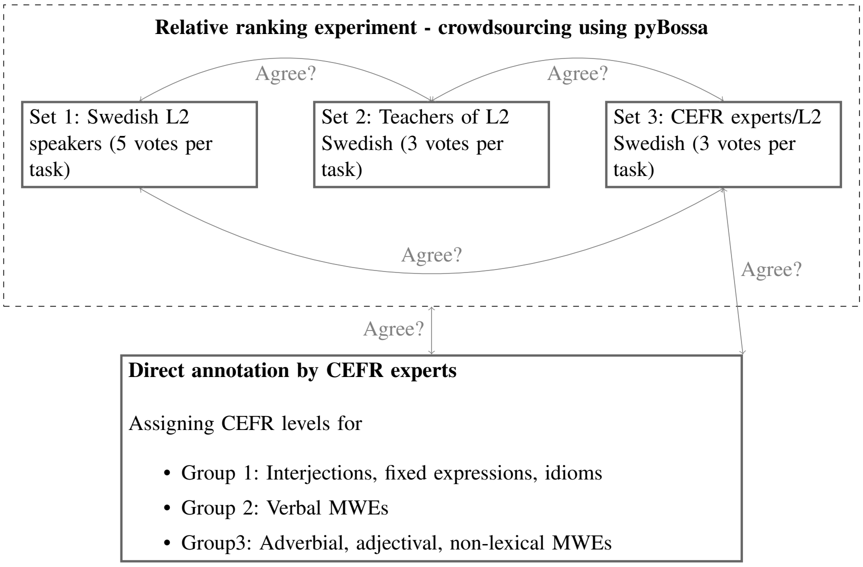

Relative ranking experiment - crowdsourcing using pyBossa
Agree?
Agree?
Set 1: Swedish L2 speakers (5 votes per task)
Set 2: Teachers of L2 Swedish (3 votes per task)
Set 3: CEFR experts/L2 Swedish (3 votes per task)
Agree?
Agree?
Agree?
Direct annotation by CEFR experts
Assigning CEFR levels for
• Group 1: Interjections, fixed expressions, idioms
• Group 2: Verbal MWEs
• Group3: Adverbial, adjectival, non-lexical MWEs

Figure 12.6: Overview of the experimental design

12.5.3 Experimental design

Figure 12.6 shows an overview of the experimental design. In the experiment, we wanted so see whether non-experts and experts agree with each other about the relative complexity of multiword expressions in the crowdsourcing experiment. We further subdivided the expert group into two to be able to compare experts' indirect judgement (crowdsourcing) to their direct (explicit) labeling, but this was only done with the small subgroup of CEFR experts who we therefore had to make sure were all well familiar with CEFR in connection with their work. Finally, we wanted to check whether individual explicit labels by the CEFR experts coincided with the group results from their implicit crowdsourcing experiment.

As indicated above we also asked our three CEFR experts to perform a direct labeling task. This meant that we asked them to go through all of the selected MWEs in a spreadsheet and decide at which CEFR level these MWEs could be expected to be understood. All three CEFR experts were asked to do the crowdsourcing experiment in pyBossa first and were only given access to the spreadsheet for direct labeling after they had completed that, to make their crowdsourcing experience as similar to that of the rest as possible. However, unlike the other participants the CEFR experts were asked to rank all items $(3 * 326)$ in all the three pyBossa projects. In the direct annotation experiment, they were asked to pick one level from a drop-down menu with A1, A2, B1, B2, C1, C2 or above for each item in a spreadsheet with all 180 MWEs.

12.5.4 Demographic information

To better understand whether and/or how the intuitions and judgments were influenced by the background of the participants, we collected information about our participants in a separate form (personal metadata). Since L2 Swedish is widely spread in Sweden and Finland, these two countries were our primary targets. However, we used social media and our personal professional networks to spread information about the experiment, which also encouraged participation from other countries. Out of 79 consent registrations in total, 50 crowdsourcers participated in the experiment, which constitutes a drop-out rate of 37%. Upon completing the crowdsourcing experiment, we could see the following participant characteristics (Table 2):

We attracted 27 L2 non-experts (L2 speakers) and 23 L2 experts, including the three CEFR experts. Sweden and Finland contributed with 22 participants each (at 44%). The first language of the contributors is dominated by Finnish (30%), but other first languages are also represented, including Swedish (20%), German (12%), Russian (8%), Spanish (4%), Arabic (4%), Hungarian (4%) and others. The population is well-educated having either a pre-doctoral university degree (60%) or a doctoral degree (36%). L2 speakers provided self-assessed levels of Swedish as B1 or above in 96% of the cases, with one outlier at the A1 level. 65% of the L2 experts have 10 years or more of experience of teaching or assessing Swedish as a second language. The age characteristics show that we attracted a rather "mature" population (78%), whereas people of 30 years and younger are less represented (22%). The gender representation is rather unbalanced (66% women versus 28% men with 6% who preferred not to answer that question), which can be due to a recruitment bias or – potentially – reflect gender representation within the areas of language learning and teaching. All in all, we have participants of various background profiles, which represents the target group for the intended output of the research.

The three CEFR experts recruited come from Finland since Finland appears to use CEFR more extensively in the teaching and assessment of L2 Swedish than Sweden does. All CEFR experts have Finnish as their L1 and represent: one L2 Swedish teacher, one L2 Swedish researcher (PhD) and one L2 Swedish assessor (PhD).

Table 12.2: Demographic variables

	Profiles	L2 speakers	L2 experts	Finland L2 speakers	Finland L2 experts	Sweden L2 speakers	Sweden L2 experts	Other L2 speakers	Other L2 experts
Total	50	27	23	9	13	13	9	5	1
Gender									
Female	33	15	18	7	10	7	8	1	-
Male	14	9	5	2	3	4	1	3	1
Other	3	3	-	-	-	2	-	1	-
Age									
16-20	5	5	-	3	-	2	-	-	-
21-30	6	2	4	2	4	-	-	-	-
31-40	15	10	5	4	3	3	2	3	-
41+	24	10	14	-	6	8	7	2	1
Education									
High school	2	2	-	-	-	2	-	-	-
University	30	16	14	9	8	5	6	2	-
PhD	18	9	9	-	5	6	3	3	1
Mother tongue									
Arabic	2	2	-	-	-	2	-	-	-
Dutch	1	1	-	-	-	-	-	1	-
English	1	1	-	-	-	1	-	-	-
Finnish	17	8	9	8	9	-	-	-	-
German	6	5	1	1	-	3	-	1	1
Hungarian	2	2	-	-	-	1	-	1	-
Luxembourgish	1	1	-	-	-	1	-	-	-

Table 12.2: Demographic variables continued

	Profiles	L2 speakers	L2 experts	Finland L2 speakers	Finland L2 experts	Sweden L2 speakers	Sweden L2 experts	Other L2 speakers	Other L2 experts
Norwegian	1	1	-	-	-	1	-	-	-
Russian	4	3	1	-	-	3	1	-	-
Serbian	2	-	2	-	1	-	1	-	-
Slovenian	1	1	-	-	-	-	-	1	-
Spanish	2	2	-	-	-	1	-	1	-
Swedish	10	-	10	-	3	-	7	-	-
L2 Swedish level									
A1	1	1	-	-	-	-	-	1	-
A2	-	-	-	-	-	-	-	-	-
B1	4	4	-	2	-	-	-	2	-
B2	7	7	-	2	-	4	-	1	-
C1	9	9	-	4	-	4	-	1	-
C2	6	6	-	1	-	5	-	-	-
Teaching years									
1-9	8	-	8	-	6	-	2	-	-
10-19	7	-	7	-	3	-	3	-	1
20+	4	-	4	-	1	-	3	-	-
other	4	-	4	-	4	-	1	-	-

12.5.5 Evaluation methodology

We use two modes of annotating items for their difficulty: crowdsourcing by non-experts and experts, and direct annotation by experts. Since direct expert annotation is a rather traditional approach, we use traditional ways of evaluating it, relying on metrics such as agreement and Spearman rank correlation. However, crowdsourcing is a new approach for this type of tasks, thus we explain how we evaluate and compare the results of the ranking (crowdsourcing) experiment below. The results are presented in Section 12.6.

For evaluation of the crowdsourcing, we project each expression onto a linear scale. The scale ranges from 1 to 3, with a minimum value of 1 if an expression was always classified as the easiest out of a set of four possible expressions and a maximum value of 3 if it was always classified as the most difficult out of a set of four possible expressions by all annotators. Otherwise, the expression *exp* is assigned the score $s(exp)$ as the mean of all assigned scores x according to the formula

$$s(exp) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{15}$$

With x_i being the i-th score assigned to exp and n being the total number of scores assigned to exp. The limits of 1 and 3 are predetermined by our way of measuring “the easiest” and “the hardest” expression with best-worst scaling (cf Section 12.4).

We sort the data according to the reverse order of $s(exp)$ and assign sequential ranks from 1 to 60 to the resulting ordering.

12.6 Results and analysis

In this section we present the results from the different experiments. First, we look into the results from the crowdsourcing experiment. After this we look into the results from the direct annotation (expert labeling) which we also compare to the rankings which this group of experts did in pyBossa. Finally, we also investigate how the number of votes influences the results and how much time is needed for the crowdsourcing experiment as opposed to direct expert annotations.

12.6.1 Linear scale

The experiments generated rich data for analysis. In this section we look at the results of the study from a quantitative point of view. For the purpose of comparison, we projected results of crowd-votings to linear scales based on the fact that each vote in a crowdsourcing task assigns scores to the items: either 1 (“easiest”), 3 (“most difficult”), or 2 for each of the two items in-between. Based on the numerical values, all items are listed in the order of their scores corresponding to the perceived degree of difficulty.

Based on that principle, we obtained one linear scale per participant group and one representing the whole population of crowdsourcers (mixed background rankings).

Table 12.3 shows the Spearman rank correlation coefficient between the three sets of MWEs and the three groups of participants. Spearman rank correlation coefficient has a range from -1 to +1 where -1 indicates a perfect negative correlation; zero indicates no correlation; and +1 indicates perfect positive correlation.

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
L2 speakers-L2 professionals	0.9509	0.9282	0.9203
L2 speakers-CEFR experts	0.9333	0.8115	0.8370
L2 professionals-CEFR experts	0.9386	0.8495	0.8579

Table 12.3: Agreement between voter groups in the crowdsourcing experiment

As can be gathered from Table 12.3, the highest correlations can be found between non-experts (here meaning L2 speakers/learners) and the general group of “L2 professionals” (including teachers, assessors, researchers) across all of the three MWE groups, while the correlations between non-experts (L2 speakers) and ”CEFR-experts” (i.e. the subgroup of three L2 professionals) are the lowest among all the three MWE groups. We can thus say that non-experts (L2 speakers) and experts (L2 professionals) in our experiment agree very well on the relative difficulty of MWEs, followed by L2 professionals and CEFR experts, while L2 speakers and CEFR experts tend to agree to a lesser extent. Despite these marginal fluctuations, we can see strong correlations between all of the tested target groups across all the three sets of tested MWEs. This indicates that intuitions about the difficulty of MWEs are more or less shared across all tested groups, despite the differences in background and professional competence. It seems that we can confirm that non-experts – that is, L2 speakers lacking expertise and competence in a subject (e.g. language assessment) – can be seen as on par with experts for tasks requiring high competence, something that has also been shown in approaches in citizen science (Kullenberg and Kasperowski 2016).

To get an insight into how well individuals can agree on crowdsourcing tasks we looked at the three CEFR experts in our experiment who completed the full sets of tasks in all of the three pyBossa projects. Table 12.4 shows the Spearman rank correlation based on their individual linear scales calculated from the crowdsourced data. As can be seen from Table 12.4, annotators 1 and 2 tend to agree the most, while annotators 1 and 3 tend to agree the least, with annotators 2 and 3 falling in-between. This might be a result of their different backgrounds and how often they use CEFR explicitly. The more voters we have, the less bias there is in the resulting data (e.g. Snow et al. 2008).

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
CEFR experts 1 and 2	0.8130	0.8581	0.7735
CEFR experts 1 and 3	0.7733	0.5788	0.6988
CEFR experts 2 and 3	0.7964	0.6236	0.7026

Table 12.4: Inter-annotator agreement for CEFR experts in the crowdsourcing experiment calculated with Spearman rank correlation coefficient

12.6.2 Expert labeling

If we look closer at the simple and extended percentage agreement between the CEFR expert annotators in the explicit (interchangeably called ‘direct’) labeling experiment, we can see that agreement is generally quite low for simple agreement (Tolerance 0 in Table 12.5). With a tolerance of zero, one counts exact agreement between the annotators (e.g. the same item has been assigned to the same CEFR level). However, if one relaxes the tolerance level to 1 (extended percentage agreement), meaning that positive agreement also includes cases where annotators differed by only one level (e.g. one annotator said the item was A2 while another annotator said the item was B1), we can see that agreement drastically improves, as illustrated in Table 12.5.

	Group 1 (interjections)	Group 2 (verbs)	Group 3 (adverbs)
Tolerance 0	15.00	21.70	13.30
Tolerance 1	61.70	58.30	65.00

Table 12.5: Agreement between CEFR experts in a direct labeling experiment in percent

In general, this gives us a picture that expert judgments are not ideal and that reaching an exact agreement between them is possibly an unattainable target, which also confirms the results from essay evaluation according to the

CEFR-scale as presented in e.g. Díez-Bedmar (2012). Given that direct labeling is a subjective and cognitively challenging task, more opinions than one are required (cf Snow et al. 2008; Carlsen 2012).The MWEs in the experiments are de-contextualized which might further complicate decisions. This speaks in favor of assuming tolerance level 1 since the assigned levels describe a continuum of proficiency rather than strict categories (Council of Europe 2018: p. 34). A hypothesis in connection to this is that disagreement outside tolerance 1 may indicate items that are on the periphery of the lower CEFR level, while items within tolerance 1 constitute the core vocabulary on the lower level. This is something to be explored in future research.

Results of agreement between the explicit ranking of each individual expert and their own individual implicit judgment from the crowdsourcing experiment based on a comparison of the linear scales show mixed results (Table 12.6).

	Group 1 (interjections)	Group 2 (verbs)	Group 3 (adverbs)
Expert 1	0.9095	0.9280	0.8935
Expert 2	0.8483	0.6147	0.7299
Expert 3	0.8010	0.5248	0.5540

Table 12.6: Spearman rank correlation coefficients for intra-annotator agreement between implicit and explicit modes of annotation

Expert 1 is very consistent in both annotation methods, and all annotators seem to agree with themselves most for MWE group 1, while other agreements are lower. This could indicate that expert 1 is the one with the most experience with working with CEFR-levels. The inconsistency of the results for the same expert indicates that the expert reasons differently when using different methods, and that the way of reasoning influences the results. It has been previously shown that explicit scoring is more subjective and cognitively demanding than assessing by comparing two samples to each other (Lesterhuis et al. 2017), which also seems to be confirmed in this experiment. This indicates that we should not compare the two types of annotation and that expert judgment can only give reliable annotation if a reasonably large number of experts is used to counter-balance a potential subjective bias. How large a number constitutes a "reasonable amount" is still an open question.

12.6.3 Number of votes

Aker et al. (2012) found that using one set of non-expert results (results from different annotators) outperformed using one single non-expert's results, as the diversity of the crowd might cancel a high bias present in a single annotator. In order to see how the number of votes influences the results, we randomly selected votes for the sample sizes 1, 2 and 3 (for the non-expert crowd, for which we collected 5 votes) and derived the linear scales, for each group separately as well as a randomly sampled mixed version over all three groups ('Mixed' in Tables 12.7 and 12.8). We then compared the linear scales of the different sample sizes to the linear scale derived from the full set votes (3 for experts and 5 for non-experts; for the mixed group we calculated the target linear scale from a random sample of three votes from both experts and non-experts), meaning that we compare for example the linear scale for non-experts derived from a single vote versus the linear scale for non-experts derived from 5 votes; the linear scale for non-experts derived from two votes versus the linear scale for non-experts derived from 5 votes; or the linear scale derived from randomly sampling two votes from both experts and non-experts versus the linear scale derived from randomly sampling three votes over all groups.

In order to quantify the differences between the scales, we used the *out-of-place metric m_{oop}* (Cavnar et al. 1994). This is a straightforward metric that measures the difference between two ranked lists and quantifies the difference. The reason for choosing this metric over rank correlation measures is that Spearman's correlation coefficient was very high and had similar values across all comparisons (see Table 12.7). While a high correlation is a positive result in itself, it does not allow for a detailed analysis. We surmise that using m_{oop} may give a more tangible result. It is formalized as shown in (16)

$$m_{oop} = \sum_{i=1}^{n} (|r(x_i, l_1) - r(x_i, l_2)|) \tag{16}$$

with n being the number of items in the lists (the lists to be compared are of the same length in our case), x_i being the i-th item, $r(x_i, l_1)$ being the rank of x_i in the first list and $r(x_i, l_2)$ being the rank of x_i in the second list. To illustrate this, let us consider two lists l_1 and l_2 both containing the expression A, B, C and D, but at different ranks. Figure 12.7 shows a hypothetical scenario. In order to obtain m_{oop}, one first calculates the difference in ranks between the expressions, then sums up the differences. Thus, in this example, we would have $m_{oop} = 1 + 2 + 0 + 3 = 6$. We also calculate how many items are at the exact same rank in both lists (out of 60 total).

We find that each of the sub-sampled lists compared to the full-vote list yields high Spearman rank correlation coefficients, with Spearman's ρ varying from $\rho = 0.941, p = 5^{-29}$ to $\rho = 0.997, p = 3^{-66}$. As can be gathered from Table 12.7, group 1 (interjections) shows the least amount of divergence among all three MWE groups, but also among the different crowds. Further, it can

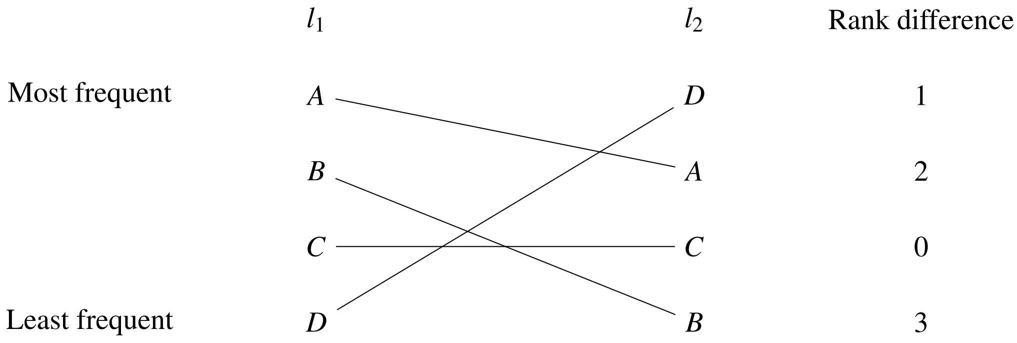


Figure 12.7: Out-of-place metric illustration

be observed that sampling over all three crowd groups produces more stable results than within-group sampling.

A more qualitative analysis reveals that for group 1 (interjections etc.) for non-experts with one vote, the hardest and easiest item is the same as with five votes, whereas with two votes, the two easiest and the three hardest are the same as with five votes. For CEFR experts with one vote, the three easiest items are the same as with three votes, whereas with two votes, the two hardest items are also the same as with five votes. For L2 professionals with one vote, the easiest item is the same as with three votes whereas with two votes, also the two hardest items are the same as with three votes. However, many of the rank differences are small, i.e. the two hardest items for group 1 (interjections etc) for L2 professionals with one vote are the reverse order of two and three votes. If one were to start from a truly unlabeled set of items without indications of level, or the number of different levels present in the data, one can only rely on relative ranks. These results indicate a certain stability when it comes to the extremes of the scale, i.e. which items are easiest and which items are hardest.

In order to account for small differences in ranks, we also compute how many items are "at the same rank" when counting as the same rank items within a difference of d, with d varying from 1 to 5 (n.b. $d = 0$ is equivalent to the same rank, column 'Same' in Table 12.7). If we take as an example the ranking in Figure 12.7, at $d = 1$, one would count as being of equal rank the item A (in addition to item C), as the rank difference is 1. At $d = 3$, one would also consider as being of equal rank items B and D, as they are below or equal to 3. Table 12.8 shows the results; we repeat $d = 0$ for comparison purposes.

It can be said that the lists derived from a sub-sample of votes are different from the lists derived from all votes. However, when relaxing the notion of "equivalence" as has been done by varying d, one can see that the difference

MWE group	Crowd	Sample size	m_{oop}	ρ	Same rank
Interj.	L2 sp.	1	150	0.98	8
		2	112	0.98	15
		3	102	0.99	16
	L2 prof.	1	160	0.97	16
		2	82	0.99	15
	CEFR exp.	1	114	0.98	18
		2	80	0.99	18
	Mixed	1	114	0.98	14
		2	78	0.99	23
Verbs	L2 sp.	1	256	0.94	6
		2	172	0.97	6
		3	114	0.98	18
	L2 prof.	1	196	0.97	8
		2	90	0.99	18
	CEFR exp.	1	200	0.96	7
		2	120	0.98	14
	Mixed	1	138	0.98	11
		2	70	0.99	26
Adverbs	L2 sp.	1	254	0.95	4
		2	154	0.98	12
		3	110	0.98	15
	L2 prof.	1	244	0.94	8
		2	132	0.98	14
	CEFR exp.	1	126	0.98	14
		2	106	0.99	13
	Mixed	1	128	0.98	14
		2	54	0.99	25

Table 12.7: Out-of-place calculations, Spearman's ρ and same rank number for different numbers of votes

is not as big as one might think at first. At $d = 2$, which means deviations of two ranks (out of 60) or less are counted as equal, around 84% of the lists are "equal" to the lists derived from full votes for the aggregated versions (82% for interjections, 80% for verbs and 90% for adverbs). Again, it can be observed that sampling over the whole crowd produces more stable results than sampling

MWE group	Crowd	Sample size	*d*					
			0	1	2	3	4	5
Interj.	L2 sp.	1	8	21	34	45	49	55
		2	15	30	43	48	51	56
		3	16	34	44	51	56	58
	L2 prof.	1	16	30	38	43	47	52
		2	15	38	53	56	58	59
	CEFR exp.	1	18	35	42	46	53	57
		2	18	42	49	55	57	59
	Mixed	1	14	30	44	49	55	57
		2	23	37	49	54	59	60
Verbs	L2 sp.	1	6	10	27	36	42	45
		2	6	26	37	42	48	52
		3	18	34	43	50	54	55
	L2 prof.	1	8	17	24	33	43	50
		2	18	37	47	54	56	58
	CEFR exp.	1	7	20	32	35	44	49
		2	14	26	38	50	56	57
	Mixed	1	11	22	36	45	52	57
		2	26	44	48	55	58	59
Adverbs	L2 sp.	1	4	16	27	31	37	40
		2	12	26	33	41	50	53
		3	15	36	46	53	54	55
	L2 prof.	1	8	17	29	36	41	46
		2	11	30	41	48	51	55
	CEFR exp.	1	14	31	44	48	51	54
		2	13	33	44	51	56	58
	Mixed	1	14	32	44	49	49	54
		2	25	48	54	59	60	60

Table 12.8: Effect of different *d* values

within a group. It can further be observed that the aggregated votes tend to be on par with expert judgments, if not surpassing them.

	Group 1	min	max	Group 2	min	max	Group 3	min	max
L2 speakers	36	3	164	38	6	260	44	3	227
L2 prof.	41	13	43	26	14	44	24	14	44
CEFR exp.	32	28	39	34	23	41	36	21	60
Average	36			32			34		

Table 12.9: Average number of seconds per task and group

12.6.4 Time investment

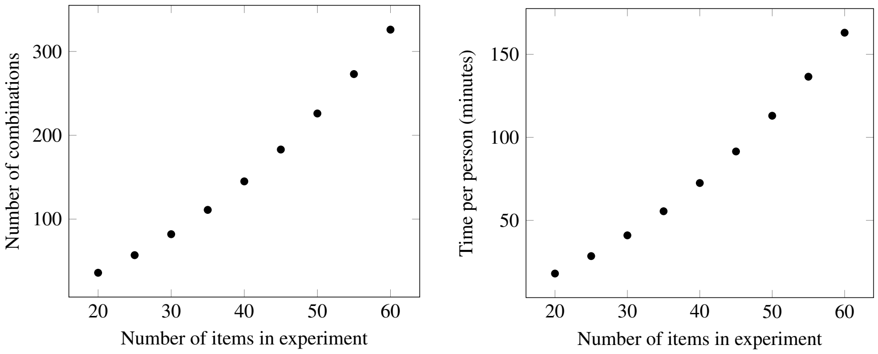
Table 12.9 shows the average time taken per crowd background and MWE group. Despite the presence of outliers in the non-expert crowd data, crowdsourcing in a best-worst scaling scenario takes on average 30-40 seconds per task. To rank 60 items presented through 326 tasks with one vote would claim ≈ 2,5-3 hours. Rankings do not seem to change drastically after the first three votes are collected, so the minimal time investment for 3 votes are estimated to approximately 8-9 hours for one project.

Table 12.10 shows the comparison between observed times in the crowdsourcing project and reported times for direct annotation by the CEFR experts. It should be noted that for expert direct annotation, the times indicated in Table 12.10 are approximated by dividing the reported time needed to finish all three lists by three. It should also be borne in mind that experts went through all 326 tasks per project. It can be observed that direct expert annotation claims 15-90 minutes per project. This is at least five times as fast as the crowdsourcing experiment.

	Group 1 (interjections)		Group 2 (verbs)		Group 3 (adverbs)	
	CS	Direct	CS	Direct	CS	Direct
Expert 1	217	≈ 90	225	≈ 90	148	≈ 90
Expert 2	155	≈ 15	129	≈ 15	117	≈ 15
Expert 3	156	≈ 20	199	≈ 20	327	≈ 20

Table 12.10: Observed (crowdsourcing, CS) and reported (direct) times for experts for the two modes of annotation, in minutes

However, reliability and consistency of a (direct) labeling depend to a larger extent upon what kind of ranking scale annotators are offered and what their backgrounds are, and the effects are difficult to account for (cf. O'Muircheartaigh, Gaskell and Wright (1995)). It is easy to fall victim to a flawed design, inexperienced annotators or face problems hiring annotators, and the cognitive load



(a) Number of combinations *(b)* Time per person

Figure 12.8: Number of combinations and time per person with varying number of items

of such an exercise is higher than in a crowdsourcing set-up (e.g. Lesterhuis et al. (2017)).

The time required to complete such a crowdsourcing experiment depends on the number of items that make up the experiment. Thus, for 20 items and 4 items per task, if one calculates with a mean response time of 30 seconds per task, it would take three crowdsourcers approximately 18 minutes each if one were to collect three votes per task.[59] Figure 12.8a shows the number of combinations in the experiment when varying the number of items from 20 to 60 in increments of 5. Figure 12.8b shows the amount of time it would take each person on average to complete the project under the above constraints. It can be noted that there seems to be a curvilinear relationship between the number of items and the number of combinations; this relationship would be exponential if it were not for the redundancy-reducing algorithm used. If one looks at the time per person in relation to the number of items, the relation seems to mimic the relation between number of items and number of combinations. Further, doubling the number of crowdsourcers (from 3 to 6) leads to a reduction of time per crowdsourcer by half: for 20 items and 4 items per task with a mean response time of 30 seconds per task, it would take six crowdsourcers approximately 9 minutes each if one were to collect three votes per task.

[59] If one wants to collect three votes per task, the minimum required number of participants is three, as no (registered) participant will be shown the same task twice.

12.7 Discussion

Among the burning questions in emerging crowdsourcing projects – within the domain of language learning – the three methodological questions below remain the most important at the current stage of development:

1. Who can be the crowd – with regards to the *background* of crowdsourcers?
2. How can *reliable* annotations be achieved with regards to design, number of answers and number of contributors? and
3. How should the *the results be interpreted* with regards to both research and practice?

The biggest gap that we have tried to fill with this study concerns the first (1) question, i.e. whether crowdsourcing as a method in language learning – within a limited domain of L2 resource annotation – could be used without explicit control for the background of the crowd.

Our results convincingly show that non-experts can perform on par with experts. We have seen that crowds with different backgrounds agree very well with each other, in comparison to previous research where CEFR raters of essays have often reached fairly low agreement. In fact, a mixed background crowd reaches "average" rankings faster. Note here that these conclusions are true of annotation carried out in a *comparative judgment* or *best-worst scaling* setting whereas previous work on essay rating has been done based on scales such as the CEFR-scale similar to our direct-labeling experiment with the CEFR-experts. To further confirm our findings, similar experiments need to be repeated for other languages, for other types of problems (e.g. annotation of texts for difficulty/readability), and for other sub-problems of a given problem (e.g. annotation of single vocabulary items for difficulty). Similar conclusions have been made in projects within citizen science, among others by Kullenberg and Kasperowski (2016) where the experimental setup was not necessarily "comparative" in character. This leaves room for further experimentation.

In relation to question (2) the *reliability* of annotations, we have seen how the design of an annotation task influences the results. Clearly, a more traditional method of annotation – using expert judgments – compares negatively to crowdsourced comparative judgments/best-worst scaling rankings. We have seen that experts do not agree with themselves when using comparative judgments versus categorical judgments, whereas the comparative judgment setting leads to homogeneous results between all groups of crowdsourcers regardless of their background, as shown in Section 12.2. According to Hovy and Lavid

(2010) reliability of annotation of language resources has two types of major consequences, namely theoretical ones for shaping, extending and re-defining theories, and practical ones for use in the classrooms, but also in teaching and assessment practices. Unreliably annotated data can lead to biased – if not erroneous – theoretical conclusions and generalizations, as well as influence teaching and assessing practices in unwanted ways, as discussed among others in Carlsen (2012).

The above-stated *theory–practice* dichotomy can be traced down to the proficiency dimension of the MWE items in our experiment. On the one hand, the ranked list represents Multi-Word Expressions according to their difficulty from the learner's point of view and can thus be assumed to reflect stepwise development of their phraseological competence, which is of immediate interest to theoretical studies on L2 development. On the other hand, the scale represents perceptions of L2 professionals and – hypothetically – reflects their reasoning about what to teach/assess and in which order to do so based on their practical experience from teaching and assessing language learners, and has an immediate relevance for practical applications in "real life", including use in automatic solutions for language learning. It is very encouraging to see that the two perspectives (theoretical-developmental and practical) produce similar results and are so much in harmony with each other. However, this harmony can be observed only as long as we view vocabulary development as a continuum as opposed to groups of items belonging to one of a number of categorical proficiency levels.

In fact, both dimensions – theoretical and practical – are equally important. To understand how to teach and what to teach (practical dimension), we need to understand how learning is happening and (among others) observe which linguistic and cognitive aspects develop and in which order. While the produced scales give us material to study development of phraseology from a theoretical point of view, it is not obvious how to apply these scales to practical use (question (3) above) in teaching, assessment and Intelligent CALL, where categorical representations of proficiency are more customary and readily applicable. There are no indications in our crowdsourcing results as to where to draw the line between one level of proficiency and the other. We are not unique in facing these troubles, even though in other areas it can be a vice versa case:

> A weakness in this line of work is that SLA researchers have most often chosen to treat proficiency as a categorical variable and then have assessed mean differences in complexity values across proficiency groupings. Yet, this practice of converting interval variables (i.e. individual proficiency scores of some kind) into categorical ones (i.e. participants grouped by nominal proficiency levels) has always been criticized by

> statisticians because it discards much useful information. More specifically, it does away with the variance of continuous scores and leads to unreliability and increased likelihood of Type II errors (e.g. Troncoso Skidmore and Thompson 2010), that is, the problem of failing to detect a difference, relationship, or effect that is in fact present because of some psychometric methodological problem, such as lack of power or (in the case at hand) lack of variance in the observations. It would be profitable in future work, therefore, to accumulate evidence from designs where both complexity and proficiency are treated as interval scales.
>
> (Ortega 2012: p. 131)

This is a current challenge that needs to be addressed in the future (e.g. Paquot, Naets and Gries 2020). Proficiency levels are always rather arbitrary (Hulstijn, Alderson and Schoonen 2010) as is also noted by the authors of CEFR (Council of Europe 2001) who caution that “any attempt to establish ‘levels’ of proficiency is to some extent arbitrary, as it is in any area of knowledge or skill. However, for practical purposes it is useful to set up a scale of defined levels to segment the learning process for the purposes of curriculum design, qualifying examinations, etc.” (p. 17). To summarize this part of the discussion, we view our results as a strong argument for treating vocabulary development as a continuum, while we also recognize the need to establish ways to partition vocabulary by levels of proficiency where these items can be taught.

On the practical side of crowdsourcing, our results show that a good and reliable agreement within a mixed crowd can be reached with two to three votes per task by at least three different voters. Considering these results, it might be interesting to use the same methodology for essay grading, especially since results from various experiments which have looked at inter-rater agreement in marking essays according to categorical proficiency levels have been less promising (cf Carlsen 2012; Díez-Bedmar 2012).

One of the limitations of the current setup lies in the use of the combinatorial algorithm which we apply to calculate the task pairings. As stated, we only achieve 77% *non-redundant* combinations, which means that certain pairs of expressions are included more than once and thus get more votes than other combinations, which might skew the picture. More involved statistical methods such as balanced incomplete block design (BIBD) (Yates 1936) can be used to circumvent this problem. However, such methods impose hard constraints on the number of items and the number of items per task and not all combinations of number of items and items per task are able to satisfy these constraints. To the best of our knowledge, there exists no solution to the BIBD constraints for 60 items with 4 items per task.

Ideally, results from an experiment should be *generalizable* even to other uses, for example be applicable to prospective automatic solutions for language learning.

The two methods of annotation – crowdsourcing by unknown crowd versus annotation by approved experts – have different dimensions of pros and cons. Here we have seen that time versus reliability can outweigh each other. In addition, one needs to consider that when using crowdsourcing, one has little control over the participant group and the time. Hence, neither method is superior on all accounts, but both are appropriate as long as one is aware of their weaknesses and strengths. If one is able to pay CEFR experts, one may get faster results. However, as seen in this study, one would need a large number of experts to reach consensus. Thus, expert knowledge can be fast and reliable *if* a large enough number of experts is consulted, to counteract the bias of individual subjective opinions, but it is also expensive. If one does not have access to experts for various reasons, one can use crowdsourcing as an alternative to derive a relative ranking of expressions. The resulting ranking is similar regardless of whether one uses non-experts or experts, thus one may be able to realize such an experiment with non-experts only. In contrast to using experts for direct annotation, crowdsourcing is cheap however it takes longer time, both regarding the implementation and the actual crowdsourcing phase. Furthermore, with the set up we have chosen, one does not get concrete CEFR levels but rather a relative ranking. This data can, however, potentially be partitioned into more or less discrete proficiency levels by various techniques, should one desire to do so. The exploration and experimentation in this direction is future work.

12.8 Conclusion

In this study, we asked whether it influences the results in a crowdsourcing experiment aimed at ranking MWEs by difficulty if crowdsourcers are experts (L2 Swedish professionals) or non-experts (L2 Swedish learners / speakers). We set up different crowdsourcing experiments for the different target groups so as to be able to compare the results of different groups. The presented experiment suggests that it does not matter for this type of experiment if the crowdsourcer is an L2 speaker or an L2 professional, as the results produced by L2 speakers of Swedish, teachers of Swedish and CEFR experts are highly correlated.

Furthermore, we explored how the number of votes influences the results and we found that with only two votes, the difference in results on a scale 1-60 is insignificant in comparison to three votes. Additionally, we found that sampling from a mixed-background group tends to produce more stable results. Indeed, using a mixed crowd produces results similar to results obtained from only expert annotations. This finding can further speed up crowdsourcing projects, since one can gather data with only one experiment instead of having to set up three distinct experiments for each target background. We also found that L2 *proficiency*, as measured by L2 professionals, does seem to correlate with L2 *development*, collected through intuitive judgments by L2 speakers.

These findings suggest that crowdsourcing might be a viable method to create a ranking of expressions by difficulty even in the absence of gold standard data. Our results suggest that there is a strong incentive in exploring crowdsourcing for other languages (if getting a scale is sufficient). For any new language and new item combination, we would suggest that the best-worst method be applied. There are reasons to believe that having strong “anchor words” for levels, i.e. words for which one knows the level with reasonable certainty, among the data can help create clusters around those with suggestions where to draw the line between one level and another, if there is a need for the pedagogical, assessment, CALL or other uses.

Future studies could investigate whether the same methodology produces the same results when applied to for example single word expressions or essays. Another direction for future research might be how to partition an unordered, unlabeled set of expressions into different proficiency levels based for example on clustering results. This might be achieved by adding certain *anchor* expressions to the experiment, i.e. expressions of which one knows with a sufficient degree of certainty their true label (i.e. target level). As a possible starting point, one could take the easiest and the hardest expressions overall from a ranking experiment such as the one presented, as the agreement at the extremes (very easy and very hard expressions) tends to be much higher than in the middle of the scale. Further, one might want to investigate how core and peripheral vocabulary can be identified based on different kinds of annotations.

13 Semi-automatic lexicographic data enrichment

This publication is discussed in section 7.1.4.

This chapter is a postprint version of the following publication:

Alfter, David, Therese Lindström Tiedemann, Elena Volodina. 2019. LEGATO: A flexible lexicographic annotation tool. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa*. Turku, 2019 . Linköping University Electronic Press.

Abstract

This article reports an ongoing project aimed at analyzing lexical and grammatical competences of Swedish as a Second language (L2). To facilitate lexical analysis, we need access to linguistic information about relevant vocabulary that L2 learners can use and understand. The focus of the current article is on the lexical annotation of the vocabulary scope for a range of lexicographical aspects, such as morphological analysis, valency, types of multi-word units, etc. We perform parts of the analysis automatically, and other parts manually. The rationale behind this is that where there is no possibility to add information automatically, manual effort needs to be added. To facilitate the latter, a tool LEGATO has been designed, implemented and currently put to active testing.

13.1 Introduction

Lexical competence has been acknowledged as one of the most important aspects of language learning (e.g. Singleton 1995; Milton 2013; Laufer and Sim 1985). Some claim that we need to understand 95–98% of the words in a text to manage reading comprehension tasks (cf. Laufer and Ravenhorst-Kalovski

2010; Nation 2006; Hsueh-Chao and Nation 2000). It has also been observed that vocabulary is actively taught at all levels of L2 proficiency courses with a tendency to be dominating at more advanced levels in comparison to other linguistic skills, see for example findings from a course book corpus COCTAILL (Volodina et al. 2014a: p.140). Lexical features have also been found to be one of the best predictors in text classification studies (e.g. Pilán and Volodina 2018; Xia, Kochmar and Briscoe 2016; Vajjala and Meurers 2012) with important implications to the area of educational NLP. Deciding on which vocabulary to use and include is thus an important part of teaching a foreign language, in designing course materials and tests. In theoretical descriptions of L2 acquisition, lexical knowledge was previously "side-lined" according to Milton, but within academic circles its place has been "significantly revised" and received an increasing amount of interest over recent decades (Milton 2013).

There are multiple characteristics of vocabulary that are interesting from the point of view of both theoretical analyses, as well as for pedagogical and NLP-based applications. Such characteristics include, among others, vocabulary size & breadth (e.g. Nation and Meara 2010; Milton 2013), corpus frequency (Dürlich and François 2018; François et al. 2016), word family relations (Bauer and Nation 1993), syllable structure, morphological characteristics, semantic relations, topical domain categorization (Alfter and Volodina 2018), and many others (e.g. Capel 2010, 2012).

While frequency information comes from corpora, most linguistic characteristics are non-trivial to acquire by automatic methods and require either manual effort or access to manually prepared resources – lexicons being the most extensive and reliable sources for that. However, dictionaries and lexicons are often proprietary resources (e.g. Sköldberg et al. 2019), which complicates automatic lexicon enrichment. Among freely available lexicons for Swedish, we can name Saldo (Borin, Forsberg and Lönngren 2013), Swesaurus (Borin and Forsberg 2014), Lexin (Hult, Malmgren and Sköldberg 2010) and a few other resources provided through Språkbanken's infrastructure Karp (Borin et al. 2012), although, even there many aspects of vocabulary are not documented, e.g. the transitivity of verbs, the morphological structure of the words (root, prefix, suffix) or the topical domain of the words.

To circumvent the problem of access to the information that may prove crucial in the context of the current project for the three outlined areas of application (theoretical studies, pedagogical studies/applied linguistics and educational NLP), we have initiated semi-automatic annotation of learner-relevant vocabulary interlinking available resources with manual controls of those, and adding missing aspects manually. The work is ongoing, and below we present the reasoning around this annotation process and the main components of the system that facilitate that.

Aspect	**Explanation / choices**	**Mode**	**Resources for auto-enrichment**
1 Adj/adv structure	comparisons: periphr.: *(mer/mest) entusiastisk*; morph.: *vacker-vackrare-vackrast*; irreg.: *god-bra-bäst*	A-M[2]	Saldo-Morphology
2 Adj declension	decl. 1 & 2, irregular, indeclinable	A-M	Saldo-Morphology
3 Morphology 1	word analysis for morphemes: *oändlig*: prefix:*o-*; root:-*änd-*; suffix:-*lig*	M[3]	
4 Morphology 2	word-building: root, compound, derivation, suppletion, lexicalized, MWE[1]	M	
5 MWE type	taxonomy under development	M	
6 Nom declension	decl. 1-6, extra	A[4]	Saldo-Morphology
7 Nom gender	common, neuter, both, N/A	A	Saldo-Morphology
8 Nom type	abstract–concrete, (un)countable, (non)collective, (in)animate, proper name, unit of measurement	M	
9 Register	neutral, formal, informal, sensitive	M	
10 Synonyms	free input, same word class	A-M	Swesaurus

Continued on next page

Table 13.1 – *Continued from previous page*

11 Topics/domains	general + 40 CEFR-related topics[5]	A-M	Lexin, COCTAILL
12 Transitivity	(in-, di-)transitive, N/A	A-M	SAOL (under negotiation)
13 Verb category	lexical, modal, auxiliary, copula, reciprocal, deponent	M	
14 Verb conjugation	conjugations 1-4, irregular, N/A	A	Saldo-Morphology
15 Verb action type	motion, state, punctual, process[6]	M	

Table 13.1: Linguistic aspects added to SenSVALex and SenSweLLex items [1]MWE = Multi-Word Entity; [2]Manual based on automatically enriched input; [3]Manual; [4]Automatic; [5]Topics come from the CEFR document (Council of Europe 2001), COCTAILL corpus (Volodina et al. 2014a), and some other resources; [6]Incl. limited and unlimited process verbs

13.2 Second language profiles project

In the current project, *Development of lexical and grammatical competences in immigrant Swedish* funded by Riksbankens Jubileumsfond, the main aim is to provide an extensive description of the lexical and grammatical competence learners of L2 Swedish possess at each CEFR[60] level, and to explore the relation between the receptive and productive scopes. The exploration of the grammatical and lexical aspects of L2 proficiency is performed based on two corpora, COCTAILL (Volodina et al. 2014a), a corpus of course books used in teaching L2 Swedish and the SweLL-pilot (Volodina et al. 2016a), a corpus of L2 Swedish essays. The corpora are automatically processed using the SPARV pipeline (Borin et al. 2016), and include, e.g., tokenization, lemmatization, POS-tagging, dependency parsing, and word sense disambiguation.

[60] CEFR = Common European Framework of Reference (Council of Europe 2001)

13.3 LEGATO tool

LEGATO[61] - **LE**xico**G**raphic **A**nnotation **TO**ol - is a web-based graphical user interface that allows for manual annotation of different lexicographic levels, e.g. morphological structure (root, affix etc), topic, transitivity, type of verb (e.g. auxiliary, motion verb), etc. The interface shows a lemgram for a given word sense, the part of speech and the CEFR level, as well as the Saldo sense and the primary and secondary sense descriptors used in Saldo (Borin, Forsberg and Lönngren 2013), and up to three example sentences taken from the COCTAILL corpus. If there are fewer than three sentences available at the target CEFR level, the maximum number of sentences found is shown. It also features search, filter and skip functionalities as well as external links to other information sources such as Karp (Ahlberg et al. 2016); SAOL, SO & SAOB via svenska.se (Malmgren 2014; Petzell 2017); and the Swedish Academy's Grammar (SAG, the main grammar of the Swedish language) (Teleman, Hellberg and Andersson 1999). Figure 13.1 shows the user interface for the annotation of *nominal type* category.

13.3.1 Data for lexicographic annotation

For lexical analysis, we generate word lists (SenSVALex and SenSweLLex) based on senses from the two linguistically annotated corpora, both lists being successors of the lemgram-based ones from the same corpora (François et al. 2016; Volodina et al. 2016b). The lists contain accompanying frequency information per CEFR level according to the level assigned to the texts/essays where they first appear. In practical terms, the task of preparing a resource for lexical studies involves:
1. labeling all items for their "target" level of proficiency – that is, the level at which the item is expected to be understood (receptive list) or actively used (productive list). The CEFR level of each item is approximated as the first level at which the item appears, i.e. the level would be B2 for entry X if it was first observed at level B2 (cf. Gala, François and Fairon 2013; Gala et al. 2014; Alfter and Volodina 2018).
2. interlinking items with other resources for enrichment, e.g. adding information on adjective declension
3. manually controlling the previous step for a subset of items to estimate the quality
4. setting up an annotation environment for adding missing information.

[61] `https://spraakbanken.gu.se/larkalabb/legato`; user "test" for testing purposes

While (1) above has been partially addressed by Alfter et al. (2016) and Alfter and Volodina (2018), steps (2–4) are described shortly in the sections below.

13.3.2 Automatic enrichment

An overview of linguistic aspects annotated using LEGATO is provided in Table 13.1. All aspects are kept as close as possible to the terminology and the description of Swedish grammar in SAG (Teleman, Hellberg and Andersson 1999). A subset of those aspects, marked as *A* or *A-M* in Table 13.1 (column "Mode") are annotated automatically using a range of available resources mentioned in the column "Resources for auto-enrichment". Other aspects are added manually (*M*) following guidelines[62] explaining choices and argumentation based on SAG and other work on the Swedish language and linguistic description in general.

To augment SenSVALex & SenSweLLex, we use different resources. Besides the information already present in these lists (word senses, Saldo descriptors, automatically derived CEFR level, part-of-speech), we use Saldo / Saldo morphology (Borin, Forsberg and Lönngren 2013), Swesaurus (Borin and Forsberg 2014), Lexin (Hult, Malmgren and Sköldberg 2010) and potentially SAOL (Malmgren 2014) to enrich the lists.

Saldo morphology is used to add nominal gender, nominal declension and verbal conjugation. Adjectival declension and adjectival (and adverbial) structure are derived from the comparative and superlative forms given in Saldo morphology and checked manually. Synonyms are added using Swesaurus. Other named resources are planned for enriching topics and transitivity patterns. The remaining categories are left to be manually annotated.

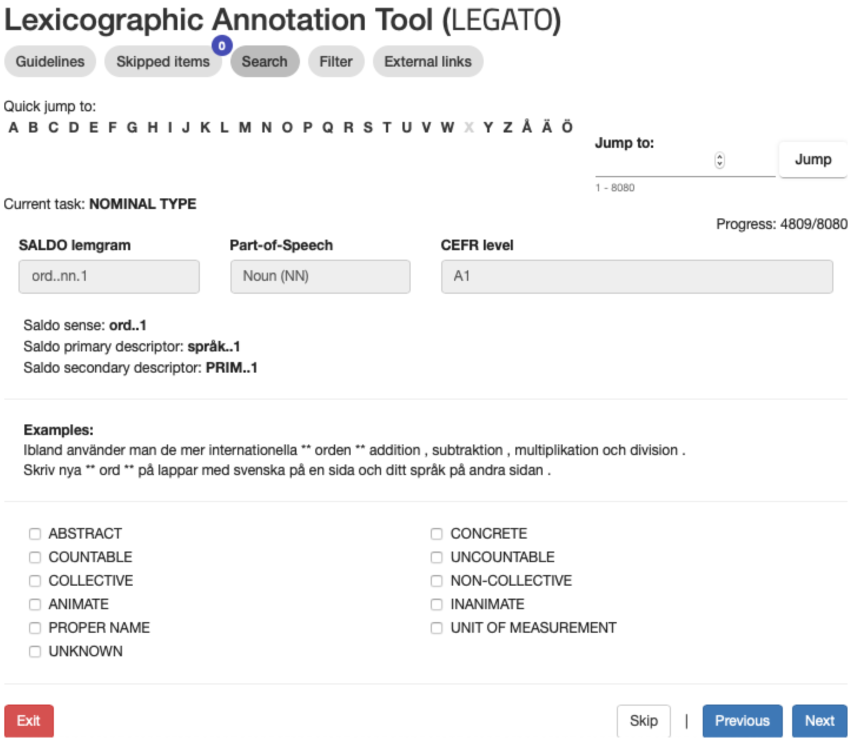
13.3.3 Tool functionality

LEGATO offers a range of useful functionalities. It allows moving forward as well as backwards through the list; to search through the list of word senses to be annotated and to filter by certain criteria; to skip words you are uncertain about. Items that are skipped are added to a dedicated 'skip list' which makes it is easy to come back to these items. It also keeps track of your progress, allowing the annotator to close the interface, come back at a later time and continue where they left. Finally, it includes (automatically generated) links to different external resources such as Saldo (through Karp), Wiktionary, svenska.se,

[62] `https://urlzs.com/PZoRm`

Lexin, synonymer.se, Korp and SAG.

For user friendliness, we keep guidelines, issue-reporting and lookup/reference materials linked to the front page of the tool. It is possible to leave comments, start issues/discussion threads, as well as see an overview of all completed tasks and tasks that are remaining.



Lexicographic Annotation Tool (LEGATO)

Jump to:
1 - 8080

Current task: **NOMINAL TYPE**

Progress: 4809/8080

SALDO lemgram: ord..nn.1
Part-of-Speech: Noun (NN)
CEFR level: A1

Saldo sense: **ord..1**
Saldo primary descriptor: **språk..1**
Saldo secondary descriptor: **PRIM..1**

Examples:
Ibland använder man de mer internationella ** orden ** addition , subtraktion , multiplikation och division .
Skriv nya ** ord ** på lappar med svenska på en sida och ditt språk på andra sidan .

□ ABSTRACT
□ COUNTABLE
□ COLLECTIVE
□ ANIMATE
□ PROPER NAME
□ UNKNOWN
□ CONCRETE
□ UNCOUNTABLE
□ NON-COLLECTIVE
□ INANIMATE
□ UNIT OF MEASUREMENT

Figure 13.1: LEGATO graphical user interface

13.3.4 Piloting the tool

To test LEGATO's functionality as well as to control that the automatic linking of items is sufficiently reliable, we carried out an experiment with 100 SenSVALex items, divided equally between nouns, verbs, adjectives and adverbs. The selected words represent all the CEFR levels available in the COCTAILL corpus, various morphological paradigms and other types of linguistically relevant patterns as shown in Table 13.1.

In order to test the tool, two of the authors volunteered as annotators. After gathering data from the intial test phase, we calculated inter-annotator agreement (IAA) between the automatic analysis and annotator one (IAA 1), as well as the inter-annotator agreement between annotator one and annotator two (IAA 2). Table 13.2 shows Cohen's κ[63] for the various categories. For IAA 1, only categories where annotator one had completed all tasks, and where automatic enrichment was used, were taken into account. For IAA 2, only categories where both of the annotators had completed all tasks were taken into account. This explains why some of the values are missing in the Table.

Category	IAA 1	IAA 2
nominal declension (6)	0.85	0.80
nominal gender (7)	0.82	0.73
nominal type (5)		0.20
verbal conjugation (14)	0.82	0.94
adjectival declension (2)	0.49	
adjectival adverbial structure (1)	0.39	
morphology 1 (3)		0.48
Overall κ	0.73	0.60

Table 13.2: Inter-annotator agreement. Numbers in brackets (Column 1) refer to the numbering of categories in Table 13.1

As can be gathered from Table 13.2, categories with closed answers, e.g. only one possible answer value, lead to higher agreement (nominal declension, nominal gender, verbal conjugation), while categories that allow multiple answers or free-text input show less agreement (nominal type, adjectival adverbial structure, morphology 1). For example, for nominal type, if one annotator selects "abstract, countable, inanimate" and another annotator select "concrete, countable, inanimate", this would be counted as disagreement. In order to address such problems, one would have to calculate partial agreement. One notable exception is adjectival declension, which only allows one value, but has low agreement between the automatic analysis and annotator one. This discrepancy could stem from the fact that all forms in Saldo morphology are automatically expanded, according to regular morphology, thus potentially producing forms that are incorrect.

As a result of the IAA calculations, a subset of categories has been deemed reliable enough to be added automatically (categories 6, 7, 14 in Table 13.1), and another subset will be offered in a semi-automatic way, where a manual control check will be performed (categories 1, 2, 10, 11, 12 in Table 13.1).

[63] While values between 0.40 and 0.60 are generally considered borderline, values of 0.75 and above are seen as good to excellent.

The experiment with the 100 items has also helped us set up and refine guidelines for more extensive annotation by project assistants, as well as improve the functionality of the tool.

13.3.5 Technical details

LEGATO is a module integrated with the LärkaLabb[64] platform. Like its parent platform, the LEGATO front-end is written in TypeScript and HTML using the Angular (previously called *Angular 2*) framework[65]. The back-end is written in Python 2. Data is stored in MySQL format.

Data preparation (i.e. automatic enrichment, see Section 13.3.2) is done outside of the LEGATO platform using a set of dedicated scripts. In a multistep process, these scripts (1) create the sense-based word list, (2) add Saldo primary and secondary descriptors, (3) add further information such as synonyms and nominal gender by linking lexical resources based on lemgram, sense and part-of-speech tuples and (4) add example sentences. The resulting data is played into the databases on the server side to reduce the number of API calls and reduce runtime. As some of these scripts have a rather long runtime (the average time per entry for example selection is 0.66 seconds on an Intel Core i5-5200U processor, resulting in about 3 hours total for the whole list), they are not distributed as an integrated part of LEGATO and we do not consider advisable to integrate them into the LEGATO platform. However, the code for running interlinking can be made available for reuse.

13.4 Concluding remarks

We are currently exploring a possibility of using Lexin (Hult, Malmgren and Sköldberg 2010) and COCTAILL (Volodina et al. 2014a) to automatically derive topical domains for vocabulary items. Furthermore, fruitful negotiations are ongoing on a potential access to parts of the SAOL database (Malmgren 2014) for semi-automatic support of annotation of transitivity patterns.

A full-scale annotation of the two lists is planned for the near future, with the results (i.e. a full resource) expected by the end of 2019. Once the resources are richly annotated, we expect to perform both quantitative and qualitative analysis of L2 lexical competence. The LEGATO tool will have a thorough testing during that time and we hope this will lead to further improvements of the tool.

[64] `https://spraakbanken.gu.se/larkalabb`

[65] `https://angular.io`

Since Legato is a module in a highly intricate and interlinked system Lärka, we do not deem it reasonable to release the code for this module only. However, in the future, we would like to make the platform available to other users by allowing them to upload their own data and define what they want to annotate.

13.5 Acknowledgements

This work has been supported by a grant from the Swedish Riksbankens Jubileumsfond (Development of lexical and grammatical competences in immigrant Swedish, project P17-0716:1).

14 MULTILINGUAL COMPARISON

This publication is discussed in section 5.5.3.

This chapter is a postprint version of the following publication:

Johannes Graën, David Alfter and Gerold Schneider. 2020. Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. In *Proceedings of LREC*.

Abstract

The Common European Framework of Reference for Languages (CEFR) defines six levels of learner proficiency, and links them to particular communicative abilities. The CEFRLex project aims at compiling lexical resources that link single words and multi-word expressions to particular CEFR levels. The resources are thought to reflect second language learner needs as they are compiled from CEFR-graded textbooks and other learner-directed texts. In this work, we investigate the applicability of CEFRLex resources for building language learning applications. Our main concerns were that vocabulary in language learning materials might be sparse, i.e. that not all vocabulary items that belong to a particular level would also occur in materials for that level, and, on the other hand, that vocabulary items might be used on lower-level materials if required by the topic (e.g. with a simpler paraphrasing or translation). Our results indicate that the English CEFRLex resource is in accordance with external resources that we jointly employ as gold standard. Together with other values obtained from monolingual and parallel corpora, we can indicate which entries need to be adjusted to obtain values that are even more in line with this gold standard. We expect that this finding also holds for the other languages.

14.1 Introduction

Graded vocabulary lists have different applications such as serving as a basis for textbook writers, learner dictionaries or as self-paced learning tool for language learners (Kilgarriff et al. 2014). Especially in a second language learning context, vocabulary knowledge is highly correlated with general language proficiency and is a prerequisite for successful communication (Nation 2013).

The main problem is that most graded vocabulary lists do not contain an evaluation of their quality and reliability. Nevertheless, as an user of such resources, one might want to know how reliable the resource is before employing it in the context of a language learning application.

The Common European Framework of Reference (CEFR) for Languages (Council of Europe 2001) is a scale of proficiency divided into three broad levels, A, B, and C, each of which is further subdivided into two sub-levels, so that the full scale ranges over 6 levels, from A1 for beginning learners over A2, B1, B2, C1 to C2 for near-native learners.

The most prominent use of the CEFR is in the form of (1) language certificates such as the Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS), and (2) CEFR-graded textbooks. While most tests have their own scoring system, they can all be mapped onto the CEFR scale. CEFR levels are also used in classroom language teaching to differentiate between different learner groups. Thus, one can have a Swedish class for B1 learners, which presupposes that learners taking the class have mastered all or most of the skills of the lower levels A1 and A2 and should have mastered all or most skills introduced at B1 after having finished the class.

Expression	PoS	A1	A2	B1	B2	C1	Total
video	noun	2.47	0.56	34.83	23.80	13.25	18.43
write	verb	934.71	378.34	760.73	536.38	713.33	549.91
empty	adjective	86.49	150.89	65.95	194.80	123.41	156.02
shopping center	noun	0	0	15.58	0	0.82	1.75
dream up	verb	0	0	0	0	0.82	0.23

Table 14.1: Sample from EFLLex

In this paper, we explore multiple hypotheses relating to the CEFRLex family, a collection of similar resources derived from CEFR-graded textbook corpora, which is available for several languages. Our first hypothesis is that similar words in two languages, i.e. good direct translations, should have similar

CEFR levels. However, this also raises the question of culture- and language-specific vocabulary. A second hypothesis is that the broader a concept is, the lower its CEFR level should be, as the possibility of knowing at least one of the possible interpretations is higher than with highly specific vocabulary. Thirdly, we also explore how the frequency as reflected in CEFRLex and textbooks relates to the frequency of expressions in actual language by looking at the British National Corpus (The BNC Consortium 2007) and the International Corpus of Learner English (ICLE) (Granger et al. 2009).

The CEFRLex resources are based on CEFR-graded textbooks, with the exception of the Swedish SweLLex, which is based on CEFR-graded learner essays. Each single word or multiword expression that has been found in textbooks and other language learner material is listed in its base form, i.e. lemmatized, together with an automatically derived part-of-speech tag. For each entry, the resource lists its normalized distribution over the respective CEFR levels as indicated by the learning material. Table 14.1 shows examples from the English EFLLex.

Other resources aligned to the CEFR that we use in this work are the KELLY lists (Kilgarriff et al. 2014), which exist for nine different languages, including English and Swedish, the Pearson Global Scale of English (GSE) (Pearson 2017) and the Cambridge English Vocabulary Profile (EVP) (Capel 2015) vocabulary lists for English.

Two obvious problems that we are facing are the absence of a gold standard for most languages, as well as a data sparseness issue. Indeed, any such list pertaining to natural language must be finite and cannot, by definition, be exhaustive. This may then result in certain expressions being only found at advanced levels, although they could have been introduced much earlier. Furthermore, textbooks may opt to introduce vocabulary of a higher level if necessary for certain tasks, which may give the impression that a word is used earlier, and thus easier to understand, than expected. In such cases, the words in question are often explained further.

In the absence of a gold standard we are using the two mentioned well-established independent English lexicons GSE and EVP as a base for comparison. For our experiments, we use both of them together as gold standard by combinding their scores. As a resource of the same kind, we expect EFLLex to correlate with the gold data.We further assume that the findings for EFLLex also hold for other CEFRLex resources, as they follow the same methodology.

Thus, the purpose of this study is to evaluate the applicability of CEFRLex resources for language learning applications. To this end, we use monolingual lexical resources in English as an external reference, and compare it to the English CEFRLex. Then, with the help of translation candidates from word-aligned parallel corpora, we identify the divergence from CEFR levels in other languages, and evaluate if the chosen features, in combination with other monolingual resources, lead to a better fit. In our experiments, we only consider single words, as multi-word expressions account for only a small share of the lexical entries (see Figure 14.2 in Section 14.3.4), and word alignment of multi-word units in parallel corpora is less accurate.

14.2 Related work

The KELLY project (KEywords for Language Learning for Young and adults alike) aimed at creating a language learning tool for nine different languages (Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian, and Swedish) (Kilgarriff et al. 2014). To this end, approximately 9,000 keywords were collected for each language, based on their frequency in large corpora. After ordering the list by frequency, it was divided into six equally-sized parts and assigned CEFR levels, from A1 for the most frequent items, to C2 for the least frequent items.

KELLY vocabulary can be seen as "core" vocabulary, i.e. vocabulary that should be known by a prototypical learner of a certain proficiency level. Each resource was also manually translated into all other eight languages. As an added effect of interlinking the lists through translation, it is also possible to identify expressions that occur in all lists ("universal vocabulary"), expressions that occur in most of the lists ("common vocabulary") and expressions that only occur in certain language pairs or only in one single list ("language-specific vocabulary") (Volodina and Kokkinakis 2012).

For English, two of the most prominent resources which, among other things, link expressions to CEFR levels are the English Vocabulary Profile (EVP) (Cambridge University Press 2015), and the Global Scale of English (GSE) Teacher Toolkit (Pearson 2017). While the GSE Teacher Toolkit is freely accessible, EVP requires a (free) subscription.

A possible application for CEFR-graded word lists is, for instance, the readability assessment of texts including visualization of words of different CEFR levels. Projects that employ such a methodology are, inter alia, Duolingo CEFR checker[66] for English and Spanish, Texteval[67] for Swedish and the CEFRLex Lexical Complexity Analyzer[68] for English, Spanish, French and Dutch. Each of these tools highlights words of different CEFR levels in different colors. The first two also incorporate a readability estimation algorithm, which predicts an overall CEFR level for the text, while the latter lets the user select a

[66] https://cefr.duolingo.com/

[67] https://spraakbanken.gu.se/larka/texteval

[68] https://cental.uclouvain.be/cefrlex-demo/analyze

target CEFR level and highlights all words that belong to a level higher than the chosen one.

In readability assessment research, lexical features have repeatedly shown to be one of the most prominent predictors of readability (Beinborn, Zesch and Gurevych 2014; François and Fairon 2012; Heilman et al. 2007; Huang et al. 2011; Pilán, Alfter and Volodina 2016; Volodina, Pilán and Alfter 2016). It has also been shown that replacing "traditional" frequency-based word lists by CEFRLex-derived resources significantly improves results of automatic essay grading (Pilán, Alfter and Volodina 2016).

While tools such as readability assessment of texts can be useful not only to teachers but also to learners, a more learner-targeted application of CEFRLex resources is the automatic generation of exercises, as for example exemplified by the Lärka platform (Alfter et al. 2019) where multiple different exercises such as listening exercises or word guess exercises are automatically generated, or the multilingual particle verb exercise described in (Alfter and Graën 2019), which connects different language resources, all of which are taking into account the level of proficiency of the learner as well as the estimated proficiency level at which a learner can understand certain words as given by the CEFRLex resource.

Some of our English analyses in this paper have already been conducted for Dutch (Tack et al. 2018), such as frequency effects and word length effects. We go beyond their approach by comparing to a soft gold standard, by comparing several algorithms for calculating the learning level, by suggesting possible changes to the CEFR level, and by profiting from multilingual resources.

14.3 Resources

In this work, we use EFLLex for English (Dürlich and François 2018), FLELex for French (François et al. 2014), and SVALex for Swedish (François et al. 2016), all available online.[69] We are aware that CEFRLex resources for Dutch (Tack et al. 2018) and Spanish (François and De Cock 2018) have been compiled, and that there is ongoing work on creating CEFRLex resources for German and Portuguese as well, but for the scope of this paper, we have chosen to limit ourselves to those three language resources that have officially been released.

It should be noted that there are two different versions of the French CEFRLex resource, differing only in the choice of part-of-speech tagger, and that we have chosen the CRF (Conditional Random Field) version, as this tagger is said to be more accurate (François et al. 2014). It should also be noted that

[69] https://cental.uclouvain.be/cefrlex/

only the French CEFRLex resource covers all six CEFR levels, from A1 to C2. All other CEFRLex resources disregard the C2 level, as it is notoriously difficult to find textbooks pertaining to the highest level of proficiency. At this level, learners have attained near-native proficiency and they have, thus, little need for textbooks.

14.3.1 Word alignments from a general corpus

The Sparcling corpus (Graën 2018; Graën et al. 2019a) consists of parallel texts in 16 different languages. It comprises the debates of the European Parliament for a time span of 15 years, originally published as Europarl corpus by Koehn (2005) and released in a cleaner version with document-level alignment by Graën, Batinic and Volk (2014). The corpus features alignment on several levels, from documents down to bilingual word alignment for each language pair. Word alignment has been performed with four different word aligners, namely GIZA++ (Och and Ney 2003), the Berkeley Aligner (Liang, Taskar and Klein 2006), fast_align (Dyer, Chahuneau and Smith 2013) and efmaral (Östling and Tiedemann 2016). For the present work, we only used those alignment links that were supported by all four aligners, thus strongly favoring precision over recall.

Based on those word alignments, we derive the conditional probability of a token with lemma λ_s in one language being aligned with a token with lemma λ_t in another language (Graën 2018: Section 3.2.1). With f_a being the frequency of two lemmas being connected via word alignment of their corresponding tokens, the conditional probability p_a of the target lemma λ_t given the source lemma λ_s is calculated as:

$$p_a(\lambda_t|\lambda_s) = \frac{f_a(\lambda_s, \lambda_t)}{\sum_{\lambda_{t'}} f_a(\lambda_s, \lambda_{t'})}$$

If we also take assigned part-of-speech tags θ into account, this equation extends to:

$$p_a\left((\lambda_t, \theta_t)|(\lambda_s, \theta_s)\right) = \frac{f_a((\lambda_s, \theta_s), (\lambda_t, \theta_t))}{\sum\limits_{(\lambda_{t'}, \theta_{s'})} f_a((\lambda_s, \theta_s), (\lambda_{t'}, \theta_{s'}))}$$

For example, the alignment probability of the French noun ‘vaccin’ given the English noun ‘vaccine’ is high (94%). Other correspondences of the English source lemma identified via word alignment are the verb ‘vacciner’ (to vaccinate), the noun ‘vaccination’ (vaccination), and, with a single occurrence each, the nouns ‘grippe’ (influenza) and ‘médicament’ (medicine/pharmaceutical). The other way round, the alignment probability of English ‘vaccine’ given French ‘vaccin’ is also high (91%). Alternative alignments are ‘vaccination’ and, with a single occurrence, ‘inoculate’.

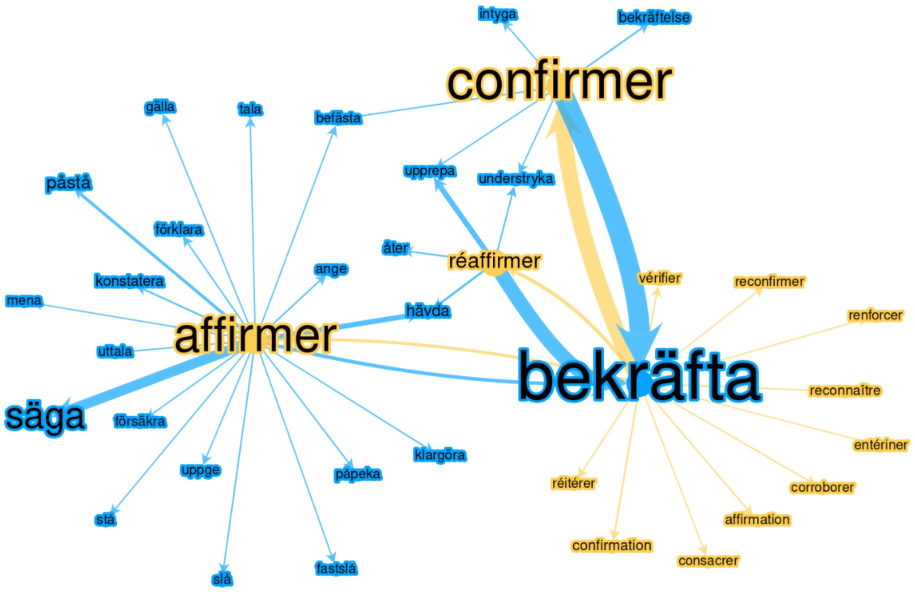


Figure 14.1: Alignment probabilities for Swedish (blue) and French (yellow) words. The size of the nodes represents corpus frequency and the sizes of the connecting lines relates to conditional alignment probability.

While the lemma ‘vaccine’ shows a strong alignment and thus translation preference for ‘vaccin’, and vice versa, other correspondences are not as straightforward. The most frequent alignment of French ‘confimer’ (to confirm) to Swedish is ‘bekräfta’, which also holds for the opposite direction. However, ‘bekräfta’ given ‘confirmer’ is considerably more probable (93%) than ‘confirmer’ given ‘bekräfta’ (70%). Other frequent correspondences of ‘bekräfta’ are ‘réaffirmer’ (10%) and ‘affirmer’ (8%). Figure 14.1 depicts the alignment probability for those words (Graën and Schneider 2020).

In case of compounds in one language that correspond to two or more tokens in a second language, the alignment probability is distributed to all constituents of the corresponding expression, e.g. for English ‘waste management’ and Swedish ‘avfallshantering’, we see a probability of 50% for ‘waste’ given ‘avfallshantering’ and 31% for ‘management’ given ‘avfallshantering’.

For each pair of lexical entries in two CEFRLex resources, we determine the respective conditional alignment probability for both directions from the Sparcling corpus. As we are looking for standard translations, we set a threshold of 25%, below which we ignore alignment probabilities. The alignments of both ‘waste’ and ‘management’ with ‘avfallshantering’ would pass, but for ‘bekräfta’, we would only accept ‘confirmer’ with its value of 70%.

14.3.2 Multilingual core vocabulary

In this paper, we consult the English and Swedish KELLY lists. The KELLY lists can be regarded as core vocabulary, i.e. vocabulary that should be known by a generic learner of a certain level (Kilgarriff et al. 2014). KELLY lists are frequency-based and the assigned CEFR levels directly result from a frequency-ranking of expressions as found in large web corpora.

The English KELLY list was compiled from the UK-Web-as-Corpus (ukWaC) corpus (Ferraresi et al. 2008) and British National Corpus (BNC) (The BNC Consortium 2007). UkWaC contains over 2 billion words, while BNC contains almost 100 million words. The English KELLY list comprises the 7,549 most frequent lemmas, although they are not evenly distributed across the six CEFR levels.[70]

The Swedish KELLY list was compiled from the Swedish-as-a-Web corpus (SweWaC) containing 114 million words. It comprises the 8,425 most frequent lemmas distributed evenly across the six CEFR levels.[71]

14.3.3 Independent English lexicons

We regard the English Vocabulary Profile (EVP) and the graded vocabulary part of Pearson’s Global Scale of English (GSE) Teacher Toolkit as independent lexical resources. Through their respective web interfaces, one can query words and phrases, and the results include, among other information, assigned CEFR levels.[72] It should be noted that both EVP and GSE list word senses.

Since EFLLex does not distinguish between senses, we have chosen to conflate EVP and GSE senses in such a way as to assume the first level of any polysemous word as the target level.

While EVP seems to be targeting productive knowledge, given that it is mainly based on the Cambridge Learner Corpus (Nicholls 2003), GSE is slightly more unclear as to whether it targets productive or receptive knowledge.

[70] A1: 789, A2: 921, B1: 1383, B2: 1107, C1: 948, C2: 2401

[71] Each level from A1 to C1 contains 1404 entries while level C2 contains 1405 entries.

[72] GSE uses a more fine-grained numerical scale from 11 to 89, but also maps this scale onto CEFR levels.

14.3.4 Data overview

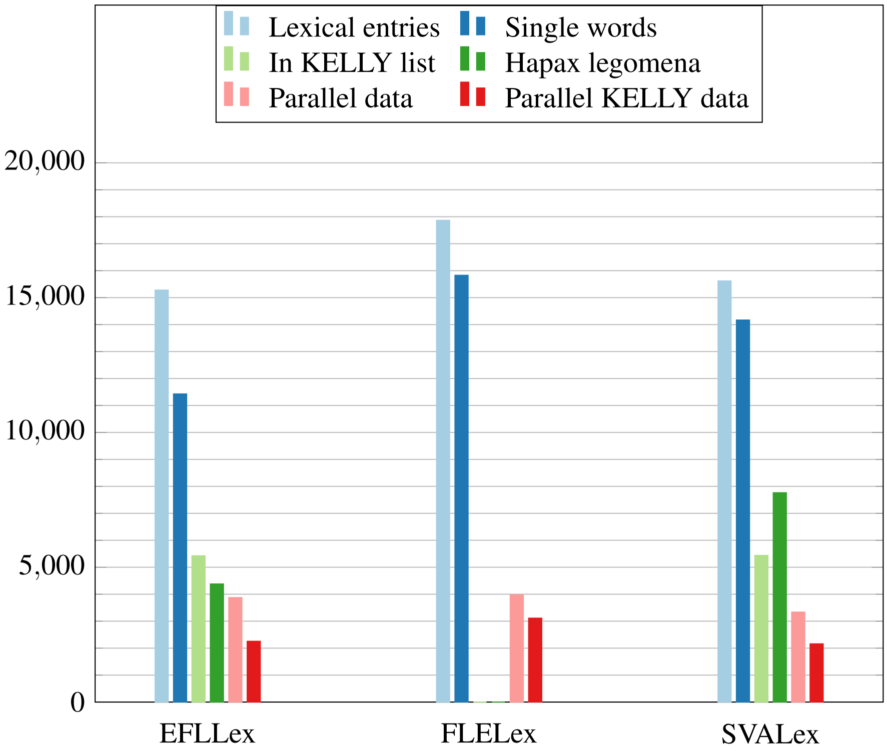


Figure 14.2: Vocabulary sizes for the three languages

In Figure 14.2, we show vocabulary sizes from the CEFRLex resources that we use. The French resource, FLELex, unlike the other two, does not include absolute frequencies, hence we cannot detect hapax legomena. Furthermore, there is no KELLY list for French. We do, however, frequently find the translations of French vocabulary entries into the other two languages in their respective KELLY lists (‘Parallel KELLY data’ in Figure 14.2). For English and Swedish, there can only be one translation that appears in a KELLY list due to the nonexistence of a French list. This is why we find more translations of French entries in the other languages’ lists.

The KELLY lists of English and Swedish comprise less than half of the single-word entries in each language. Between 3,000 and 4,000 entries of each language have parallel correspondences (see Section 14.3.1).

14.3.5 CEFRLex combined

In addition to extracting pairwise language combinations as described in Section 14.3.1, we also created a combined aligned list with entries from all three resources.

To this aim, we start from one language pair, for example French/Swedish, and for each entry we look up possible translations in the third language, English in this case, from the other two resources. Thus, if we start with the French/Swedish list, we retrieve English translations from the aligned English/Swedish and English/French lists. Each entry can have zero, one or multiple translations.

For each translation, we then retrieve its translation probabilities. In case there are multiple translation candidates, we create separate entries. For example, for the French/Swedish noun entry ‘question’ (French) – ‘fråga’ (Swedish), two English correspondences are available, namely ‘question’ and ‘issue’. We thus create two aligned entries as (additional information omitted from the example for readability):

PoS	English	French	Swedish
NOUN	question	question	fråga
NOUN	issue	question	fråga

We repeat this process for each of the three paired lists and merge the resulting lists, removing duplicate entries in the process. The resulting list can still contain partial entries (entries with only two languages) that are covered by more complete entries. This is due to the fact that we only perform a single-step translation look-up, and that some translations might only be reachable under certain circumstances. Thus, in a second step, we remove partial entries which are covered by more complete entries.

This results in a final list of 6,077 entries. While the original pairwise files also contain non-lexical part-of-speech entries such as conjunctions, lexical part-of-speech entries (nouns, verbs, adjectives and adverbs) constitute the majority of entries, as listed in Table 14.2.

Entries can be sparse, i.e. if we do not have a translation candidate in the third language for any given language pair, the entry will only contain the original language pair data. The final combined list contains at least two languages per entry with at least two translation probabilities, up to three languages per entry and $3 \times 2 = 6$ translation probabilities.

pair	entries	lexical entries
en/fr	4012	3976
en/sv	3329	3298
fr/sv	3350	3319

Table 14.2: Number of (lexical) entries per language pair

14.4 Methods

In CEFRLex, each lexical entry (i.e. a pair of lemma and part-of-speech tag) is listed with a distribution of observed frequencies by CEFR level. The frequencies are indicated as relative, normalized, adjusted using dispersion, per-level and per-million-word frequencies (François et al. 2016). The distributions come in different shapes. Figure 14.3 shows the distribution of ‘smör’ (butter) in SVALex.[73] We see a peak at B1 level, but the first occurrence of that word in SVALex is at A2.

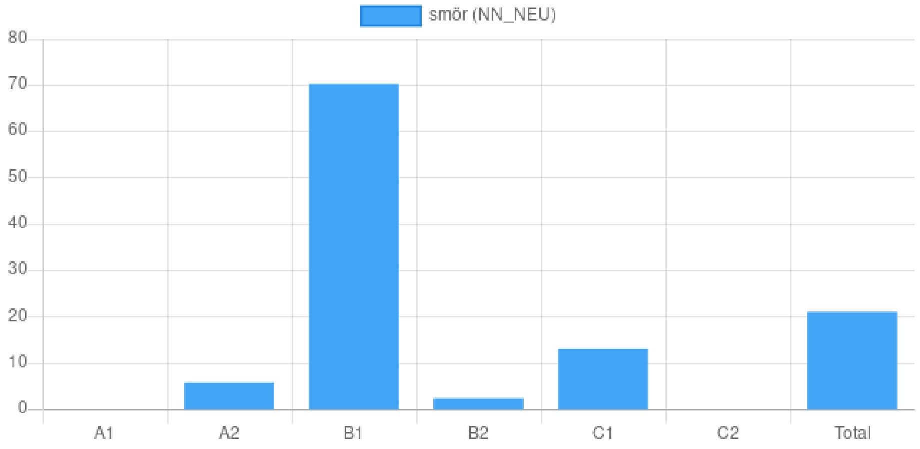


Figure 14.3: The distribution of the noun ‘smör’ (butter) in SVALex over CEFR levels from A1 to C1. The numbers represent the expected frequency in one million running words.

The most straightforward strategy to determine the corresponding level for each entry is to go by the first occurrence, in our example that is A2. In some distributions, however, we see a very small number ($\ll 1$) at the first and a considerably larger number at the second occurrence (occasionally enclosing

[73] These charts are generated by the interactive CEFRLex lookup tool located at `https://cental.uclouvain.be/cefrlex-demo/search`.

an intermediate level without any reported occurrence). We assume that those might be cases where a word or expression of a higher level has been required for a lower-level text. To account for those cases, we define thresholds of 1%, 5% and 10% of the sum of frequencies over all levels.[74] We refer to the first-occurrence CEFR level as C, to those levels determined by the thresholds of 1%, 5% and 10% as C_1, C_5 and C_{10}, respectively.

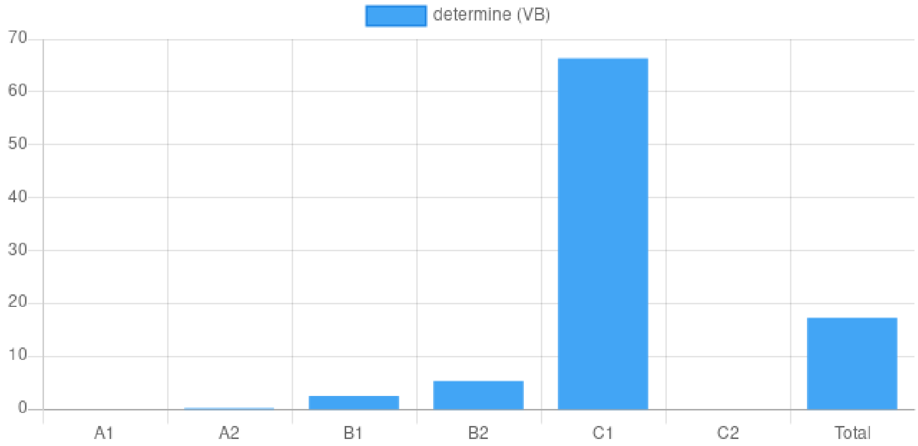


Figure 14.4: The English verb ‘determine’ shows comparably low frequencies at lower levels and a peak at C1.

In most cases (83%), the resulting levels among all thresholds are the same as the first-occurrence level. Figure 14.4 shows one of very few cases (4), where all four resulting levels are different. If we go by first occurrence, we would assign the level A2 to the verb ‘determine’. With a threshold of 1%, we would skip the A2 frequency (0.29 per million words) and assign the level B1. With the highest threshold of 10%, we finally would skip all lower levels and assign C1 as CEFR level.

As described in Section 14.3.1, we identify pairs of lemmas and part-of-speech tags with an alignment probability p_a greater than 25%. For each lexical entry from one of the CEFRLex resources where we find at least one corresponding pair with identical part-of-speech tags, we calculate the minimum (p_{min}), maximum (p_{max}) and average (p_{avg}) of those conditional probabilities in both directions. Minimum and maximum correspond to the directions with a lower and greater probability. For the adjective ‘hungry’, for instance, we find

[74] The ‘total’ number, which forms part of each CEFRLex resource (also shown in Figure 14.3), does not correspond to the sum of frequencies, as each frequency has been normalized to per-million-words over all entries at that level and adjusted by taking dispersion into account. As the total takes all levels into account and the number of observed words per level are different, the numbers do not add up to the ‘total’ number.

the Swedish correspondence 'hungrig'. While 92% of the occurrences of 'hungrig' are aligned to 'hungry', 'hungry' is also frequently translated as 'svälta' (to starve) to Swedish, which leaves a 56% probability for 'hungrig'. The minimum is thus 56%, the maximum 96% and the average 74%. In the cases where we find corresponding entries in two languages, we use the average of both minimal, maximal and average values.

Out of 2976 entries that have correspondences in all three languages (with $p_a > 25\%$ in both directions), 406 show the same CEFR level (C) in all three, and 1981 show the same CEFR level in at least two languages. The remaining 995 entries have different levels. The verbs 'work', 'travailler' and 'arbeta', for instance, are all classified as A1, while 'paralyse', 'paralyser' and 'förlama' are classified as C1. On the other hand, we find different levels for 'adventurous', 'aventurier' and 'äventyrlig' (A2, A1 and C1, respectively) or 'linguist', 'linguiste' and 'lingvist' (A2, B1 and C1, respectively).

We calculate the average difference in terms of CEFR levels between our respective target language and the other languages available, and normalize it to the range from -1 to +1, once by dividing it by the maximal distance of 4 levels[75] (δ) and once by using a sigmoidal function (δ_σ), which projects to the same range, but has a more abrupt gradient, thus giving less relative weight to smaller differences.

For each lexical entry (lemma plus part-of-speech tag), we have assembled the following features:

- The four derived CEFR levels (C, C_1, C_5 and C_{10}) mapped to a linear scale (1 = A1, 2 = A2, ...)
- The CEFR level as defined by KELLY (if available)
- A flag whether a word is only seen once in the corpus (hapax legomenon)[76]
- Three values derived from alignment probabilities (p_{min}, p_{max} and p_{avg})
- The number of languages with an alignment probability of more than 25% in both directions
- The number of languages for which we find a corresponding entry in KELLY[77]

[75] For reasons of comparability, we disregard the near-native level C2, which is only available for French words.

[76] Not available for FLELex as the absolute frequencies are unknown

[77] As KELLY does not cover French, the value is either 0 or 1 for English and Swedish.

- The average difference of corresponding lemmas from the Sparcling corpus (δ) in terms of CEFR levels (normalized to values between -1 and +1)
- The same difference projected to the range -1 to +1 by a sigmoidal function (δ_σ)
- The entry's length in terms of letters
- The entry's frequency from BNC (The BNC Consortium 2007), both from the entire BNC (100 million words), and just the spontaneous conversation section (4 million words)
- The entry's frequency from ICLE (Granger et al. 2009), a corpus of Learner English, with over 3 million words.

In the absence of a hard gold standard, we rely on some of the best industry efforts and best practices, namely GSE and EVP. These two resources are strongly correlated (the Pearson correlation is 0.85), but there are also differences. Following the logic of ensemble approaches (Dietterich 1997) or of the four-eye principle, namely that independent systems typically make partly different errors, offer a different perspective and are thus a good base for triangulation, we have decided to predict the sum of GSE and EVP, i.e. their linear combination, to which refer as GSE&EVP in the following.

In order to assess the correlation of EFLLex to GSE and EVP and other correlations, and in order to test out hypothesis that we can further improve EFLLex, we had to restrict our data sets to those cases where we found an entry in both GSE and EVP, and where we obtained a CEFR level. This gives us a data set of 1,571 lemmas. In the smallest lexical resource, KELLY, which mainly reflects core vocabulary, we replaced the frequent null entries by the highest level (C2) in order not to have to restrict our data set further.

14.5 Results

In this section, we report our results. We use EFLLex, and our suggested changes due to multilingual alignment, and we evaluate using GSE and EVP in combination (GSE&EVP, see previous section) as gold standard, and EFLLex and other features as correlated variables and as predictors.

Feature	Pearson Correlation
Frequency(BNC)	-0.1237227
log(Frequency(BNC))	-0.5081583
log(Frequency(BNC spoken))	-0.7820845
log(Frequency(ICLE))	-0.4028432
word length	0.4515713
log(word length)	0.4572295
C	0.7077803
C_1	0.7061353
C_5	0.7027760
C_{10}	0.6802382
KELLY	0.6464615

Table 14.3: Correlations of individual features to GSE&EVP

14.5.1 Correlations

Among the larger set of features that we have tested, we found high correlations between GSE&EVP and the following features: token frequency, word length, C, and our suggested changes to C.

Correlations to individual baseline features are given in Table 14.3. They confirm and partly extend the findings of (Tack et al. 2018) on Dutch. Concerning frequency, BNC spoken correlates better than the complete BNC, and also better than ICLE, a corpus of Learner English essays. The logarithm of frequency correlates much better, which is in line with psycholinguistic experiments (Smith and Roger Levy 2013) and Zipf's law. A plot of *log(Frequency(BNC spoken))* vs. GSE&EVP/2 is given in Figure 14.5. Concerning word length, length (in letters) and its logarithm correlate very similarly. C shows a strong correlation, albeit less well than the trivial feature of frequency from BNC spoken. This fact already indicates that CEFR levels can be approximated further to our assumed gold standard, thus indicating which entries are more reliable and which may need manual verification. C and C_1 correlate almost equally well (C slightly better), increasing the threshold further leads to a decrease in correlation.

In the next step, we test if the model would fit better, if CEFR levels were closer to their counterparts in other languages. We have tested δ and δ_σ using several feature weights, in order to find out: does the correlation increase if we add the suggested correction? What is the approximate optimal weight of the correction?

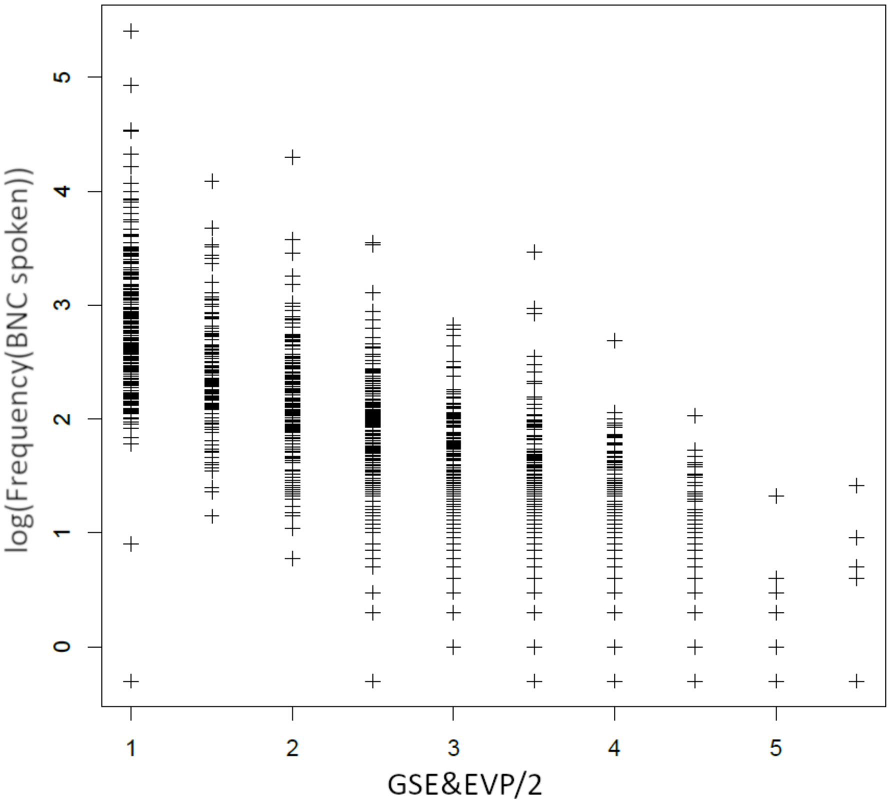


Figure 14.5: Plot of the correlation between GSE&EVP/2 and log(Frequency(BNC Spoken)). Each dot is a word type.

Figure 14.6 shows that δ correlates better than δ_σ, and that optimal weights seem to be around 1.5 or 2.0.[78] The best correlation is 0.759, 0.05 higher than the C baseline. In terms of coefficient of determination (r^2) the proportion of variance increases from $C^2 = 0.708^2 = 50.1\%$ to $0.759^2 = 57.6\%$. Hypothesis 1 has thus been proven.

Correlations can be increased further by adding more features, and adapting the weights of the features. The highest correlations to GSE&EVP reach about 0.85, which is also the correlation between GSE and EVP. A selection of combinations is given in Table 14.4. The last two lines are baseline feature combinations, indicating that an increase of about 3% can be obtained by our approach.

[78] Note that δ and δ_σ are normalized and take values between 0 (no difference found in parallel data) and 1 (a difference of 4 levels, i.e. between A1 and C1).

Features	Pearson Correlation
$C+\delta\times 1.5$	0.7591618
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})$	0.8352488
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})\times 1.2$	0.8393453
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})\times 1.2+\log(\mathrm{word\ length})/3.2$	0.8404394
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})\times 1.2+\log(\mathrm{word\ length})/3.2+(C_1+\delta/20)$	0.8405208
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})\times 1.2+\log(\mathrm{word\ length})/3.2+(C_1+\delta/20)+\mathrm{hapax}/4$ $-$ PoS is NOUN/2 $+$ PoS is VERB/7 $-$ PoS is ADJ/3 $-$ PoS is ADV/3	0.8457225
$(C+\delta\times 1.5-\log(\mathrm{f(BNC\ spoken)})\times 1.2+\log(\mathrm{word\ length})/3.2+(C_1+\delta/20)+\mathrm{hapax}/4$ $-$ PoS is NOUN/2 $+$ PoS is VERB/7 $-$ PoS is ADJ/3 $-$ PoS is ADV/3 $+$ KELLY/5	0.8517668
$\log(\mathrm{word\ length})/3.2-\log(\mathrm{f(BNC\ spoken)})\times 1.2$	0.7828156
$\log(\mathrm{word\ length})/3.2+C\times 1.2-\log(\mathrm{f(BNC\ spoken)})\times 1.2$	0.8210981

Table 14.4: Correlations of weighted feature combinations to GSE&EVP

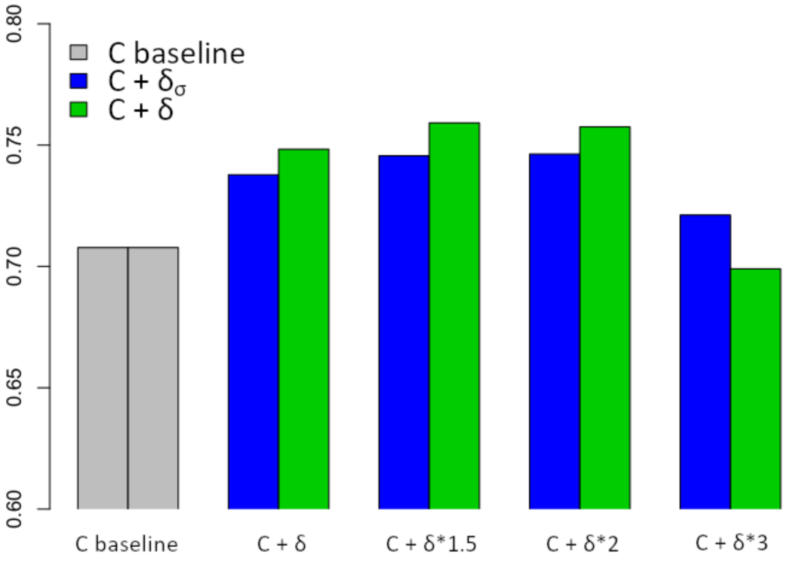


Figure 14.6: Correlation between CEFR values from GSE&EVP and different combinations of C (CEFR) with the relative CEFR level differences δ (div-lin) and δ_σ (div-curve) from parallel data

14.5.2 Regression models

Instead of manually tuning feature weights, linear regression models find optimal weights automatically, and distinguish between significant and non-significant features. For example, the flag indicating hapax legomena is not a significant feature.

We discuss five models in the following: First, a low baseline model, predicting GSE&EVP from word length and log of frequency from BNC spoken. Second, an upper baseline which adds C. Third, the model which includes our best correction, $C + \delta \times 1.5$. Fourth, a model which additionally includes $C_1 + \delta \times 1.5$. Fifth, a model which includes all significant features.

First, the low baseline model which uses word length and frequency is given in Figure 14.7 at the top. It reaches an R^2 value of 61.3%, which can be interpreted as the percentage of the data that is explained by the model.

Second, the model which adds C, but without our suggested CEFR level change, given in Figure 14.7 at the bottom. Its R^2 is 68.7%.

```
lm(formula = GSEplusEVP ~ log10(length) + logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3371 -0.9637  0.0376  0.9198  5.1082

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.23589    0.27435  30.019  < 2e-16 ***
log10(length)  0.72853    0.27476   2.652  0.00809 **
logsf         -1.96662    0.04856 -40.495  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.429 on 1568 degrees of freedom
Multiple R-squared:  0.6134,Adjusted R-squared:  0.6129
F-statistic:  1244 on 2 and 1568 DF,  p-value: < 2.2e-16

lm(formula = GSEplusEVP ~ cefr + log10(length) + logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4810 -0.9199 -0.0176  0.8521  4.6896

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.98622    0.27365  21.875   <2e-16 ***
cefr           0.63892    0.03340  19.131   <2e-16 ***
log10(length)  0.62439    0.24752   2.523   0.0117 *
logsf         -1.39332    0.05302 -26.279   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.287 on 1567 degrees of freedom
Multiple R-squared:  0.6866,Adjusted R-squared:  0.686
F-statistic:  1144 on 3 and 1567 DF,  p-value: < 2.2e-16
```

Figure 14.7: Baseline Regression models

Third, the lower baseline plus our best-performing CEFR change, $C+\delta\times 1.5$. This factor (δ) is highly significant, as the top half of Figure 14.8 shows. It reaches an R^2 value of 70.64%

Fourth, our correlation experiments indicated that adding a correction based on C_1, although less well correlated to GSE&EVP than CEFR-based corrections, may help the model. This is indeed the case, as the bottom half of Figure 14.8 shows. Note that both factors, although highly correlated, stay highly significant.

Fifth, the model including all relevant features also adding PoS tags and KELLY information, but neither the hapax legomena flag, nor C_5-based measures, etc. This model reaches R^2 of 72.9%.

Finally, a word on the quality of prediction is due. We consider the output of the fifth model here. The mean of the absolute value of the difference between GSE&EVP/2 to our prediction is 0.46. This means that a prediction is on average off by 0.46 levels. The residuals (for the second and fifth model), given in Figure 14.9 show a normal distribution, indicating a good model fit. The dif-

```
lm(formula = GSEplusEVP ~ cefr.a2lin + log10(length) + logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7547 -0.8644 -0.0319  0.8033  4.5176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.93961    0.28120  17.566  < 2e-16 ***
cefr.a2lin     0.96377    0.04325  22.284  < 2e-16 ***
log10(length)  0.69285    0.23951   2.893  0.00387 ** 
logsf         -1.21912    0.05401 -22.571  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.246 on 1567 degrees of freedom
Multiple R-squared:  0.7064,Adjusted R-squared:  0.7059 
F-statistic:  1257 on 3 and 1567 DF,  p-value: < 2.2e-16

lm(formula = GSEplusEVP ~ cefr.a2lin + cefr01.a2lin + log10(length) + 
    logsf, data = cefrnozero2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7162 -0.8455 -0.0570  0.8069  4.5308 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.92344    0.28001  17.583  < 2e-16 ***
cefr.a2lin     0.50652    0.12682   3.994 6.79e-05 ***
cefr01.a2lin   0.47196    0.12312   3.833 0.000131 ***
log10(length)  0.63995    0.23887   2.679 0.007460 ** 
logsf         -1.21920    0.05378 -22.671  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.24 on 1566 degrees of freedom
Multiple R-squared:  0.7091,Adjusted R-squared:  0.7084 
F-statistic: 954.6 on 4 and 1566 DF,  p-value: < 2.2e-16
```

Figure 14.8: Central Factor Regression models

ferences between model 2 and 5 are statistically significant (Welch two sample t-test, $p = 0.0004$), tested on the residuals of the second and fifth model, see Figure 14.9. This means that the residuals are significantly smaller on the fifth model than on the second model.

Also confusion matrices confirm the improvement. If we round the prediction of the linear models to the nearest integer, we obtain the confusion matrices given in Figure 14.5. The upper baseline predicts 827 words correctly (out of 1,571), the fifth model 883.

The upper baseline model (the second model) is off by 0.51 levels on average. *C* on it own is off by 0.83 levels on average, partly due to the fact that *C* is 0.6 levels higher than GSE&EVP/2. A model predicting GSE&EVP/2 from *C* only is off by 0.65 levels.

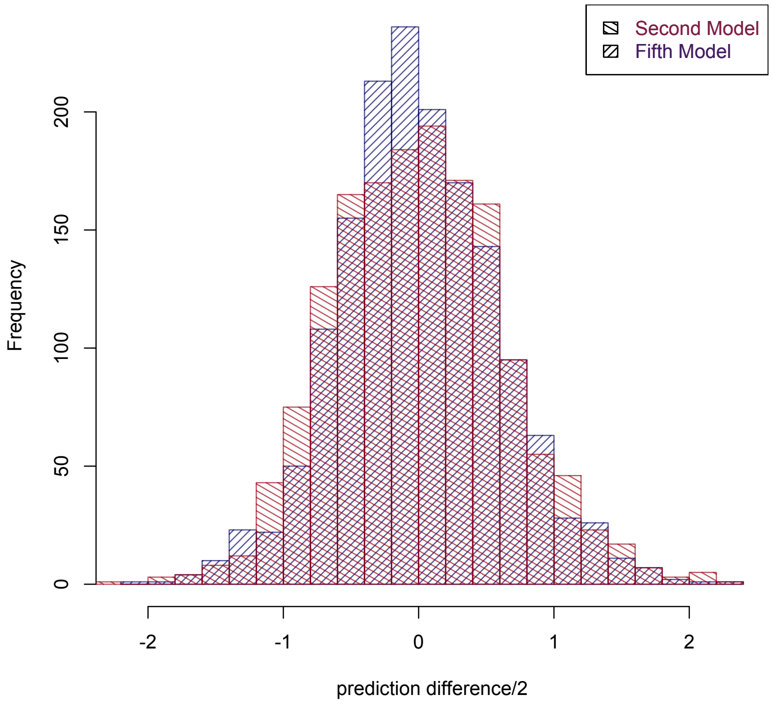


Figure 14.9: Residuals of the second and fifth model

	Upper Base=2nd Model						5th Model					
	0	1	2	3	4	5	0	1	2	3	4	5
1	5	80	155	1	2	0	2	143	95	3	0	0
2	1	34	354	144	13	1	0	51	314	167	15	0
3	0	0	46	153	20	2	0	0	35	167	19	0
4	0	1	40	202	237	59	0	0	23	210	253	53
5	0	0	0	0	6	3	0	0	0	0	3	6
6	0	0	0	4	1	7	0	0	0	1	3	8

Table 14.5: Confusion Matrix of Upper Baseline (second model) vs. fifth model. Predicted is the horizontal, actual in the vertical axis. While the actual range is between 1 (A1) and 6 (C2), the models predict (rounded) values between 0 (below A1) and 5 (C1).

14.6 Discussion

The reduction of prediction difference from 0.51 levels off (upper baseline = second model) to 0.46 levels off (fifth model) seems modest. But we need to bear in mind that we are dealing with several ceiling effects. First and foremost, word length and word frequency (particularly from BNC spoken) are very strong, and partly orthogonal predictors. With a correlation of -0.78 to BNC spoken, word frequency stays the strongest predictor in all models. In the third model, our C-based correction almost reaches the weight of frequency from BNC spoken. At a correlation of 0.71, also C itself is a strong predictor. It is remarkable that our suggestions can lead to a further approximation to our assumed gold standard of GSE and EVP in combination.

The fact that also GSE and EVP, although best efforts and the achievements of best practice from several decades of teaching experience, cannot be regarded as a clear gold standard, but maximally a good proxy to one, is a major limitation of our study.

14.7 Conclusion and future work

In this study, we examined the correlation between the English CEFRLex resource and a soft gold standard. We have found that the CEFRLex-derived levels are highly congruent with our gold standard. The observed deviations are to be expected, as the combined scores of GSE&EVP seem to model productive knowledge, while CEFRLex reflects receptive knowledge; vocabulary is expected to first be understood receptively before it is used productively.

In the future, we would like to include evaluations with French resources, include psycholinguistic variables such as age-of-acquisition, imageability, concreteness, etc., and add eye-tracking reading times. Furthermore, given that our study suggests a good correlation of CEFR levels across three languages, it would be interesting to try and project CEFR levels from these resources to other, possibly under-resourced, languages for which there are no CEFRLex resources.

All features that we calculated, our derived best-fit CEFR level and the multilingual combined entries from the three CEFRLex resources are available at `http://pub.cl.uzh.ch/purl/multiCEFRLex`.

14.8 Acknowledgments

This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1_184212 through the project "From parallel corpora to multilingual exercises: Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications", conducted at Pompeu Fabra University in Barcelona (with Grael, Grup de Recerca en Aprenentatge i Ensenyament de Llengües, and at the University of Gothenburg (with Språkbanken Text).

This research report has made use of the English Vocabulary Profile (EVP). EVP is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile program, which aims to provide evidence about language use that helps to produce better language teaching materials. See `http://www.englishprofile.org` for more information.

15 Automatic exercise generation

This publication is discussed in sections 7.1.1, 7.1.2.1, 7.1.2.2, 7.1.2.3, 7.1.3, and 7.1.4 .

In addition to the information presented in this publication, section 7.1.2.3 gives a more detailed account of the Word guess prototype including information about the experimental inclusion of images, image selection using multilingual resources and unsupervised translation, and how to use crowdsourcing to find good matches between words and images.

This chapter is a postprint version of the following publication:

Alfter, David, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann and Elena Volodina. 2019. Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. In *Selected papers from the CLARIN Annual Conference 2018*, Pisa, 8-10 October 2018 (No. 159, pp. 1-14). Linköping University Electronic Press.

Abstract

Lärka is an Intelligent Computer-Assisted Language Learning (ICALL) platform developed at Språkbanken, as a flexible and a valuable source of additional learning material (e.g. via corpus-based exercises) and a support tool for both teachers and L2 learners of Swedish and students of (Swedish) linguistics. Nowadays, Lärka is being adapted into a building block in an emerging second language research infrastructure within a larger context of the text-based research infrastructure developed by the national Swedish Language bank, Språkbanken, and SWE-CLARIN.

Lärka has recently received a new responsive user interface adapted to dif-

ferent devices with different screen sizes. Moreover, the system has also been augmented with new functionalities. These recent additions aim at improving the usability and the usefulness of the platform for pedagogical purposes. The most important development, though, is the adaptation of the platform to serve as a component in an e-infrastructure supporting research on language learning and multilingualism. Thanks to Lärka's *service-oriented architecture*, most functionalities are also available as web services which can be easily re-used by other applications.

15.1 Introduction

Lärka[79] is an Intelligent Computer-Assisted Language Learning (ICALL) platform developed at the CLARIN B Center Språkbanken Text (University of Gothenburg, Sweden). Lärka development started in the project *A system architecture for ICALL* (Volodina et al. 2012), the initial goal being to re-implement a previous tool, ITG, used up until then for teaching grammar (Borin and Saxena 2004) with modern technology. The new application, Lärka, gradually developed into a platform for language learning covering two groups of learners – second/foreign language learners of Swedish and students of (Swedish) linguistics. Lärka is an openly available web-based tool that builds on a variety of existing SWE-CLARIN language resources such as Korp (Borin, Forsberg and Roxendal 2012) for querying corpora, Karp (Borin et al. 2013) for querying lexical resources and language technology tools (Borin et al. 2017). Thanks to its service-oriented architecture, Lärka functionalities can be re-used in other applications (Volodina et al. 2014b).

In parallel to exercise generation functionalities, Lärka has been evolving into a research tool with a number of supportive modules for experimentation and visualization of research results, such as for selection of best corpus examples for language learners, for readability analysis of texts aimed at or produced by language learners, for prediction of single-word lexical difficulty, as well as for facilitating text-level annotation of language learner corpora, but also to collect data from exercises where learner interaction with the platform and their input have been used in research on metalinguistic awareness. Lärka is actively used in teaching grammar to university students, where we can report only those uses that we have explicitly been told about. As we do not require login to the platform, we do not know who our users are, but we can deduce from the logs that Lärka is being used beyond the reported schools and universities.

[79] `https://spraakbanken.gu.se/larka`

Nowadays, Lärka is being adapted into a building block in an emerging second language research infrastructure SweLL (Volodina et al. 2018), within a larger context of the text-based research infrastructure developed by the national Swedish Language bank, Språkbanken, and SWE-CLARIN. This addresses an obvious need within CLARIN, as evidenced by the interest in the recent CLARIN workshop on "Interoperability of Second Language Resources and Tools".[80]

The current paper describes the new version of Lärka that was released in 2016, replacing the 2013 version, and illustrates improved and newly added functionalities.

15.2 Related work

There have been some attempts to combine exercise platforms with different types of data collection. The *Writing Mentor* Google Docs add-on, for example, allows users to get feedback on their writing in different categories such as coherence, topic development or use of sources to back up claims. The application uses natural language processing tools to provide users with feedback but at the same time collects the texts and all subsequent modifications to the texts that have been analyzed (Madnani et al. 2018). However, accessibility of the data for SLA research is limited.

The FeedBook project (Rudzewitz et al. 2017) is based on an English text book and presents the text book in a digitized interactive web platform that has been enriched with natural language processing to provide immediate fine-grained feedback to the students concerning both form and meaning errors. Teachers can also see their students' progress and provide individual feedback. The data is logged and is used iteratively for further improvement of the system, the data access so far being limited to the researchers involved in the project.

Most applications, however, are purely pedagogical. An outstanding example is the *Language Muse* Activity Palette (Burstein and Sabatini 2016; Burstein et al. 2017). It allows teachers to upload texts and automatically generates exercises based on these texts. Texts are analyzed using natural language processing algorithms to identify different linguistic features such as multi-word expressions, syntactic relations and discourse structure. Based on the analysis, the platform creates over twenty different activities for the teacher to choose from, such as antonym exercises, homonym exercises or verb tense exercises. Teachers have full control over which texts are used, and are offered

[80] See `https://sweclarin.se/eng/workshop-interoperability-l2-resources-and-tools`

a possibility to edit automatically suggested exercise items. In that way, teachers can build a 'palette' of activities from the original text that best suits their and their students' needs.

Perez and Cuadros (2017) propose a framework for automatic exercise generation from user-specified texts that works with Spanish, Basque, English and French. Users can use texts of their own choosing in four different languages. The framework can generate three different kinds of tasks, namely gap exercises, multiple-choice exercises and sentence rearrangement exercises. Furthermore, the framework automatically generates hints for the gap exercise and allows for the adjustment of the number of distractors for multiple-choice exercises. Exercises are also exportable in Moodle's CLOZE[81] format, increasing its appeal.

On the other hand, there are multiple examples of SLA and psycholinguistic experiments that are staged through exercises that elicit certain types of data from language learners – data that helps researchers to address particular research questions, e.g. Andersson, Sayehli and Gullberg (2018) investigating the influence of the native language on the processing of the word order in Swedish or Kerz and Wiechmann (2017) studying individual differences in L2 processing of multi-word phrases.

We argue that exercise generation platforms/applications have a capacity to mediate between language learners and researchers, bringing interests of the two groups together. We aim to foster this collaboration through the Lärka platform.

Lärka started as an exercise generation platform for learners of Swedish, and later it was extended to support the development and visualization of new algorithms in support of language learning. Now we are taking a new direction, combining research interests from Second Language Acquisition (SLA), Learner Corpus Research (LCR) and language learning into one and building an infrastructure supporting the collection of L2 data through exercises.

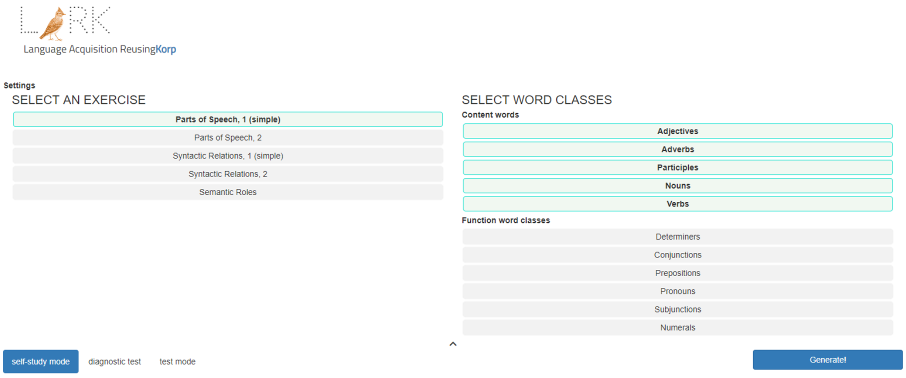
In the next sections, we delineate how Lärka can be and is used as a pedagogical tool in teaching students of Swedish linguistics (Sections 15.3.1 and 15.4), the different exercises in support of research aimed at L2 Swedish (Section 15.3.2), and the various components that constitute the research infrastructure facet of the platform (Section 15.5).

[81] https://docs.moodle.org/23/en/Embedded_Answers_(Cloze)_question_type

15.3 Lärka for learning and teaching

One of the main functionalities of Lärka is the automatic generation of exercises based on real-life authentic language examples from corpora. Exercise generation is aimed at two groups of learners: students of (Swedish) linguistics and learners of Swedish as a second or foreign language (L2).

15.3.1 Exercises for students of linguistics



Settings

SELECT AN EXERCISE

Parts of Speech, 1 (simple)
Parts of Speech, 2
Syntactic Relations, 1 (simple)
Syntactic Relations, 2
Semantic Roles

SELECT WORD CLASSES

Content words

Adjectives
Adverbs
Participles
Nouns
Verbs

Function word classes

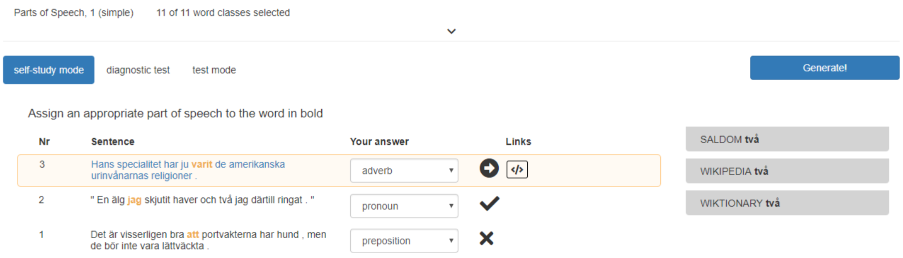
Determiners
Conjunctions
Prepositions
Pronouns
Subjunctions
Numerals

Figure 15.1: Exercises for linguists

Students learning grammatical analysis are in constant need of exercises and feedback on their analysis. Lärka offers exercises for linguistic analysis of parts of speech (word classes), syntactic relations and semantic roles. The exercises are based on authentic texts, which can make them more difficult than textbook examples. However, they are authentic examples of the type of texts the students are expected to be able to analyze in the future. Through on-the-spot feedback, students' learning is enhanced, especially if exercises are done at least partly in a class room setting with the possibility of consulting a teacher and/or the possibility of discussing one's analysis with a fellow student and together trying to make sense of why the automatic feedback said that they got it right or wrong (Lindström Tiedemann, Volodina and Jansson 2016).

As mentioned above, Lärka offers students 3 types of exercises: parts of speech, syntactic relations and semantic roles. The first two offer two levels of difficulty (beginner and intermediate), whereas the third exercise, semantic roles, is only available as one level (Figure 15.1).

Exercises are presented as shown in Figure 15.2 with a sentence and a word or phrase highlighted in another colour. The learner then has to select from a multiple choice drop down box which answer is correct given the highlighted word or phrase. In part-of-speech exercises for example, learners have to select the correct part of speech for the highlighted word.



Parts of Speech, 1 (simple) 11 of 11 word classes selected

Assign an appropriate part of speech to the word in bold

Nr	Sentence	Your answer	Links
3	Hans specialitet har ju **varit** de amerikanska urinvånarnas religioner .	adverb	➔ </>
2	" En älg **jag** skjutit haver och två jag därtill ringat . "	pronoun	✓
1	Det är visserligen bra **att** portvakterna har hund , men de bör inte vara lättväckta .	preposition	✗

Figure 15.2: Execises for linguists

The exercises are available in three different modes: self-study, diagnostic test or test. Students can choose whichever mode they want to use. In self-study mode, answers can be revised as often as desired and needed, even after submitting the answer. In this way, if the answer was incorrect, it is possible to find the correct answer. In test mode, answers cannot be changed after submitting and the correct answer will be shown immediately after submitting. In diagnostic mode, as in test mode, answers cannot be changed after submitting. In addition, the number of exercise items is limited to three items of each main category, e.g. the part-of-speech exercise type covers eleven part-of-speech categories, resulting in a total of 33 diagnostic exercise items. Exercise generation stops after completion of all items in diagnostic mode and a summary is provided which can be emailed to the teacher for further comments or to oneself in order to study the examples further or to be able to track one's learning. In contrast, the other two modes generate exercises infinitely. In both self-study and test mode the actual categories practiced can also be chosen (e.g. one can select to only practice adjectives and adverbs for part-of-speech exercises), whereas the diagnostic test automatically selects all available categories.

In order to avoid exercise item repetition, a sentence will be shown only once during the same session.

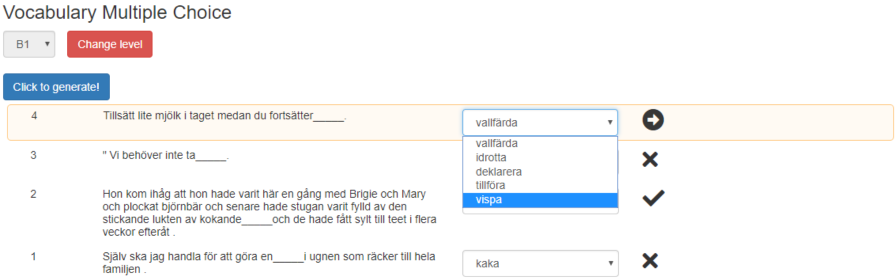
15.3.2 Exercises for language learners

Lärka offers a number of exercises for learners of L2 Swedish as illustrated in the following paragraphs. For all learner exercises, target vocabulary items are sampled from SVALex (François et al. 2016) and SweLLex (Volodina et al. 2016b). SVALex presents a list of lemmata occurring at the different CEFR (Common European Framework of Reference for Languages (Council of Europe 2001)) levels in the textbook corpus COCTAILL (Volodina et al.

2014a). Similarly, SweLLex is based on the pilot SweLL corpus (Volodina et al. 2016a), a corpus of learner essays. We map each distribution to a single CEFR level according to two approaches, namely *first-occurrence* (Gala, François and Fairon 2013; Gala et al. 2014) and *threshold* (Alfter et al. 2016).

The exercises target lexical knowledge of Swedish L2 learners, and speaking pedagogically, train lexical knowledge from various points of view, namely: listening and spelling of lexical items, recognition of an appropriate item for a given context, morphological inflectional behaviour of individual lexical items, and linking definitions/translations with words. There are certainly a many other conceivable exercises that target different word knowledge aspects that we have not implemented. While even the exercise types that we currently offer are still in need of evaluation with teachers and learners, we do believe that they are useful. The session logs for the listening and spelling and word guess exercises show that there is interest in these types of exercises.

15.3.2.1 Vocabulary and inflection


Vocabulary Multiple Choice

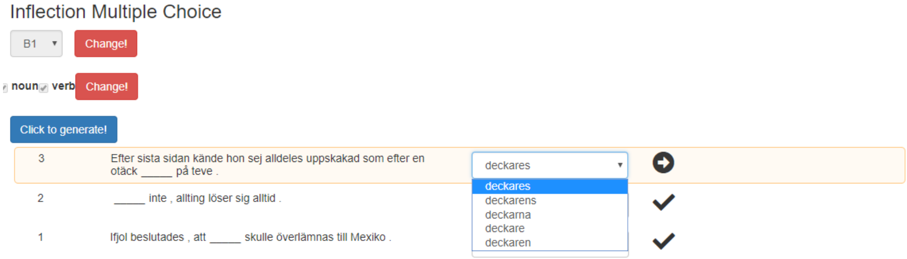
4 Tillsätt lite mjölk i taget medan du fortsätter_____. vallfärda

3 " Vi behöver inte ta_____. ✗

2 Hon kom ihåg att hon hade varit här en gång med Brigie och Mary och plockat björnbär och senare hade stugan varit fylld av den stickande lukten av kokande_____och de hade fått sylt till teet i flera veckor efteråt . ✓

1 Själv ska jag handla för att göra en_____i ugnen som räcker till hela familjen . kaka ✗

Figure 15.3: Vocabulary multiple choice


Inflection Multiple Choice

3 Efter sista sidan kände hon sej alldeles uppskakad som efter en otäck _____ på teve . deckares

2 _____ inte , allting löser sig alltid . ✓

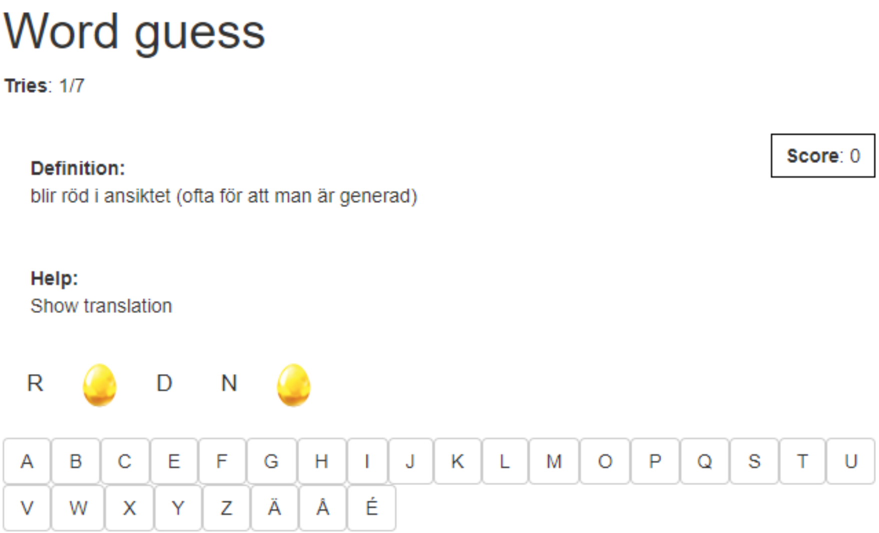
1 Ifjol beslutades , att _____ skulle överlämnas till Mexiko . ✓

Figure 15.4: Inflection multiple choice

Vocabulary exercises and inflection exercises have a multiple-choice format. Each item consists of a sentence containing a gap, as well as a list of five answer alternatives, of which one is correct and four are *distractors*, i.e. incorrect options (Figure 15.3). For vocabulary, distractors are chosen of the same word class as the target word. This morphological selection is further restricted by requiring that distractors be of the same number and/or definiteness as the target item for nouns or the same voice and/or tense for verbs. In case the restriction on the distractors returns too few results, these constraints can be relaxed or dropped.

For inflection exercises, we look up all morphological forms of the target word in Saldo's morphology (Borin, Forsberg and Lönngren 2013) and use a subset of those as distractors. Figures 15.3 and 15.4 show the vocabulary and inflection multiple choice exercise respectively.

15.3.2.2 Word guess



Word guess

Tries: 1/7

Score: 0

Definition:
blir röd i ansiktet (ofta för att man är generad)

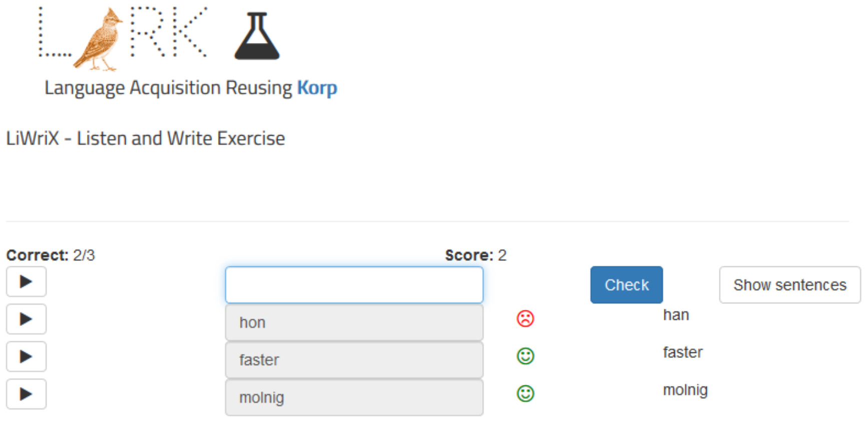
Help:
Show translation

Figure 15.5: Word guess

A recent addition to our platform is a simple word-level exercise, *Word guess*, that takes a step towards gamified learning. Word guess re-implements the well-known Hangman game format: users are presented with a number of hidden characters and the definition of the word in Swedish, and their task is to guess letters contained in the word, which eventually helps them guess the word itself, as shown in Figure 15.5. Every time the guessed character is not in the word, users receive penalty points. In our learning-oriented version of the game, users can choose to receive clues such as the translation of the word (into a range of different languages). Both the definition and translations are retrieved from *Lexin*, a core-vocabulary lexicon for immigrants (Gellerstam 1999). This game is a simple example of reusing information from lexical resources for gamified language learning activities.

15.3.2.3 Liwrix

Another exercise is the listening exercise *Liwrix* (Volodina and Pijetlovic 2015). This exercise makes use of Text-to-Speech (TTS) technology by SitePal[82] to dynamically generate audio of single words and multi-word expressions. In the future, we also intend to include phrases and sentences, as was done in the previous version of Lärka. The delay is caused by the newly introduced hint system which needs to be modified in order to work with phrases and sentences.



Language Acquisition Reusing Korp

LiWriX - Listen and Write Exercise

Correct: 2/3 **Score:** 2

▶			
▶	hon	☹	han
▶	faster	☺	faster
▶	molnig	☺	molnig

Figure 15.6: Liwrix

Figure 15.6 shows the exercise. By clicking on the button on the left, a word or multi-word expression is played and the answer is to be entered into the textfield. In addition, hints are available: As a first hint, but also to avoid problems with homonyms or possible mispronunciations, users can get “clues” in the form of a number of sentences in which the word(s) to be guessed appear in context. As a second hint, learners can choose to have the initial letter of the target word revealed.

[82] sitepal.com

Feedback is given in the form of a green smiley if the answer was correct and a red smiley if the answer was incorrect. In test mode (as in Figure 15.6) the correct answer is also shown irrespective of the correctness of the learner input.

15.4 Lärka in practice

Lärka for linguists has been used in introductions to grammar and linguistics in Sweden and Finland (Volodina et al. 2014b; Lindström Tiedemann, Volodina and Jansson 2016). In Uppsala the platform was often used in lab sessions first so that students had a chance to consult a teacher when they had questions and they were also encouraged to discuss their analysis and the automatic feedback they got with their fellow students.

In Helsinki students have sometimes been encouraged to use it independently on courses in Swedish grammar where they have then been asked to hand in some of their analysis to their teacher or simply been told to use it to get more practice which is something they clearly cannot get too much of in learning grammatical analysis. Some exercise books might not even come with a key, which means that all exercises must be treated in class if the students are to find out what they did right or wrong. In comparison, Lärka material is better suited in this case than many exercise books since it provides authentic texts accompanied by immediate automatic feedback.

The students felt that this was of great use and definitely thought that the platform should be used in the future. In a study with 45 students, Lärka was generally well received. Figure 15.7 shows that the majority of students were in favor of keeping Lärka as part of lab sessions with 34 students (78%) responding strongly in favor of keeping Lärka (scores 5–6), while 10 students (22%) showed more reservation (scores 3–4). No students voted against keeping Lärka (scores 1–2). Similarly, Figure 15.8 shows that 80% of students would recommend Lärka to a fellow student while 20% showed reservations.

A more recent analysis of the linguistic exercise log data collected through the 2016 version of the platform shows that during the time span from October 2016 to May 2018, there were 2086 sessions. One session is counted as a user using the platform from the moment of opening the page to closing it. As we do not require users to login, we create anonymous session identifiers each time a user opens the page. Thus, multiple sessions can stem from the same user. There were 126 sessions in the period from October 2016 to December 2016, 1449 sessions during 2017 and 511 sessions from January 2018 to May 2018.

During those 2086 sessions, a total of 47082 interactions were carried out.

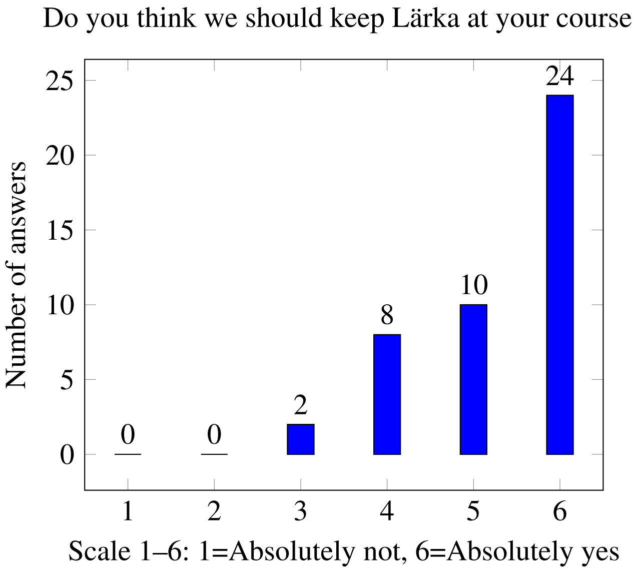


Figure 15.7: Evaluation results 1

Exercise type	# interactions	Operating system	# interactions
Part of Speech 1	28,544	Android	1,081
Part of Speech 2	6,717	Linux	1,093
Semantic roles	553	Mac OS X	11,516
Syntactic relations 1	8,426	Windows	18,962
Syntactic relations 2	2,842	iOs	2,102

(a) Exercise type *(b)* Operating system

Table 15.1: Interaction by exercise type (a) and operating system (b)

One interaction counts as an exercise item being completed. Interaction counts do not include self-corrections, mode changes or helps consulted. Table 15.1 (a) shows the breakdown of the interactions per exercise type. A logging feature that was added later[83] was the logging of whether the page was accessed from a mobile device and which operating system was used to access the page. The logs show that the linguist exercise was accessed 3,184 times (~10%) from a mobile device, as opposed to 31,571 times from a non-mobile device. Table 15.1 (b) shows the breakdown of interactions by operating system.

Furthermore, we can see that the platform was mainly accessed from Swe-

[83] That is why the total is lower than 47,082

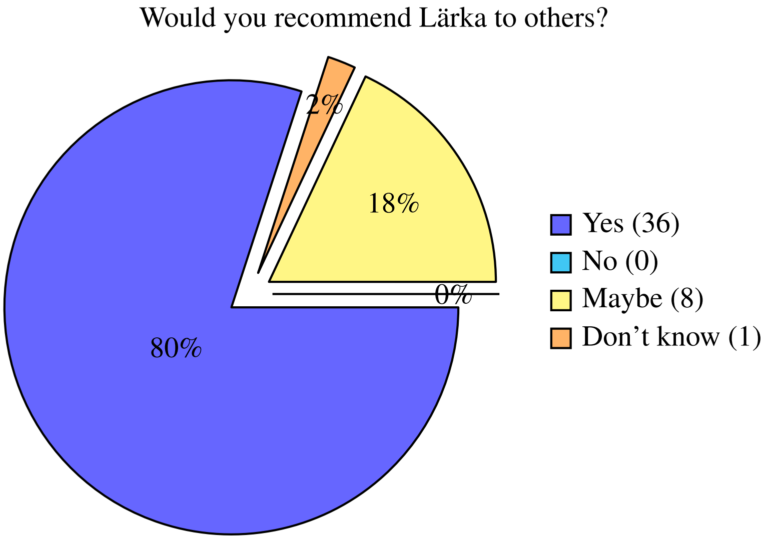


Figure 15.8: Evaluation results 2

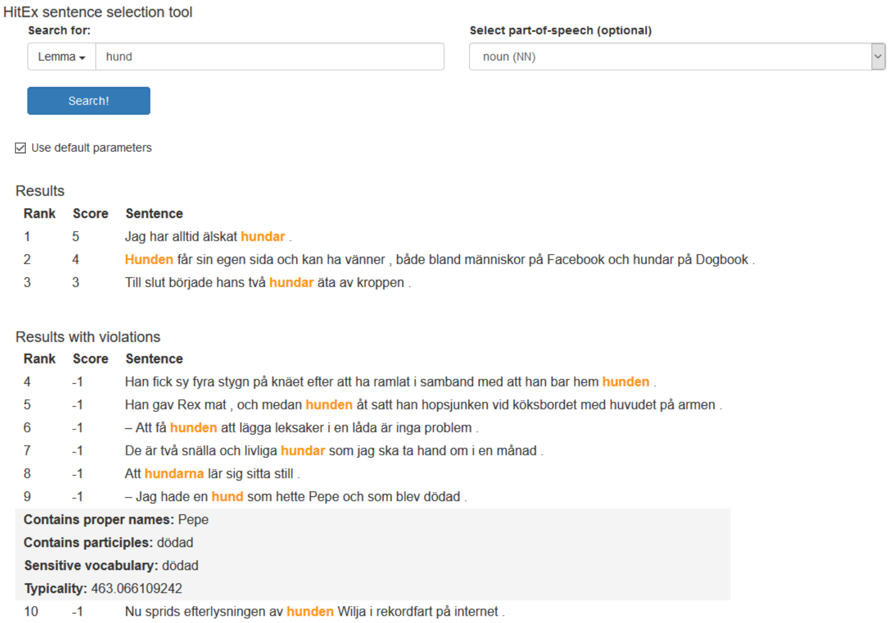
den (91%) and Finland (8%), but also from other countries such as the US, Poland, Germany, the Netherlands, Turkey, Estonia, the UK, India, Belgium, Switzerland, Japan, Canada and Russia, together making up the remaining 1%.

15.5 Lärka as research infrastructure

Lärka is being developed to serve as one of the e-infrastructure components offered to the research community by the Swedish CLARIN B-centre Språkbanken Text at the University of Gothenburg. Specifically Lärka is intended to be used as an infrastructure for research in (Swedish as) L2 acquisition. Currently Lärka offers modules for (1) collection of data from learners through their interaction with the platform, i.e. exercise logs; (2) text-level annotation of learner essays and course book texts; as well as (3) experimentation and visualization of the ongoing research in support of language learning.

With these modules, materials and exercises can be tailored drawing on vast collections of naturally occurring language, in a precise yet flexible as well as replicable way, and students' responses and reactions can be recorded in detail for subsequent quantitative and qualitative analysis. In order to achieve the necessary combination of precision and flexibility, we integrate natural language processing tools and algorithms for corpus example selection, text assessment and automatic exercise generation. These aspects are described in more detail below. A recent direction is "profiling" lexical and grammatical competences that learners of Swedish have, where we experiment with different lexical resources for exercise creation, and in the near future expect to integrate research on grammar profiles.[84]

15.5.1 Corpus example selection



HitEx sentence selection tool

Search for: Lemma hund

Select part-of-speech (optional): noun (NN)

☑ Use default parameters

Results

Rank	Score	Sentence
1	5	Jag har alltid älskat **hundar** .
2	4	**Hunden** får sin egen sida och kan ha vänner , både bland människor på Facebook och hundar på Dogbook .
3	3	Till slut började hans två **hundar** äta av kroppen .

Results with violations

Rank	Score	Sentence
4	-1	Han fick sy fyra stygn på knäet efter att ha ramlat i samband med att han bar hem **hunden** .
5	-1	Han gav Rex mat , och medan **hunden** åt satt han hopsjunken vid köksbordet med huvudet på armen .
6	-1	– Att få **hunden** att lägga leksaker i en låda är inga problem .
7	-1	De är två snälla och livliga **hundar** som jag ska ta hand om i en månad .
8	-1	Att **hundarna** lär sig sitta still .
9	-1	– Jag hade en **hund** som hette Pepe och som blev dödad .

Contains proper names: Pepe
Contains participles: dödad
Sensitive vocabulary: dödad
Typicality: 463.066109242

Rank	Score	Sentence
10	-1	Nu sprids efterlysningen av **hunden** Wilja i rekordfart på internet .

Figure 15.9: Corpus example selection tool HitEx: Results

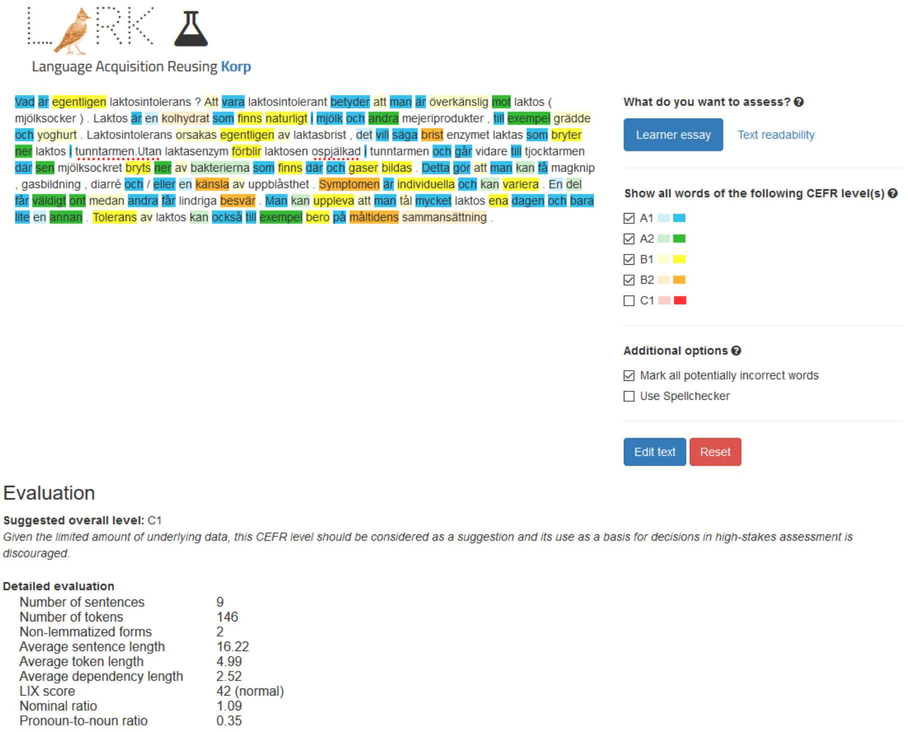
In Lärka, the automatically generated exercises for language learners rely on *HitEx* (*Hitta Exempel* 'find examples'), a tool for selecting and ranking corpus examples (Pilán, Volodina and Borin 2017). The main purpose of HitEx is to identify sentences from generic corpora which are suitable as exercise items for L2 learners. The suitability of the sentences is determined based on a number of parameters that reflect different linguistic characteristics of the sentences. Through a graphical user interface, it is also possible to conduct a sentence search based on parameters customized by the user. The selection criteria include a wide variety of linguistic aspects such as the desired difficulty level based on CEFR, typicality based on word co-occurrence measures, as well as the absence of anaphoric expressions and sensitive vocabulary (e.g. profanities), just to name a few. It is also possible to use a set of default param-

[84] https://spraakbanken.gu.se/eng/l2-profiling

eters for searching. Figure 15.9 shows the results of HitEx. Sentences which fulfill all the required parameter constraints are shown on top while results that violate one or more constraints are shown under ‘Results with violations’. Upon clicking on one of the sentences, more information is shown.

15.5.2 Text complexity evaluation

Another functionality, *TextEval*, offers an interface to automatically assess Swedish texts for their degree of complexity according to the CEFR. Texts can be either learner productions (e.g. essays) or texts written by experts as reading material for learners. The machine learning based automatic analysis returns an overall CEFR level for the text, as well as a list of linguistic indicators relevant for measuring text complexity, such as the average length of sentences and tokens, LIX score and nominal ratio. In addition, it is possible to add a color-enhanced highlighting for words per CEFR levels which provides users with a straightforward visual feedback about the lexical complexity of a text.



Vad är egentligen laktosintolerans ? Att vara laktosintolerant betyder att man är överkänslig mot laktos (mjölksocker) . Laktos är en kolhydrat som finns naturligt i mjölk och andra mejeriprodukter , till exempel grädde och yoghurt . Laktosintolerans orsakas egentligen av laktasbrist , det vill säga brist enzymet laktas som bryter ner laktos i tunntarmen.Utan laktasenzym förblir laktosen ospjälkad i tunntarmen och går vidare till tjocktarmen där sen mjölksockret bryts ner av bakterierna som finns där och gaser bildas . Detta gör att man kan få magknip , gasbildning , diarré och / eller en känsla av uppblåsthet . Symptomen är individuella och kan variera . En del får väldigt ont medan andra får lindriga besvär . Man kan uppleva att man tål mycket laktos ena dagen och bara lite en annan . Tolerans av laktos kan också till exempel bero på måltidens sammansättning .

What do you want to assess?

Learner essay Text readability

Show all words of the following CEFR level(s)

☑ A1 🟦🟦
☑ A2 🟩🟩
☑ B1 🟨🟨
☑ B2 🟧🟧
☐ C1 🟥🟥

Additional options

☑ Mark all potentially incorrect words
☐ Use Spellchecker

Edit text Reset

Evaluation

Suggested overall level: C1

Given the limited amount of underlying data, this CEFR level should be considered as a suggestion and its use as a basis for decisions in high-stakes assessment is discouraged.

Detailed evaluation

Number of sentences	9
Number of tokens	146
Non-lemmatized forms	2
Average sentence length	16.22
Average token length	4.99
Average dependency length	2.52
LIX score	42 (normal)
Nominal ratio	1.09
Pronoun-to-noun ratio	0.35

Figure 15.10: Text complexity evaluation tool TextEval

Figure 15.10 shows the analysis of a text with word-level CEFR highlighting. We use the aforementioned lists SVALex and SweLLex to mark up receptive and productive vocabulary respectively. For each CEFR level, a darker and a lighter shade of the same color represents productive and receptive vocabulary respectively at the given level.

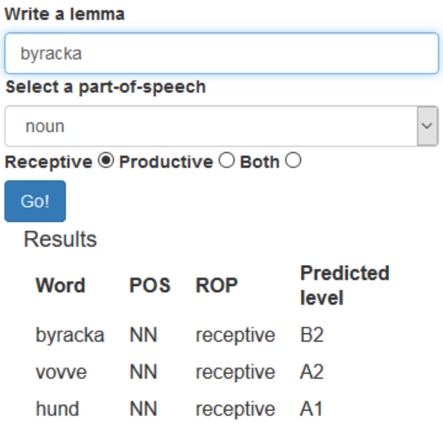
15.5.3 Lexical complexity prediction

Based on the word lists SVALex and SweLLex, which have been transformed so as to map each word to a single CEFR level as described in Alfter et al. (2016), we have built a module capable of predicting the complexity of any Swedish word, not only words occurring in the word lists (Alfter and Volodina 2018). For each word, we extract both traditional word-based features such as length, number of syllables, number of homonyms and also information about topics, i.e. which topics a word belongs to. For example, the word *fisk* 'fish' would occur in the topics 'Animals' and 'Food'. We then feed a machine learning algorithm these feature vectors as well as the predicted mapped single CEFR level of the word and let the algorithm learn how to map from these features to CEFR levels.

An interested user can test a bespoke interface to get predictions about the complexity of a word and its target level (receptive versus productive), as shown in Figure 15.11. This user interface can be used for getting predictions of any word, not only words present in the word lists. The input word is transformed into a feature vector as described above and then fed into the classifier, which predicts a label. Figure 15.11 shows the predictions for *hund* 'dog', *vovve* 'doggy' (childish or endearing term for 'dog') and *byracka* 'mutt' (derogatory term for 'dog').

15.5.4 Annotation editor

Lärka contains an annotation editor that can be used for XML markup of textbooks. The editor provides an intuitive menu that makes adding XML tags easy. The editor keeps track of current settings in order to make adding new elements as easy as possible. It also automatically increments lesson counters and other counters. The editor offers the possibility to download the annotated text as an XML file. The current version of the editor also includes the possibility to save one's progress and continue working on it at a later moment in time without the need to login. The SweLL corpus pilot project (Volodina et al. 2016a) and the COCTAILL corpus project (Volodina et al. 2014a) used



Write a lemma

byracka

Select a part-of-speech

noun

Receptive ● Productive ○ Both ○

Results

Word	POS	ROP	Predicted level
byracka	NN	receptive	B2
vovve	NN	receptive	A2
hund	NN	receptive	A1

Figure 15.11: User interface for lexical complexity prediction

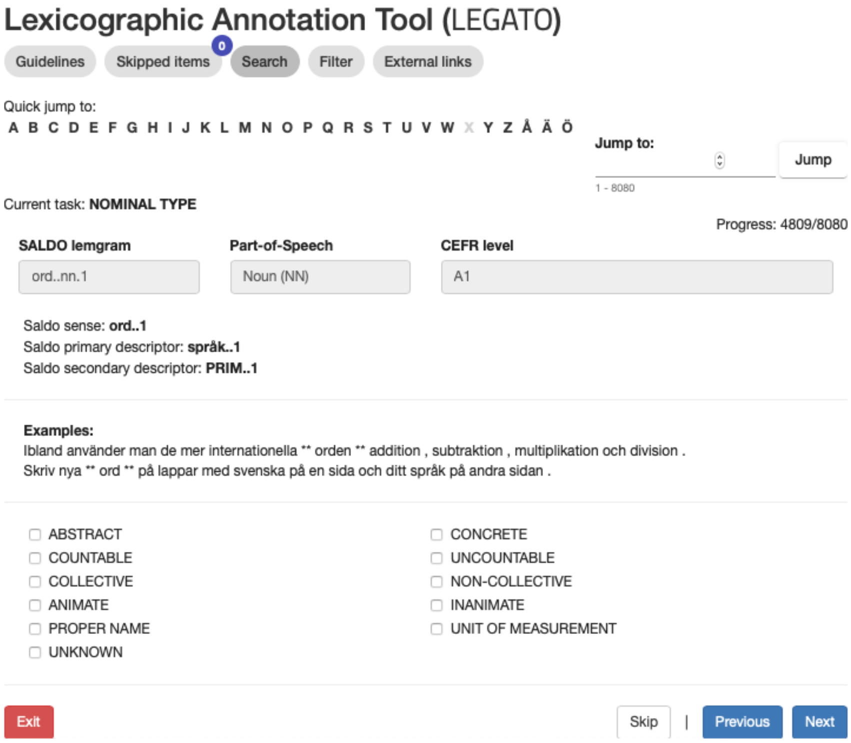
a previous version of the annotation editor to achieve consistent XML markup of essays and course books as well as to simplify the annotation process by providing an intuitive and intelligent user interface.

15.5.5 Lexicographic annotation tool

Another annotation tool that has recently been added to Lärka is the Lexicographic Annotation Tool, Legato. This tool can be used to annotate words or word senses on different lexicographic levels. Figure 15.12 shows the tool in the ‘register’ annotation mode. Here, the annotator is presented with a SALDO sense (viz. *gammal* ‘old’), its part-of-speech (adjective) and the predicted CEFR level (A1). In addition, the tool shows the primary and secondary SALDO descriptors, if available. As different senses of a word can still be ambiguous as to the category to be annotated, we also show an example sentence where the word sense is highlighted, in this case surrounded by two asterisks (**). The example sentences have been selected to be of the same CEFR level as the word sense in question.

The main part of the interface shows the annotation possibilities. In the example shown, different options for register are shown. The annotator can select none, one, or more than one of these possibilities.

Finally, using the buttons at the bottom, annotators can leave the interface to annotate either another lexicographic category or to stop annotating altogether. Items can be skipped if the annotator is unsure about the annotation. In this case, the item will be added to the list of skipped items which can be accessed



Lexicographic Annotation Tool (LEGATO)

Current task: **NOMINAL TYPE**

Progress: 4809/8080

SALDO lemgram	Part-of-Speech	CEFR level
ord..nn.1	Noun (NN)	A1

Saldo sense: **ord..1**
Saldo primary descriptor: **språk..1**
Saldo secondary descriptor: **PRIM..1**

Examples:
Ibland använder man de mer internationella ** orden ** addition , subtraktion , multiplikation och division .
Skriv nya ** ord ** på lappar med svenska på en sida och ditt språk på andra sidan .

- ☐ ABSTRACT
- ☐ COUNTABLE
- ☐ COLLECTIVE
- ☐ ANIMATE
- ☐ PROPER NAME
- ☐ UNKNOWN
- ☐ CONCRETE
- ☐ UNCOUNTABLE
- ☐ NON-COLLECTIVE
- ☐ INANIMATE
- ☐ UNIT OF MEASUREMENT

Figure 15.12: Lexicographic annotation tool Legato

by clicking the button on top next to the ‘Guidelines’ button. This opens up a side menu which shows all the skipped items. By clicking on any of these items, the interface returns to the item in question. The interface also offers a search functionality which makes searching through the list of items easy.

In addition, the interface keeps track of different annotators and their progress across different annotation categories. Thus, if an annotator annotates ten items in ‘morphology’, then returns to the main screen and annotates ten items in ‘nominal gender’, then returns to morphology, the interface will resume at item number eleven. This also works across sessions. Thus, annotation does not have to be done in one fell swoop but can be done intermittently. The skipped items are also saved per annotator and category. For example, if annotator A skips *gammal* ‘old’ in ‘register’ but not in ‘morphology’, it will turn up for annotator A under ‘register’ until it is resolved. All data is saved to a data base on the server.

Besides fully manual annotation, the tool also offers a semi-automatic annotation mode where some of the values have been automatically extracted by linking together various resources. In this annotation mode, if values have been found, the annotator’s task is to check whether the values are correct and correct them if necessary. If no values have been found, the annotator proceeds as in manual mode.

15.6 Ongoing work and planned extensions

Besides the activities described in this paper, the addition of new exercise formats and the implementation of a diagnostic placement test are currently under development. In the near future we plan to add a login functionality as well as an infrastructure to log more specific user data. This would enable us to create a valuable resource for modeling learners (e.g. L1-specific errors, learners’ development over time) and to offer adaptive exercises.

16 Particle verb exercise

This publication is discussed in section 7.1.2.4.

This chapter is a postprint version of the following publication:

Alfter, David and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa*. Turku, 2019 . Linköping University Electronic Press.

Abstract

In this paper, we present a prototype for an online exercise aimed at learners of English and Swedish that serves multiple purposes. The exercise allows learners of these languages to train their knowledge of particle verbs receiving clues from the exercise application. At the same time, we collect information which will help us judge the accuracy of our graded word lists. As resources, we use lists with annotated levels from the proficiency scale defined by the Common European Framework of Reference (CEFR) and a multilingual corpus with syntactic dependency relations and word alignments for all language pairs. From the latter resource, we extract translation equivalents for particle verb constructions together with a list of parallel corpus examples that are used as clues in the exercise.

16.1 Introduction

Combinations of verbs and particles have been studied extensively in various aspects, e.g. particle placement with regard to cognitive processes (Gries 2003), the relation between syntactical and semantic structure (Roßdeutscher 2011) and their compositionality with respect to syntactic argument structure

(Bott and Schulte im Walde 2015). In the field of language learning, verb-particle combinations have been investigated in matters of their use of language learners of English (EFL) (Gilquin 2015; Liao and Fukuya 2004), also in comparison to native language speakers (Schneider and Gilquin 2016) and with regard to pedagogical suggestions for language learning and teaching (Gardner and Davies 2007).

The term 'phrasal verb' is used in most publications to refer to an English verb-particle combination that "behaves as a semantic unit" (Gilquin 2015), while for other (mostly Germanic) languages term such as 'verb-particle constructions', 'verb-particle expressions' (Toivonen 2002) or simply 'particle verbs' prevail (Zeller 2001). Dehé (2015) compares particle verbs in Germanic languages and regards these terms as synonyms. We will thus refer to construction of verb and particle as particle verbs.

Particle verbs are especially difficult for learners since they present no discernible pattern in the selection of the particle. Gardner and Davies (2007) observe that "many nonnative English speakers actually avoid using phrasal verbs altogether, especially those learners at the beginning and intermediate levels of proficiency." Not all verbs and particles are equally likely to take part in particle verbs. In English, "a number of lexical verbs such as take, get, come, put and go are particularly productive and frequent when they combine with adverbial particles" (Deshors 2016). Gardner and Davies (2007) recommend learners to memorize those verbs and particles that occur frequently in verb-particle combinations.

Recently, so-called Games With A Purpose (GWAPs) (Lafourcade, Joubert and Le Brun 2015) have been used to collect information from players while offering a ludic interface that promotes participation. For example, JeuxDeMots (Lafourcade and Joubert 2008; Lafourcade 2007) has been used to find lexico-semantic relations between words, ZombiLingo (Fort, Guillaume and Chastant 2014) for the annotation of dependency syntax in French corpora, RigorMortis (Fort et al. 2018) for the identification of multi-word expression by (untrained) learners, relying on their subjective opinion.

With the six reference levels of the Common European Framework of Reference (CEFR) (Council of Europe 2001), henceforth CEFR levels, we can classify learners according to their level of proficiency. In Section 16.2.1, we introduce two resources that we build upon, which provide lists of vocabulary units together with their estimated distribution over CEFR levels. In Section 16.2.2, we explain how we look up translation equivalents in several languages in a word-aligned multiparallel corpus, followed by a manual reassessment step described in Section 16.2.4.

In continuation, we present an application that implements a gamified exercise based on particle verbs in English and Swedish, their translation equivalents and corpus examples that demonstrate their use in authentic translations (Section 16.3). Learners playing the game try to not lose while the game automatically adapts to their current predicted knowledge level. The application keeps track of decisions taken by the user during the course of the game to provide them with feedback regarding their language skills, and points to potential weaknesses and (language-specific) factors for confusions. At the same time, we expect that a sufficiently large collection of decisions will help us assess the CEFR levels of our lexical resources and provide insights for future extensions.

16.2 Data preparation

We extract particle verbs for CEFR levels from A1 to C1 from two lexical resources, one for English and one for Swedish.[85] For each particle verb that we find in these resources, we look up potential translation variants for several other languages, from a large multilingual word-aligned corpus. Since word alignment is less reliable when it comes to function words, we need to review the lists of translation variants and adjust word order and missing function words in multiword variants manually.

16.2.1 Lexical resources

The CEFRLex project[86] offers lists of expressions extracted from graded textbook corpora for different languages. The languages currently available are French, Swedish and English. For this project, we use the Swedish list SVALex François et al. (2016) and the English list EFLLex Dürlich and François (2018) from the CEFRLex project. Each resource lists single-word and multiword expressions, as recognized by a syntactic parser, and their frequency in textbooks of different CEFR levels. Table 16.1 shows examples from the EFLLex list.

We extract particle verbs from both lists. For EFLLex, we use regular expressions to match all two-word expressions that are tagged as verbs. Manual inspection of the results shows that most expressions extracted this way are indeed particle verbs; we only had to exclude four expressions.[87]

For SVALex, we consider the subset of expressions tagged as verbal multiword expressions. Since not all verbal multi-word expressions are particle

[85] No particle verb has been classified as C2.

[86] `http://cental.uclouvain.be/cefrlex/`

[87] Those are ‘finger count’, ‘deep fry’, ‘go lame’ and ‘tap dance’, which use other part of speech than particles.

Expression	PoS	A1	A2	B1	B2	C1	C2
video	noun	65.19	0	67.87	81.76	111.06	90.93
write	verb	758.66	1421.51	1064.47	682.26	1104.72	1053.96
empty	adjective	0	28.83	28.65	102.29	37.84	61.88
shopping center	noun	0	45.12	9.80	0	15.50	11.45
dream up	verb	0	0	0	0	0.82	0.24

Table 16.1: Example entries from EFLLex.

verbs, we cross-check for the existence of each expression in the upcoming version of Saldo,[88] which includes particle verbs. Upon manual inspection of the resulting list we removed two reflexive particle verbs.[89] In total, we extracted 221 English and 362 Swedish particle verbs. As we are, among other things, interested in seeing how CEFR levels correlate with self-proclaimed proficiency, we assign each particle verb the CEFR level at which it first occurs in the respective resource, as has been previously done in various other experiments Gala, François and Fairon (2013); Gala et al. (2014); Alfter et al. (2016); Alfter and Volodina (2018).

16.2.2 Translation equivalents from parallel corpus data

The exercise is based on finding the correct particle for a particle verb in the target language based on translations in the source language. In other words, it means that, for example, learners of Swedish (target language) with knowledge of English (source language) will have to guess Swedish particle verbs based on English translations. For identifying translation equivalents in multiple languages, we use the Sparcling corpus (Graën 2018; Graën et al. 2019b), which, in addition to standard annotation such as part-of-speech tagging, features dependency relations from syntactic parsing in a number of languages (including English and Swedish) and bilingual word alignment for all language pairs. We use dependency relations to identify pairs of particles and their head verb matching the list that we extracted from EFLLex and SVALex.

For each occurrence of those pairs in the corpus, we look up aligned tokens in all other languages available to spot corresponding translation equivalents. We then filter the aligned tokens for content words, that is, in terms of universal part-of-speech tags Petrov, Das and McDonald (2012), verbs, nouns, adjectives or adverbs. Functional parts of multi-word expressions are noto-

[88] `https://spraakbanken.gu.se/eng/resource/saldo`

[89] To wit ‘ge sig ut’ *go out* and ‘klamra sig fast’ *cling to*.

riously misaligned if the syntactic patterns of the corresponding expressions differ. For instance, English ‘to cry out (for sth.)’ can be expressed in Spanish with the fixed expression ‘pedir (algo) a gritos’. In this case, we often see ‘cry’ aligned with ‘pedir’ and ‘gritos’, and the particle ‘out’ with the preposition ‘a’. A similar expression is ‘llevar (algo) a cabo’ *‘get through (sth.)’*, where ‘carry’ is aligned with ‘llevar’ and ‘out’ with ‘cabo’; the preposition ‘a’ often remains unaligned in this case.

By filtering out function words, we systematically miss any preposition, determiner or particle that forms part of the equivalent expression. Not filtering them out, on the other hand, leads to considerably noisier lists. The missing functional parts need to be added back later and the set of lemmas needs be put in the preferred lexical order (see Section 16.2.4). We retrieve lemmas of the aligned tokens as a set, disregarding their relative position in the text, and calculate frequencies for each translation equivalent. Translation equivalents are most frequently single verbs. The Swedish particle verb ‘ha kvar’ (literally *‘have left’*), for instance, is aligned to the English verbs ‘retain’ 49 times, to ‘maintain’ 31 times and to ‘remain’ 26 times.

16.2.3 Example sentence selection

Alongside other options (see Section 16.3), we want to provide learners with authentic examples where the given particle verb is used as translation of a particular expression in another language. We typically find several example sentences per translation correspondence in the Sparcling corpus. The question now is how to select the most adequate one for the respective learner. In previous works, we have used the length of the candidate sentence pair as ranking criterion, downgrading those pairs that showed a substantial deviation in length Schneider and Graën (2018); Clematide, Graën and Volk (2016).

While there is a substantial amount of previous work on finding good example sentences for use in dictionaries (e.g. GDEX Kilgarriff et al. (2008)) or for language learners (e.g. HitEx Pilán, Volodina and Borin (2017)), most of the features they use are language-specific, such as blacklists, ‘difficult’ vocabulary, or recognizing and excluding anaphoric expressions without referent in the same sentence.

For the purpose of this study, we have thus opted for a simple heuristics which works well across a number of different languages. We use sentence length and a weighted measure for lexical proficiency required to understand the target language sentence (since we do not have gradings for most of the source languages).

16.2.4 Manual revision

Manual correction involves the removal of irrelevant translations, the re-ordering of words, in case a particle verb has been aligned to multiple other words, and the insertion of missing words into the translation variants (as in ‘llevar *a* cabo’). In addition, we judge example sentences with regard to adequacy.

While the translation candidate extraction could be restricted to allow only verbal translations for particle verbs, this is a constraint that we do not want to impose. Indeed, certain languages tend towards more nominal ways of expression while other languages tend towards more verbal ways of expression Azpiazu Torres (2006). Thus, imposing such a constraint could possibly induce non-idiomatic or unnatural translation candidates.

Having multiple part-of-speech possibilities for translation variants also allows us to potentially control the difficulty of the exercise by only giving verbal translation variants to beginners while, as the learner progresses and improves, other part-of-speech variants could be included.

16.3 Crowdsourcing and gamification

We use our gamified system to assess knowledge of language learners in their L2 (English or Swedish), and to judge the accuracy of the automatically assigned CEFR labels. The game presents one base verb each round, together with a list of particles to choose from and one initial clue in form of a translation variant for the particle verb that the player is supposed to guess. The player can gain points by choosing the right particle and loose points by choosing a wrong one. Additional clues can be traded off against points. These clues can also be example sentences in the target language or the elimination of several of the non-fitting particles.

The learner assessment is achieved by monitoring how players of certain self-proclaimed proficiency levels deal with expressions that they are supposed to master, according to the automatic CEFR level assignment method. If learners systematically struggle with expressions of their self-chosen proficiency level, we assume that they overvalued their level and provide feedback accordingly. If they show little or no difficulties in dealing with expressions deemed of their current self-proclaimed proficiency level, we assume that their actual proficiency is higher, and gradually increase the challenge by using particle verbs of higher levels and more difficult clues (e.g. less frequent translation variants).

The accuracy of the automatically assigned CEFR labels is measured by aggregating results over all players. We also take into account response times for individual exercises. Significantly large deviance from the average answering time or the average number of points used for 'trading' clues for particle verbs of the supposedly same proficiency level suggests that the particle verb in question could belong to a different level.

Before the actual game starts, learners have to choose the language that they want to train. They are also asked to indicate their mother tongue and any other languages they know, including a self-assessment of their proficiency in the respective languages (beginner, intermediate, advanced). This rough scale is translated to the levels A1 and A2, B1 and B2 and C1 respectively.

Having finished the self assessment, the learner gets a predefined amount of points, as a virtual currency. More points can be gained each round by finding the right particle for the given verb with as few clues as possible. A wrong answer is worth an equally negative amount of points that could have been gained by choosing the right answer. We employ a function to calculate the reward based on hints used and difficulty of the hints in terms of language knowledge, i.e. a clue in a lower-rated language will cost the learner less points than, for instance, in his mother tongue. The game ends when the player is out of points or the game is out of particle verbs. The final score is used to create an entry on a leaderboard.

16.4 Discussion and future work

With the development of new CEFR graded multi-word expression lists, including a wider range of expressions, the exercise can be extended to other types of expressions. With the advent of CEFR graded multi-word lists in other languages, the exercise can also be extended to encompass a more diverse set of languages.

One aspect that is not specifically addressed in this study is the issue of polysemy. Indeed, a particle verb can have multiple meanings, and thus multiple different translations. This aspect will prove problematic when the particle verbs are shown in context, as one has to ensure that both the original as well as the translation pertain to the same sense of the expression.

Another question concerns the accuracy of the automatic assignment of CEFR levels based on the method used. While we surmise that we can gain insights about the accuracy of the assigned levels through the proposed prototype, a separate investigation should be carried out. One could possibly compare the automatically assigned levels from EFLLex to the levels given in English Vocabulary Profile.[90]

[90] http://www.englishprofile.org

Acknowledgements

The authors would like to express their gratitude towards Lars Borin for making available the upcoming version of the Saldo lexical resource. This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1_184212 through the project “From parallel corpora to multilingual exercises – Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications”.

BIBLIOGRAPHY

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp and Geoffrey Irving 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Ädel, Annelie and Britt Erman 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for specific purposes* 31 (2): 81–92.

Ahlberg, Malin, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher and Jonatan Uppström 2016. Språkbanken's Open Lexical Infrastructure. *SLTC 2016*.

Ahmad, Tuan Sarifah Aini Syed, Anealka Aziz Hussin and Ghazali Yusri 2018. B1 Single Sentence Descriptors. *International Journal of Modern Languages and Applied Linguistics* 1 (1): 23–27.

Aker, Ahmet, Mahmoud El-Haj, M-Dyaa Albakour, Udo Kruschwitz et al. 2012. Assessing Crowdsourcing Quality through Objective Tasks. *LREc*, 1456–1461. Citeseer.

Alfter, David 2015. Language Segmentation. Master's thesis, Universität Trier.

Alfter, David, Yuri Bizzoni, Anders Agebjörn, Elena Volodina and Ildikó Pilán 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 1–7. Linköping University Electronic Press.

Alfter, David, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann and Elena Volodina 2019. Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. *Selected papers from the CLARIN Annual Conference 2018*, 1–14. Linköping University Electronic Press.

Alfter, David and Johannes Graën 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 321–326.

Alfter, David and Ildikó Pilán 2018. SB@GU at the Complex Word Identification Task 2018. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*.

Alfter, David and Elena Volodina 2018. Towards Single Word Lexical Complexity Prediction. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, 79–88.

Allén, Sture 1971. *Nusvensk frekvensordbok baserad på tidningstext: Frequency Dictionary of Present-day Swedish based on Newspaper Material*. Almqvist & Wiksell.

Allén, Sture 1972. Tiotusen i topp. *Sweden: Almqvist & Wiksell.*

Allwood, Jens 1999. Talspråksfrekvenser. *Gothenburg Papers in Theoretical Linguistics S*, vol. 21.

Alpaydin, Ethem 2020. *Introduction to machine learning*. MIT press.

Andersson, Annika, Susan Sayehli and Marianne Gullberg 2018. Language background affects online word order processing in a second language but not offline. *Bilingualism: Language and Cognition*, pp. 1–24.

Artstein, R. and M. Poesio 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 (4): 555–596.

Avdiu, Drilon, Vanessa Bui, Klára Ptacinová Klimci et al. 2019. Predicting learner knowledge of individual words using machine learning. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 1–9. Linköping University Electronic Press.

Azpiazu Torres, Susana 2006. Stylistic-contrastive analysis of nominality and verbality in languages. *Studies in Contrastive Linguistics: Proceedings of the 4th International Contrastive Linguistics Conference*, 69–77. Universidade de Santiago de Compostela.

Bachman, Lyle F 1990. *Fundamental considerations in language testing*. Oxford University Press.

Bachman, Lyle F and Adrian S Palmer 2010. *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Baldwin, Timothy and Su Nam Kim 2010. Multiword expressions. *Handbook of natural language processing* 2: 267–292.

Baldwin, Timothy and Aline Villavicencio 2002. Extracting the unextractable: A case study on verb-particles. *Proceedings of the 6th conference on Natural language learning-Volume 20*, 1–7. Association for Computational Linguistics.

Bauer, Laurie 1983. *English word-formation*. Cambridge university press.

Bauer, Laurie and Paul Nation 1993. Word families. *International journal of Lexicography* 6 (4): 253–279.

Beacco, Jean Claude, Béatrice Blin, Emmanuelle Houles, Sylvie Lepage and Patrick Riba 2011. *Niveau B1 pour le français (apprenant/utilisateur indépendant): niveau seuil*. Editions Didier.

Beacco, Jean-Claude, Simon Bouquet and Rémy Porquier 2004. *Niveau B2 pour le français: un référentiel*. Editions Didier.

Beacco, Jean-Claude, Sylvie Lepage, Rémy Porquier and Patrick Riba 2008. *Niveau A2 pour le français: un référentiel*. Editions Didier.

Beacco, Jean-Claude and Rémy Porquier 2007. *Niveau A1 pour le français (utilisateur/apprenant élémentaire): un référentiel*. Editions Didier.

Beinborn, Lisa, Torsten Zesch and Iryna Gurevych 2014. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics* 165 (2): 136–162.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent and Christian Jauvin 2003. A neural probabilistic language model. *Journal of machine learning research* 3 (Feb): 1137–1155.

Bengio, Yoshua, Ian Goodfellow and Aaron Courville 2017. *Deep learning*. Volume 1. MIT press Massachusetts, USA.

Bhagoliwal, B 1961. Readability formulae: Their reliability, validity and applicability in Hindi. *Journal of Education and Psychology* 19: 13–26.

Bhalla, Vishal and Klara Klimcikova 2019. Evaluation of automatic collocation extraction methods for language learning. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 264–274.

Bhatia, Archna, Choh Man Teng and James Allen 2017. Compositionality in verb-particle constructions. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 139–148.

Bird, Steven, Ewan Klein and Edward Loper 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bird, Steven and Edward Loper 2004. NLTK: the natural language toolkit. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 69–72. Association for Computational Linguistics.

Björnsson, Carl Hugo 1968. *Läsbarhet*. Liber.

Blache, Philippe 2011. A computational model for linguistic complexity.

Bloomfield, Leonard 1914. Sentence and word. *Transactions and Proceedings of the American Philological Association*, Volume 45, 65–75. JSTOR.

Bloomfield, Leonard 1935. Linguistic aspects of science. *Philosophy of science* 2 (4): 499–517.

Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5: 135–146.

Bolander, Maria 1989. Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. *Bilingualism across the lifespan: Aspects of acquisition, maturity, and loss*, pp. 73–86.

Bongers, Herman 1947. *The History and Principles of Vocabulary Control as it affects the Teaching of Foreign Languages in general and of English in particular*. Volume 1. Wocopi.

Borin, Lars . Is a cross-linguistic typology of multiword expressions useful or even possible?

Borin, Lars 2005. Mannen är faderns mormor: Svenskt associationslexikon reinkarnerat. *LexicoNordica*, no. 12.

Borin, Lars and Markus Forsberg 2014. Swesaurus; or, The Frankenstein approach to Wordnet construction. *Proceedings of the Seventh Global Wordnet Conference*, 215–223.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer and Anne Schumacher 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, 17–18.

Borin, Lars, Markus Forsberg and Lennart Lönngren 2008. The hunting of the BLARK–SALDO, a freely available lexical database for Swedish language technology. *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, no. 7:21–32.

Borin, Lars, Markus Forsberg and Lennart Lönngren 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation* 47 (4): 1191–1211.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson and Jonatan Uppström 2013. The lexical editing system of Karp. *Proceedings of the eLex 2013 conference*, 503–516.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson and Jonatan Uppström 2012. The open lexical infrastructure of Spräkbanken. *LREc*, 3598–3602.

Borin, Lars, Markus Forsberg and Johan Roxendal 2012. Korp-the corpus infrastructure of Spräkbanken. *PLRroceedings of Ec*, 47 20124–478.

Borin, Lars and Anju Saxena 2004. Grammar, incorporated. Peter Juel Henrichsen (ed.), *CALL for the Nordic languages*, 125–145. Copenhagen: Samfundslitteratur.

Borin, Lars, Nina Tahmasebi, Elena Volodina, Stefan Ekman, Caspar Jordan, Jon Viklund, Beáta Megyesi, Jesper Näsman, Anne Palmér, Mats Wirén, Kristina Nilsson Björkenstam, Gintarė Grigonytė and Gusta 2017. Swe-Clarin: Language Resources and Technology for Digital Humanities. *Digital Humanities 2016. Extended Papers of the International Symposium on Digital Humanities (DH 2016) Växjö, Sweden, November, 7-8, 2016. Edited by Koraljka Golub, Marcelo Milra*, vol. Vol-2021.

Bott, Stefan and Sabine Schulte im Walde 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, 34–39.

Bourque, Yves Stephen 2014. Toward a typology of semantic transparency: The case of French compounds. Ph.D. diss., University of Toronto.

Brabham, Daren C 2013. *Crowdsourcing*. MIT press Massachusetts, USA.

Breiman, Leo 2001. Random Forests. *Machine Learning* 45 (1): 5–32 (October).

Brezina, Vaclav and Dana Gablasova 2015. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics* 36 (1): 1–22.

Brooke, Julian, Vivian Tsang, David Jacob, Fraser Shein and Graeme Hirst 2012. Building readability lexicons with unannotated corpora. *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, 33–39.

Browne, Charles 2014. A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction* 3 (2): 1–10.

Brugmann, Karl 1892. *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen: Bd. Wortbildungslehre (Stammbildungs-und Flexionslehre). 1. Hälfte Vorbemerkungen. Nominalcomposita. Reduplicierte Nominalbildungen. Nomina mit Stammbildenden Suffixen. Wurselnomina*. Volume 2. Karl J Trübner.

Buhrmester, Michael, Tracy Kwang and Samuel D Gosling 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? Alan E Kazdin (ed.), *Methodological issues and strategies in clinical research*. American Psychological Association.

Bulté, Bram and Alex Housen 2012. Defining and operationalising L2 complexity. Alex Housen, Folkert Kuiken and Ineke Vedder (eds), *Dimensions* *of L2 Performance and Proficiency*, Volume 32 of *Language Learning & Language Teaching*, 21–46. John Benhamins.

Burger, Harald 1998. *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt Verlag.

Burstein, Jill, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers and Kelsey Dreier 2017. Generating Language Activities in Real-Time for English Learners using Language Muse. *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 213–215. ACM.

Burstein, Jill and John Sabatini 2016. The Language Muse Activity Palette. *Adaptive educational technologies for literacy instruction*, pp. 275–280.

Cai, Guoyong and Binbin Xia 2015. Convolutional neural networks for multimedia sentiment analysis. *Natural Language Processing and Chinese Computing*, 159–167. Springer.

Calzolari, Nicoletta, Charles J Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod and Antonio Zampolli 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. *LREc.*

Cambridge University Press 2015. English Vocabulary Profile. `https://www.englishprofile.org/wordlists`. Accessed: 2019-11-11.

Capel, Annette 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal* 1 (1): 1–11.

Capel, Annette 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal* 3: 1–14.

Capel, Annette 2015. The English Vocabulary Profile. Julia Harrison and Fiona Barker (eds), *English Profile in Practice*, 9–27. Cambridge University Press.

Carlsen, Cecilie 2012. Proficiency level—A fuzzy variable in computer learner corpora. *Applied Linguistics* 33 (2): 161–183.

Carroll, John Bissell, Peter Davies and Barry Richman 1971. *The American Heritage word frequency book.* Houghton Mifflin.

Cavnar, William B, John M Trenkle et al. 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, Volume 161175. Citeseer.

Chamberlain, Jon, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade and Massimo Poesio 2013. Using games to create language resources: Successes and limitations of the approach. *The People's Web Meets NLp*, 3–44. Springer.

Chamberlain, Jon, Massimo Poesio and Udo Kruschwitz 2008. Phrase detectives: A web-based collaborative annotation game. *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, 42–49.

Chapelle, Carol A and Dan Douglas 2006. *Assessing Language through Computer Technology*. Cambridge Language Assessment. Cambridge University Press.

Chen, Yu-Hua and Paul Baker 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14 (2): 30–49.

Chen, Yu-Hua and Paul Baker 2016. Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics* 37 (6): 849–880.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches.

Chollet, François et al. 2015. Keras. `https://keras.io`.

Chollet, Francois 2018. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG.

Chomsky, Noam and Morris Halle 1965. Some controversial questions in phonological theory. *Journal of linguistics* 1 (2): 97–138.

Chrzan, Keith and Megan Peitz 2019. Best-Worst Scaling with many items. *Journal of choice modelling* 30: 61–72.

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim and Jong C Park 2013. Enhancing readability of web documents by text augmentation for deaf people. *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, 1–10.

Čibej, Jaka, David Alfter, Iztok Kosem and Elena Volodina . Multi-word expressions and language learning: validity of a crowdsourcing approach. In preparation.

Cieślicka, Anna B 2015. Idiom acquisition and processing by second/foreign language learners.

Clematide, Simon, Johannes Graën and Martin Volk 2016. Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora. Gloria Corpas Pastor (ed.), *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia computacional y basada en corpus: perspectivas monolingües y multilingües*, 447–455. Tradulex.

Cohen, Adam 2011. FuzzyWuzzy: Fuzzy string matching in Python. *ChairNerd Blog*, vol. 22.

Cohen, Yoav and Anat Ben-Simon 2011. The Hebrew language project: Automated essay scoring & readability analysis. *IAEA Annual Conference*. Vienna, Austria.

Coltheart, Max 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33 (4): 497–505.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Accessed 09.03.2019 from `www.coe.int/lang-cefr`.

Council of Europe 2020. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume. Accessed 29.10.2020 from `www.coe.int/lang-cefr`.

Cowie, Anthony P 1992. Multiword lexical units and communicative language teaching. *Vocabulary and applied linguistics*, 1–12. Springer.

Coxhead, Averil 1998. *An academic word list*. Volume 18. School of Linguistics and Applied Language Studies.

Coxhead, Averil 2000. A new academic word list. *TESOL quarterly* 34 (2): 213–238.

Coxhead, Averil 2011. The academic word list 10 years on: Research and teaching implications. *Tesol Quarterly* 45 (2): 355–362.

Croft, William 2007. Construction grammar. *The Oxford handbook of cognitive linguistics*.

Crossley, Scott, Tom Salsbury and Danielle McNamara 2010. The development of polysemy and frequency use in English second language speakers. *Language Learning* 60 (3): 573–605.

Csikszentmihalyi, Mihaly 1975/2000. *Beyond boredom and anxiety*. Jossey-Bass.

Culbertson, Gabriel, Solace Shen, Erik Andersen and Malte Jung 2017. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 286–296.

Curtotti, Michael, Eric McCreath, Tom Bruce, Sara Frug, Wayne Weibel and Nicolas Ceynowa 2015. Machine Learning for Readability of Legislative Sentences. *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, 53–62. New York, NY, USA: Association for Computing Machinery.

Cutler, Anne 1983. Lexical complexity and sentence processing.

Dale, Edgar and Jeanne S Chall 1949. The concept of readability. *Elementary English* 26 (1): 19–26.

Davis, Colin J 2005. N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior research methods* 37 (1): 65–70.

De Belder, Jan, Koen Deschacht and Marie-Francine Moens 2010. Lexical simplification.

De Belder, Jan and Marie-Francine Moens 2010. Text simplification for children. *Proceedings of the SIGIR workshop on accessible search systems*, 19–26. ACM; New York.

De Hertog, Dirk and Anaïs Tack 2018. Deep Learning Architecture for Complex Word Identification. *Thirteenth Workshop of Innovative Use of NLP for Building Educational Applications*, 328–334. Association for Computational Linguistics (ACL); New Orleans, Louisiana.

Dehé, Nicole 2015. Particle verbs in Germanic. Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen and Franz Rainer (eds), *Word-formation: an international handbook of the languages of Europe*, Volume 1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, 40, 611–626. De Gruyter Mouton.

Deléger, Louise and Pierre Zweigenbaum 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, 2–10.

Dell'Orletta, Felice, Simonetta Montemagni and Giulia Venturi 2011. Read-it: Assessing readability of italian texts with a view to text simplification. *Proceedings of the second workshop on speech and language processing for assistive technologies*, 73–83. Association for Computational Linguistics.

Dell'Orletta, Felice, Martijn Wieling, Giulia Venturi, Andrea Cimino and Simonetta Montemagni 2014. Assessing the readability of sentences: which corpora and features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 163–173.

Del Río Gayo, Iria 2019. Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning*. Linköping University Proceedings.

Derwing, Bruce L 1976. Morpheme recognition and the learning of rules for derivational morphology 1. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 21 (1): 38–66.

Deshors, Sandra C. 2016. Inside phrasal verb constructions: A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations. *International Journal of Learner Corpus Research* 2 (1): 1–30.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, Siobhan 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Dietterich, Thomas G. 1997. Machine learning research: Four current directions. *AI Magazine* 18 (4): 97–136.

Díez-Bedmar, María Belén 2012. The use of the common european framework of reference for languages to evaluate compositions in the english exam section of the university admission examination. *Revista de Educación* 357: 55–79.

Dixon, Robert MW, Alexandra Y Aikhenvald et al. 2002. Word: a typological framework. *Word: A cross-linguistic typology*, pp. 1–41.

Dong, Fei, Yue Zhang and Jie Yang 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. *Proceedings of the 21st conference on computational natural language learning (conll 2017)*, 153–162.

Dos Santos, Cicero and Maira Gatti 2014. Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 69–78.

Dürlich, Luise and Thomas François 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. *11th International Conference on Language Resources and Evaluation (LREC 2018).*

Dyer, Chris, Victor Chahuneau and Noah A. Smith 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 644–649. Association for Computational Linguistics (ACL).

Ehara, Yo, Issei Sato, Hidekazu Oiwa and Hiroshi Nakagawa 2012. Mining words in the minds of second language learners: learner-specific word difficulty. *Proceedings of COLING 2012*, 799–814.

Ehara, Yo, Issei Sato, Hidekazu Oiwa and Hiroshi Nakagawa 2018. Mining

Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing* 26: 267–275.

Ellis, Rod 2003. *Task-based language learning and teaching*. Oxford University Press.

Eskildsen, Søren W 2009. Constructing another language—usage-based linguistics in second language acquisition. *Applied linguistics* 30 (3): 335–357.

EU Commission 2016. General data protection regulation. *Official Journal of the European Union* 59: 1–88.

Evert, Stefan and Brigitte Krenn 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language* 19 (4): 450–466.

Fazly, Afsaneh and Suzanne Stevenson 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 9–16.

Fellbaum, Christiane 1998. WordNet: an electronic lexical database. *Language, Speech, and Communication*.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni and Silvia Bernardini 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. *Proceedings of the 4th Web as Corpus Workshop (WAC) – Can we beat Google?*, 47–54.

Flesch, Rudolph 1948. A new readability yardstick. *Journal of applied psychology* 32 (3): 221.

Forsberg Lundell, Fanny 2020. Krävande krav. Vad ska språkkrav vara bra för?

Forsbom, Eva 2006. A swedish base vocabulary pool. *Swedish Language Technology conference, Gothenburg*.

Fort, Karën 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Fort, Karën, Bruno Guillaume and Hadrien Chastant 2014. Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. *Gamification for Information Retrieval Workshop (GamifIR)*.

Fort, Karën, Bruno Guillaume, Mathieu Constant, Nicolas Lefebvre and Yann-Alan Pilatte 2018. "Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform. *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, 207–213.

Foss, Donald J and Charles M Jenkins 1973. Some effects of context on the comprehension of ambiguous sentences. *Journal of verbal learning and verbal behavior* 12 (5): 577–589.

François, Thomas 2012. Lexical and syntactic complexities: a difficulty model for automatic generation of language exercises in FFL. Ph.D. diss., Université.

François, Thomas and Barbara De Cock 2018. ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*.

François, Thomas and Cédrick Fairon 2012. An AI readability formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–477. Association for Computational Linguistics.

François, Thomas and Cédrick Fairon 2017. Introducing NT2Lex: A Machine-readable CEFR-graded Lexical Resource for Dutch as a Foreign Language. *Computational Linguistics in the Netherlands 27 (CLIN27)*.

François, Thomas, Núria Gala, Patrick Watrin and Cédrick Fairon 2014. FLELex: a graded lexical resource for French foreign learners. *LREc*, 3766–3773.

François, Thomas, Elena Volodina, Ildikó Pilán and Anaïs Tack 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. *LREc*.

François, Thomas and Patrick Watrin 2011. On the contribution of MWE-based features to a readability formula for French as a foreign language. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 441–447.

Gala, Núria, Thomas François, Delphine Bernhard and Cédrick Fairon 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. *TALN 2014*, 91–102.

Gala, Núria, Thomas François and Cédrick Fairon 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Garcia, Ignacio 2013. Learning a language for free while translating the web. does duolingo work? *International Journal of English Linguistics* 3 (1): 19.

Gardner, Dee and Mark Davies 2007. Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis. *TESOL quarterly* 41 (2): 339–359.

Gellerstam, Martin 1999. LEXIN-lexikon för invandrare. *LexicoNordica*, no. 6.

Geurts, Pierre, Damien Ernst and Louis Wehenkel 2006. Extremely randomized trees. *Machine Learning* 63 (1): 3–42 (Apr).

Gilquin, Gaëtanelle 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11 (1): 51–88.

Goldberg, Adele E 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Gooding, Sian and Ekaterina Kochmar 2018. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. *Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications*, 184–194.

Graën, Johannes 2018. Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning. Ph.D. diss., University of Zurich.

Graën, Johannes, David Alfter and Gerold Schneider 2020. Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. *Proceedings of The 12th Language Resources and Evaluation Conference*, 346–355. Marseille, France: European Language Resources Association.

Graën, Johannes, Dolores Batinic and Martin Volk 2014. Cleaning the Europarl Corpus for Linguistic Applications. *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 222–227. Stiftung Universität Hildesheim.

Graën, Johannes, Tannon Kew, Anastassia Shaitarova and Martin Volk 2019a. Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection. Peter Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen and Caroline Iliadi (eds), *Challenges in the Management of Large Corpora (CMLC)*. Leibniz-Institut für Deutsche Sprache.

Graën, Johannes, Tannon Kew, Anastassia Shaitarova and Martin Volk 2019b. Modelling Large Parallel Corpora:The Zurich Parallel Corpus Collection. Piotr Bański et al. (eds), *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC)*.

Graën, Johannes and Gerold Schneider 2020. Exploiting multiparallel corpora as measure for semantic relatedness to support language learners. David Levey (ed.), *Strategies and Analyses of Language and Communication in Multilingual and International Contexts*. Cambridge Scholars Publishing.

Graesser, Arthur C, Danielle S McNamara, Max M Louwerse and Zhiqiang

Cai 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36 (2): 193–202.

Granger, Sylviane 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14 (1): 58–72.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot 2009. International corpus of learner English.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomas Mikolov 2018. Learning Word Vectors for 157 Languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Grefenstette, Gregory and Pasi Tapanainen 1994. What is a word, what is a sentence?: problems of Tokenisation. Technical Report, Rank Xerox Research Centre, Meylan, France.

Gries, Stefan Thomas 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement.* Open Linguistics. A&C Black.

Gu, Jiuxiang, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai et al. 2018. Recent advances in convolutional neural networks. *Pattern Recognition* 77: 354–377.

Gu, Peter Yongqi 2003. Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ* 7 (2): 1–25.

Gurrutxaga, Antton and Inaki Alegria 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. *Proceedings of the 9th Workshop on Multiword Expressions*, 116–125.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie and Yorick Wilks 2006. A closer look at skip-gram modelling. *Lrec*, Volume 6, 1222–1225.

Guyon, Isabelle, Jason Weston, Stephen Barnhill and Vladimir Vapnik 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46 (1-3): 389–422.

Habibi, Hanieh 2019. LARA Portal: a Tool for Teachers to Develop Interactive Text Content, an Environment for Students to improve Reading Skill. *Proceedings of the 12th annual International Conference of Education, Research and Innovation*, 8221–8229.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H Witten 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11 (1): 10–18.

Hall, Mark Andrew 1999. Correlation-based feature selection for machine learning.

Haspelmath, Martin 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica* 45 (1): 31–80.

Hawkins, John A and Paula Buttery 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, vol. 1.

Hawkins, John A and Luna Filipović 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Volume 1. Cambridge University Press.

Hearst, Marti A. 1998. Support Vector Machines. *IEEE Intelligent Systems* 13 (4): 18–28 (July).

Heilman, Michael, Kevyn Collins-Thompson, Jamie Callan and Maxine Eskenazi 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 460–467.

Heimann Mühlenbock, Katarina 2013. I see what you mean—Assessing readability for specific target groups. *Data linguistica*, no. 24.

Hirschman, Isidore Isaac and David V Widder 2012. *The convolution transform*. Courier Corporation.

Hochreiter, Sepp and Jürgen Schmidhuber 1997. Long short-term memory. *Neural computation* 9 (8): 1735–1780.

Housen, Alex and Folkert Kuiken 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics* 30 (4): 461–473.

Hovy, Eduard and Julia Lavid 2010. Towards a ‘science’of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22 (1): 13–36.

Howarth, Peter 1998. Phraseology and second language proficiency. *Applied linguistics* 19 (1): 24–44.

Howcroft, David M and Vera Demberg 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 958–968.

Hsueh-Chao, Marcella Hu and Paul Nation 2000. Unknown vocabulary density and reading comprehension. *Reading in a foreign language* 13 (1): 403–30.

Huang, Yi-Ting, Hsiao-Pei Chang, Yeali Sun and Meng Chang Chen 2011. A robust estimation scheme of reading difficulty for second language learners. *2011 IEEE 11th International Conference on Advanced Learning Technologies*, 58–62. IEEE.

Hulstijn, Jan H, J Charles Alderson and Rob Schoonen 2010. Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, pp. 11–20.

Hult, Ann-Kristin, Sven-Göran Malmgren and Emma Sköldberg 2010. Lexin-a report from a recycling lexicographic project in the North. *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010).*

Hultman, Tor G and Margareta Westman 1977. *Gymnasistsvenska.* Liber.

Im Walde, Sabine Schulte, Stefan Müller and Stefan Roller 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, Volume 1, 255–265.

Jackendoff, Ray 1997. *The architecture of the language faculty.* MIT Press.

Jansson, Håkan, Sofie Johansson Kokkinakis, Judy Ribeck and Emma Sköldberg 2012. A Swedish Academic Word List: Methods and Data. *Proceedings of the 15th EURALEX International Congress*, 7–11.

Jensen, John T 1990. *Morphology: Word structure in generative grammar.* Volume 70. John Benjamins Publishing.

Jensen, Kristian TH 2009. Indicators of text complexity. Susanne Göpferich, Arnt Lykke Jakobsen and Mees Inger M. (eds), *Behind the Mind: Methods, Models and Results in Translation Process Research*, 61–80. Samfundslitteratur.

Jiang, Yuchao, Daniel Schlagwein and Boualem Benatallah 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice. *PACIs*, 180.

Johnson, Neil 2009. *Simply complexity: A clear guide to complexity theory.* Oneworld Publications.

Kalchbrenner, Nal and Phil Blunsom 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584.*

Karpov, Nikolay, Julia Baranova and Fedor Vitugin 2014. Single-sentence readability prediction in Russian. *International Conference on Analysis of Images, Social Networks and Texts*, 91–100. Springer.

Kasule, Daniel 2011. Textbook readability and ESL learners. *Reading & Writing-Journal of the Reading Association of South Africa* 2 (1): 63–76.

Kerz, Elma and Daniel Wiechmann 2017. Individual Differences in L2 Processing of Multi-word Phrases: Effects of Working Memory and Personality. *International Conference on Computational and Corpus-Based Phraseology*, 306–321. Springer.

Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi and Elena Volodina 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation* 48 (1): 121–163.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell and Pavel Rychlỳ 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX International Congress*.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers and Brad S Chissom 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report, Naval Technical Training Command Millington TN Research Branch.

Klare, George R 1974. Assessing readability. *Reading research quarterly*, pp. 62–102.

Koehn, Philipp 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit*, Volume 5, 79–86. Asia-Pacific Association for Machine Translation (AAMT).

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology. Chapter 12*. Beverly Hills, CA: Sage.

Kuhn, Tanara Zingano, Peter Dekker, Branislava Šandrih, Rina Zviel-Girshin, Špela Arhar and Tanneke Schoonheim Holdt 2019. Crowdsourcing corpus cleaning for language learning resource development. *EUROCALL Conference 2019*, 163.

Kullenberg, Christopher and Dick Kasperowski 2016. What is citizen science?–A scientometric meta-analysis. *PloS one* 11 (1): e0147152.

Kusseling, Françoise and Deryle Lonsdale 2013. A corpus-based assessment of French CEFR lexical content. *Canadian modern language review* 69 (4): 436–461.

LaBontee, Richard 2019. Strategic Vocabulary Learning in the Swedish Second Language Context. Ph.D. diss., University of Gothenburg.

Lafourcade, Mathieu 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. *7th International Symposium on Natural Language Processing (snlp)*, 7.

Lafourcade, Mathieu and Alain Joubert 2008. JeuxDeMots: un prototype ludique pour l'émergence de relations entre termes. *Journées internationales d'Analyse statistiques des Données Textuelles (JADT)*, 657–666.

Lafourcade, Mathieu, Alain Joubert and Nathalie Le Brun 2015. *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Lai, Siwei, Liheng Xu, Kang Liu and Jun Zhao 2015. Recurrent convolutional neural networks for text classification. *Twenty-ninth aaai conference on artificial intelligence*.

Langacker, Ronald W 1972. *Fundamentals of linguistic analysis*. Harcourt Brace Jovanovich New York.

Laporte, Éric 2018. Choosing features for classifying multiword expressions. Manfred Sailer and Stella Markantonatou (eds), *Multiword expressions*, 143–186. Language Science Press.

Larsson, Kent, Valentina Rosén and Carin Anderson 1985. *Frekvensordbok över svenska elevtexter*. FUMS och UDCL.

Laufer, Batia and Paul Nation 1999. A vocabulary-size test of controlled productive ability. *Language testing* 16 (1): 33–51.

Laufer, Batia and Geke C Ravenhorst-Kalovski 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a foreign language* 22 (1): 15–30.

Laufer, Batia and Donald D Sim 1985. Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign language annals* 18 (5): 405–411.

Lee, John and Chak Yan Yeung 2018. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 1–4. IEEE.

Leńko-Szymańska, Agnieszka 2015. The english vocabulary profile as a benchmark for assigning levels to learner corpus data. *Learner corpora in language testing and assessment*, pp. 115–140.

Lesk, Michael 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, 24–26. ACM.

Lesterhuis, Marije, San Verhavert, Liesje Coertjens, Vincent Donche and Sven De Maeyer 2017. Comparative judgement as a promising alternative to score competences. *Innovative practices for higher education assessment and measurement*, 119–138. IGI Global.

Lété, Bernard, Liliane Sprenger-Charolles and Pascale Colé 2004. MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers* 36 (1): 156–166.

Li, Hanhong and Alex C Fang 2011. Age tagging and word frequency for learners' dictionaries. *Corpus-based studies in language use, language learning, and language documentation*, 157–173. Brill Rodopi.

Liang, Percy, Ben Taskar and Dan Klein 2006. Alignment by Agreement. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 104–111. Association for Computational Linguistics (ACL).

Liao, Yan and Yoshinori J. Fukuya 2004. Avoidance of phrasal verbs: The case of Chinese learners of English. *Language learning* 54 (2): 193–226.

Lindström Tiedemann, Therese, Elena Volodina and Håkan Jansson 2016. Lärka: ett verktyg för träning av språkterminologi och grammatik. *LexicoNordica* 23: 161–181.

Llozhi, Lorena 2016. SweLL list. A list of productive vocabulary generated from second language learners' essays. Master's Thesis. University of Gothenburg.

Lönngren, Lennart 1988. *Svenskt associationslexikon.* Volume Rapport UCDL-R-88-2. Centrum för datorlingvistik. Uppsala universitet.

López-Jiménez, María Dolores 2013. Multi-word lexical units in L2 textbooks. *Revista española de lingüística aplicada*, no. 26:333–348.

LoPucki, Lynn M 2014. System and method for enhancing comprehension and readability of legal text. US Patent 8,794,972.

Louviere, Jordan J, Terry N Flynn and Anthony Alfred John Marley 2015. *Best-worst scaling: Theory, methods and applications.* Cambridge University Press.

Lyding, Verena, Lionel Nicolas, Branislav Bédi and Karën Fort 2018. Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing Techniques (enetCollect). *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL* 2018: 176–181.

Madnani, Nitin, Jill Burstein, Norbert Elliot, Beata Beigman Klebanov, Diane Napolitano, Slava Andreyev and Maxwell Schwartz 2018. Writing Mentor: Self-Regulated Writing Feedback for Struggling Writers. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 113–117.

Malmgren, Sven-Göran 2014. Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan. *LexicoNordica*, vol. 21.

Martinez, Ron and Norbert Schmitt 2012. A phrasal expressions list. *Applied linguistics* 33 (3): 299–320.

Mashhady, Habibollah, Behruz Lotfi and Mahbobeh Noura 2012. Word Type Effects on L2 Word Retrieval and Learning: Homonym versus Synonym Vocabulary Instruction. *Iranian Journal of Applied Language Studies* 3 (1): 97–118.

McCarthy, Diana, Sriram Venkatapathy and Aravind Joshi 2007. Detecting compositionality of verb-object combinations using selectional preferences. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).*

Mc Laughlin, G Harry 1969. SMOG grading-a new readability formula. *Journal of reading* 12 (8): 639–646.

Meara, Paul 1995. The importance of an early emphasis on L2 vocabulary. *The Language Teacher* 19: 8–11.

Meara, Paul 2002. The rediscovery of vocabulary. *Second Language Research* 18 (4): 393–407.

Merriam-Webster . word. `https://www.merriam-webster.com/dictionary/word`. Accessed: 2019-09-16.

Miestamo, Matti, Kaius Sinnemäki and Fred Karlsson 2008. *Language complexity: Typology, contact, change*. Volume 94. John Benjamins Publishing.

Mikolov, Tomas and Jeff Dean 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Miller, George and Christiane Fellbaum 1998. Wordnet: An electronic lexical database.

Miller, George A 1995. WordNet: a lexical database for English. *Communications of the ACM* 38 (11): 39–41.

Milton, James 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. Camilla Bardel, Lindqvist Christina and Batia Laufer (eds), *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*, 57–78. EuroSLA monograph series 2.

Moirón, María Begona Villada 2005. Data-driven identification of fixed expressions and their modifiability. Ph.D. diss., GRODIL, Secretary Department of General Linguistics.

Mühlenbock, Katarina Heimann and Sofie Johansson Kokkinakis 2012. SweVoc-a Swedish vocabulary resource for CALL. *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, 28–34. Linköping University Electronic Press.

Nation, Paul 2006. How large a vocabulary is needed for reading and listening? *Canadian modern language review* 63 (1): 59–82.

Nation, Paul 2013. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, Paul and Paul Meara 2010. Vocabulary. *An introduction to applied linguistics*, pp. 34–52.

Nation, Paul and Paul Meara 2013. Vocabulary. Norbert Schmitt (ed.), *An introduction to applied linguistics*, 44–62. Routledge.

Nicholls, Diane 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 conference*, Volume 16, 572–581.

Nicolas, Lionel, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous and Sangat 2020. Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning. *Proceedings of The 12th Language Resources and Evaluation Conference*, 268–278. Marseille, France: European Language Resources Association.

Nieto Piña, Luis 2019. Splitting rocks: Learning word sense representations from corpora and lexica. Ph.D. diss., University of Gothenburg.

Nunberg, Geoffrey, Ivan A Sag and Thomas Wasow 1994. Idioms. *Language* 70 (3): 491–538.

Nystrand, Martin 1979. Using readability research to investigate writing. *Research in the Teaching of English* 13 (3): 231–242.

Oakland, Thomas and Holly B Lane 2004. Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing* 4 (3): 239–252.

Och, Franz Josef and Hermann Ney 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29 (1): 19–51.

O'Dell, Felicity, John Read, Michael McCarthy et al. 2000. *Assessing vocabulary*. Cambridge university press.

Ogden, Charles Kay 1944. *Basic English: A general introduction with rules and grammar*. Volume 29. K. Paul, Trench, Trubner.

O'Muircheartaigh, Colm, George Gaskell and Daniel B Wright 1995. Weighing anchors: Verbal and numeric labels for response scales. *Journal of official statistics* 11: 295–308.

O'Regan, J Kevin and Arthur M Jacobs 1992. Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance* 18 (1): 185.

Ortega, Lourdes 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics* 24 (4): 492–518.

Ortega, Lourdes 2012. Interlanguage complexity. *Linguistic complexity: Second language acquisition, indigenization, contact.*, vol. 13.

Ortiz-Zambranoa, Jenny A and Arturo Montejo-Ráezb 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.

Östling, Robert and Jörg Tiedemann 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics* 106: 125–146.

Östling, Robert and Mats Wirén 2013. Compounding in a Swedish Blog Corpus.

Ozasa, Toshiaki, G Weir and Masayasu Fukui 2007. Measuring readability for Japanese learners of English. *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*, 122–125. Citeseer.

Paetzold, Gustavo and Lucia Specia 2016. SemEval 2016 Task 11: Complex Word Identification. *SemEval at NAACL-HLt*, 560–569.

Palmero Aprosio, Alessio, Stefano Menini and Sara Tonelli 2020. Adaptive Complex Word Identification through False Friend Detection. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, 192–200. New York, NY, USA: Association for Computing Machinery.

Paquot, Magali 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35 (1): 121–145.

Paquot, Magali and Sylviane Granger 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32: 130–149.

Paquot, Magali, Hubert Naets and Stefan Gries 2020. Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb+object structures in LONGDALE. *Second Language Acquisition and Learner Corpora.*

Paquot, Magali et al. 2007. Towards a productively-oriented academic word list.

Parent, Kevin 2009. Polysemy: A second language pedagogical concern.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga and Adam Lerer 2017. Automatic differentiation in PyTorch. *NIPS-w*.

Pearson 2017. GSE Teacher Toolkit. `https://www.english.com/gse/teacher-toolkit/user/lo`. Accessed: 2019-11-11.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (Oct): 2825–2830.

Perez, Naiara and Montse Cuadros 2017. Multilingual CALL Framework for Automatic Language Exercise Generation from Free Text. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 49–52.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Petersen, Sarah E and Mari Ostendorf 2007. Text simplification for language learners: a corpus analysis. *Workshop on Speech and Language Technology in Education*.

Petrov, Slav, Dipanjan Das and Ryan McDonald 2012. A Universal Part-of-Speech Tagset. Nicoletta Calzolari et al. (eds), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Petzell, Erik M 2017. Svenska Akademiens ordbok på nätet. *LexicoNordica*, vol. 24.

Pilán, Ildikó, Sowmya Vajjala and Elena Volodina 2015. A readable read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *Research in Computing Science*.

Pilán, Ildikó and Elena Volodina 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, 49–58.

Pilán, Ildikó, Elena Volodina and Lars Borin 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *TAL* 57 (3/2016): 67–91.

Pilán, Ildikó, Elena Volodina and Richard Johansson 2013. Automatic selection of suitable sentences for language learning exercises. *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*, 218–225. Research-publishing Dublin.

Pilán, Ildikó, Elena Volodina and Torsten Zesch 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2101–2111.

Pilán, Ildikó, David Alfter and Elena Volodina 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. *Proceedings of the workshop on Computational Linguistics for Linguistic Complexity (CL4LC).* COLING 2016. Osaka, Japan.

Pintard, Alice and Thomas François 2020. Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 85–92.

Ramisch, Carlos, Aline Villavicencio and Christian Boitet 2010. Mwetoolkit: a framework for multiword expression identification. *LREc*, Volume 10, 662–669. Valletta.

Rayner, Keith and Susan A Duffy 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition* 14 (3): 191–201.

Read, John 1988. Measuring the vocabulary knowledge of second langauge learners. *RELC journal* 19 (2): 12–25.

Read, John 2004. Plumbing the depths: How should the construct ofvocabulary knowledge be defined? *Vocabulary in a second language: Selection, acquisition, and testing* 10: 209.

Řehůřek, Radim and Petr Sojka 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. `http://is.muni.cz/publication/884893/en`.

Renner, Vincent and Jesús Fernández-Domínguez 2011. Coordinate compounding in English and Spanish. *Poznań Studies in Contemporary Linguistics PSiCL* 47: 873.

Reynolds, Robert 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 289–300.

Richards, Jack C 1974. Word lists: Problems and prospects. *RELC journal* 5 (2): 69–84.

Robinson, Peter 2001. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics* 22 (1): 27–57.

Roßdeutscher, Antje 2011. Particle Verbs and Prefix Verbs in German: Linking Theory versus Word-syntax. *Leuvense Bijdragen* 97: 1–53.

Rudzewitz, Björn, Ramon Ziai, Kordula De Kuthy and Detmar Meurers 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa, Gothenburg, 22nd May 2017*, 36–46. Linköping: Linköping University Electronic Press.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger 2002. Multiword Expressions: A Pain in the Neck for NLP. *International Conference on Intelligent Text Processing and Computational Linguistics*, 1–15. Springer.

Schmitt, Norbert 2000. *Vocabulary in language teaching*. Ernst Klett Sprachen.

Schmitt, Norbert 2010. *Researching vocabulary: A vocabulary research manual*. Springer.

Schneider, Gerold and Gaëtanelle Gilquin 2016. Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research* 2 (2): 177–204.

Schneider, Gerold and Johannes Graën 2018. NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners' Collocational Skills. *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*, 69–78. Linköping Electronic Conference Proceedings.

Severyn, Aliaksei and Alessandro Moschitti 2015. Twitter sentiment analysis with deep convolutional neural networks. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 959–962.

Shardlow, Matthew 2013. A Comparison of Techniques to Automatically Identify Complex Words. *ACL (Student Research Workshop)*, 103–109.

Shardlow, Matthew 2014. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. *LREc*, 1583–1590.

Singleton, David 1995. Introduction: A critical look at the critical period hypothesis in second language acquisition research. *The age factor in second language acquisition*, pp. 1–29.

Sjöholm, Johan 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linkö.

Skehan, Peter 2001. Tasks and language performance. *Researching pedagogic tasks: Second language learning, teaching, and testing*, pp. 167–185.

Skehan, Peter and Pauline Foster 1999. The influence of task structure and processing conditions on narrative retellings. *Language learning* 49 (1): 93–120.

Sköldberg, Emma and Sofie Johansson Kokkinakis 2012. A och O om akademiska ord - Om framtagning av en svensk akademisk ordlista. *Nordiske Studier i Leksikografi*, no. 11.

Sköldberg, Emma, Louise Holmer, Elena Volodina and Ildikó Pilán 2019. State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0: empirical, applied and interdisciplinary research* 7 (1): 13–24.

Smith, Edgar A 1961. Devereux Readability Index. *The Journal of Educational Research* 54 (8): 298–303.

Smith, Nathaniel J. and Roger Levy 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128 (3): 302 – 319.

Smith, Samuel L, David HP Turban, Steven Hamblin and Nils Y Hammerla 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859.*

Snow, Rion, Brendan O'connor, Dan Jurafsky and Andrew Y Ng 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Solemon, Badariah, Izyana Ariffin, Marina Md Din, Rina Md Anwar et al. 2013. A review of the uses of crowdsourcing in higher education. *International Journal of Asian Social Science* 3 (9): 2066–2073.

Specia, Lucia, Sujay Kumar Jauhar and Rada Mihalcea 2012. SemEval-2012 Task 1: English Lexical Simplification. *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, 347–355. Stroudsburg, PA, USA: Association for Computational Linguistics.

Spinner, Patti and Susan M Gass 2019. *Using judgments in second language acquisition research.* Routledge.

Squillante, Luigi 2014. Towards an empirical subcategorization of multiword expressions. *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, 77–81.

Stegbauer, Christian, Elisabeth Bauer, Elisabeth Kartashova and Alexander Rausch 2009. *Wikipedia.* Springer.

Stemle, Egon W., Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén and Elena Volodina 2019. Working together towards an ideal infrastructure for language learner corpora. Andrea Abel, Aivars Glaznieks, Verena Lyding and Lionel Nicolas (eds), *Widening the scope of learner corpus research*, Corpora and Language in Use, 427–468. France: Presses universitaires de Louvain.

Sweedler-Brown, Carol O 1985. The influence of training and experience on holistic essay evaluations. *The English Journal* 74 (5): 49–55.

Swinney, David A and Anne Cutler 1979. The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior* 18 (5): 523–534.

Tack, Anaïs, Thomas François, Piet Desmet and Cédrick Fairon 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 137–146.

Tack, Anaïs, Thomas François, Anne-Laure Ligozat and Cédrick Fairon 2016a. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. *La 23ème Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2016).*

Tack, Anaïs, Thomas François, Anne-Laure Ligozat and Cédrick Fairon 2016b. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. *LREc*.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov and Afsaneh Fazly 2017. Investigating the opacity of verb-noun multiword expression usages in context. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 133–138.

Teleman, Ulf, Staffan Hellberg and Erik Andersson 1999. *Svenska akademiens grammatik*. Svenska akademien.

Tenfjord, Kari, Paul Meurer and Knut Hofland 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 1821–1824.

The BNC Consortium 2007. The British National Corpus, version 3 (BNC XML Edition). `http://www.natcorp.ox.ac.uk/`. Distributed by Bodleian Libraries, University of Oxford.

Toivonen, Ida 2002. Verbal particles and results in Swedish and English. *Pro-*

ceedings of the West Coast Conference in Formal Linguistics, Volume 21, 457–470.

Troncoso Skidmore, Susan and Bruce Thompson 2010. Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement* 70 (5): 777–795.

Uitdenbogerd, A 2005. Readability of French as a foreign language and its uses. *ADCS 2005: The Tenth Australasian Document Computing Symposium*, 19–25. University of Sydney.

Vajjala, Sowmya and Detmar Meurers 2012. On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the seventh workshop on building educational applications using NLp*, 163–173. Association for Computational Linguistics.

Vajjala, Sowmya and Detmar Meurers 2014. Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 288–297.

Van de Cruys, Tim and Begona Villada Moirón 2007. Semantics-based multiword expression extraction. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 25–32. Association for Computational Linguistics.

Vapnik, Vladimir and Vlamimir Vapnik 1998. Statistical learning theory. *Wiley* 1: 624.

Venkatapathy, Sriram and Aravind K Joshi 2005. Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 899–906. Association for Computational Linguistics.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart and Carlos Ramisch 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1034–1043.

Volodina, Elena 2008. From Corpus to Language Classroom: reusing Stockholm Umeå Corpus in a vocabulary exercise generator SCORVEX. Master's thesis, University of Gothenburg.

Volodina, Elena, Lars Borin, Hrafn Lofsson, Birna Arnbjörnsdóttir and Guðmundur Örn Leifsson 2012. Waste not; want not: Towards a system architecture for ICALL based on NLP component re-use. *Proceedings of* *the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, 47–58. Linköping University Electronic Press.

Volodina, Elena, Lena Granstedt, Sofia Johansson, Beáta Megyesi, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg and Mats Wirén 2018. Annotation of learner corpora: first SweLL insights. *Proceedings of Swedish Language Technology Conference (SLTC).*

Volodina, Elena, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg and Mats Wirén 2019. The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology (NEJLT)* 6 (4): 67–104.

Volodina, Elena and Sofie Johansson Kokkinakis 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 1040–1046.

Volodina, Elena and Dijana Pijetlovic 2015. Lark Trills for Language Drills: Text-to-speech technology for language learners. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 107–117.

Volodina, Elena, Ildikó Pilán and David Alfter 2016. Classification of Swedish learner essays by CEFR levels. *CALL communities and culture–short papers from EUROCALL* 2016: 456–461.

Volodina, Elena, Ildikó Pilán, Stian Rødven Eide and Hannes Heidarsson 2014a. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*. Linköping University Electronic Press.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg and Monica Sandell 2016a. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 206–212.

Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse and Thomas François 2016b. SweLLex: second language learners' productive vocabulary. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 76–84. Linköping University Electronic Press.

Volodina, Elena, Ildikó Pilán, Lars Borin and Therese Tiedemann Lindström 2014b. A flexible language learning platform based on language resources and web services. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland*, 3973–3978.

Weigle, Sara Cushing 1998. Using facets to model rater training effects. *Language testing* 15 (2): 263–287.

West, Michael Philip 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green.

Wilkins, David Arthur 1972. *Linguistics in language teaching*. E. Arnold, 1973.

Wilson, Michael 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers* 20 (1): 6–10.

Witten, Ian H, Eibe Frank, Mark A Hall and Christopher J Pal 2011. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wray, Alison and Michael R Perkins 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20 (1): 1–28.

Xia, Menglin, Ekaterina Kochmar and Ted Briscoe 2016. Text readability assessment for second language learners. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 12–22.

Yancey, Kevin and Yves Lepage 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Yaneva, Victoria, Irina P Temnikova and Ruslan Mitkov 2016. Evaluating the Readability of Text Simplification Output for Readers with Cognitive Disabilities. *LREc*.

Yates, Frank 1936. Incomplete randomized blocks. *Annals of Eugenics* 7 (2): 121–140.

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack and Marcos Zampieri 2018. A Report on the Complex Word Identification Shared Task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78. New Orleans, United States: Association for Computational Linguistics.

Zedler, Johann Heinrich 1749. Grosses vollstandiges Universal-Lexikon, vol. 62. *Leipzig & Halle*, p. 488.

Zeller, Jochen 2001. *Particle verbs and local domains*. Volume 41. John Benjamins Publishing.

Zhang, Yi, Valia Kordoni, Aline Villavicencio and Marco Idiart 2006. Automated multiword expression prediction for grammar engineering. *Proceedings of the workshop on multiword expressions: Identifying and exploiting underlying properties*, 36–44. Association for Computational Linguistics.

Zotova, Elena, Montse Cuadros, Naiara Perez and Aitor García-Pablos 2020. Vicomtech at ALexS 2020: Unsupervised Complex Word Identification Based on Domain Frequency. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.

A List of other publications not included in the thesis

A.1 Abstracts

- Alfter, David 2016. Learning the Learner: User Modeling in Intelligent Computer Assisted Language Learning Systems. CEUR Workshop Proceedings, v.1618. UMAP 2016 Extended Proceedings. Halifax, Canada, July 13-16, 2016. Edited by : Federica Cena, Michel Desmarais, Darina Dicheva, Jie Zhang.

- Alfter, David and Elena Volodina 2016. Modeling Individual Learner Knowledge in a Computer Assisted Language Learning System. Proceedings of the Sixth Swedish Language Technology Conference. Umeå University, 17-18 November, 2016.

- Alfter, David and Elena Volodina 2017. Adaptive diagnostic test. EuroCALL 2017 - CALL in a climate of change, Book of Abstracts, Southampton, United Kingdom, 23-26 August.

- Alfter, David 2017. Evaluating sentence rearrangment and sentence composition tasks using Partial Tree Kernels. EuroCALL 2017 - CALL in a climate of change, Book of Abstracts, Southampton, United Kingdom, 23-26 August.

- Pilán, Ildikó and **David Alfter** and Elena Volodina 2017. Lärka: an online platform where language learning meets natural language processing. 7th ISCA Workshop on Speech and Language Technology in Education, 25-26 August 2017, Stockholm, Sweden.

- Alfter, David and Lars Borin and Ildikó Pilán and Therese Lindström Tiedemann and Elena Volodina 2018. From Language Learning Platform to Infrastructure for Research on Language Learning. Proceedings of CLARIN-2018 conference, Pisa, Italy.

- Alfter, David and Elena Volodina 2018. Is the whole greater than the sum of its parts? A corpus-based pilot study of the lexical complexity in multi-word expressions. Proceedings of SLTC 2018, Stockholm, October 7-9, 2018.

- Alfter, David and Elena Volodina 2019. From river to bank: The importance of sense-based graded word lists. EUROCALL 2019 - CALL and Complexity, Book of Abstracts, Louvain-la-Neuve, Belgium, 28-31 August, 2019.

- Alfter, David and Therese Lindström Tiedemann and Elena Volodina 2020. Expert judgments versus crowdsourcing in ordering multi-word expressions. Prooceedings of the Eighth Swedish Language Technology Conference. Gothenburg, Sweden and Online, 26-27 November, 2020.

A.2 Book reviews

- Alfter, David and Anders Agebjörn 2016. Review of developing, modelling and assessing second languages. Linguistlist.

- Agebjörn, Anders and **David Alfter** 2018. Review of advanced proficiency and exceptional ability in second languages. Linguist List, no. Jan 16.

A.3 Conference articles (peer-reviewed)

- Volodina, Elena and Ildikó Pilán and **David Alfter** 2016. Classification of Swedish learner essays by CEFR levels. Proceedings of EuroCALL 2016. 24-27th August 2016, Cyprus.

- Alfter, David and Yuri Bizzoni 2016. Hybrid Language Segmentation for Historical Documents. Proceedings CLiC-it 2016 and EVALITA 2016, Napoli, Italy, December 5-7, 2016. Edited by : Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, Rachele Sprugnoli.

- Pilán, Ildikó and **David Alfter** and Elena Volodina 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC).

A.4 Proceedings

- Pilán, Ildikó and Elena Volodina and **David Alfter** and Lars Borin 2018. Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018), November 7 2018, Stockholm, Sweden. Linköping University Electronic Press.

- Alfter, David and Elena Volodina and Lars Borin and Ildikó Pilán and Herbert Lange. Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku, Finland. Linköping University Electronic Press.

- Alfter, David and Elena Volodina and Lars Borin and Ildikó Pilán and Herbert Lange. Proceedings of the 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020), November 25 2020, Gothenburg, Sweden. Linköping University Electronic Press.

B List of abbreviations

AWL	Academic Word List
BERT	Bidirectional Encoder Representations from Transformers
CAF	Complexity, Accuracy and Fluency
CALL	Computer Assisted Language Learning
CEFR	Common European Framework of Reference
CNN	Convolutional Neural Network
ELMo	Embeddings from Language Models
ET	ExtraTrees: Extremely Randomized Trees
EVP	English Vocabulary Profile
FVP	French Vocabulary Profile
FO	First occurrence
GSE	Global Scale of English
GSL	General Service List
ICALL	Intelligent Computer Assisted Language Learning
ILR	International Language Roundtable
KELLY	KEywords for Language Learning for Young and adults alike
L1	First language

L2	Second language
LIX	Läsbarhetsindex ‘Readability index’
MLP	MultiLayer Perceptron
MLU	Multi-Lexeme Unit
MRC	Medical Research Council
MWE	Multi-Word Expression
MWU	Multi-Word Unit
NGSL	New General Service List
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NS	Native Speaker
OCR	Optical Character Recognition
PCA	Principal Component Analysis
PoS	Part-of-Speech
RCNN	Recurrent Convolutional Neural Network
RF	Random Forest
SAL	Swedish Associative Lexion
SALDO	Swedish Associative Lexicon 2
SiWoCo	Single Word Complexity
SLA	Second Language Acquisition
SMOG	Simple Measure of Gobbledygook
SOoU	Significant Onset of Use
SVM	Support Vector Machine
TTS	Text-to-speech
W1L	Within one level
WEKA	Waikato Environment for Knowledge Analysis
XML	Extensible Markup Language

C List of resources

COCTAILL
The COCTAILL corpus consists of 18 textbooks aimed at learners of Swedish in Sweden and covering different proficiency levels from beginner levels (A1) to advanced levels (C1).

SweLL-pilot
The SweLL-pilot corpus consists of 339 learner essays covering the CEFR levels A1 to C1.

CEFRLex
A family of resources derived from graded textbooks. Each resource lists the occurring lemmata and the corresponding frequencies *per level*.

SVALex
The Swedish receptive CEFRLex list based on COCTAILL.

SweLLex
The Swedish productive CEFRLex list based on SweLL-pilot.

FLELex
The French CEFRLex list.

EFLLex
The English CEFRLex list.

NT2Lex
The Dutch CEFRLex list.

SAL
The Swedish Associative Lexicon.

SALDO
The Swedish Associative Lexicon version 2. It builds on the Swedish Associative Lexicon SAL.