Ildikó Pilán

Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning

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

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

29 • 2018

Ildikó Pilán

Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning

Gothenburg 2018

Data linguistica 29
ISBN 978-91-87850-68-4
ISSN 0347-948X
E-publication <http://hdl.handle.net/2077/55895>

Printed in Sweden by
Repro Lorensberg 2018

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

Cover design by Sven Lindström

Front cover illustration by Csaba Sajti ©

Author photo on back cover by Antonio Martini

ABSTRACT

With the ever-growing presence of electronic devices in our everyday lives, it is compelling to investigate how technology can contribute to make our language learning process more efficient and enjoyable. A fundamental piece in this puzzle is the ability to measure the complexity of the language that learners are able to deal with and produce at different stages of their progress.

In this thesis work, we explore automatic approaches for modeling linguistic complexity at different levels of learning Swedish as a second and foreign language (L2). For these purposes, we employ natural language processing techniques to extract linguistic features and combine them with machine learning methods. We study linguistic complexity in two types of L2 texts: those written by experts for learners and those produced by learners themselves. Moreover, we investigate this type of data-driven analysis for the smaller unit of sentences.

Automatic proficiency level prediction has a number of application potentials for the field of Intelligent Computer-Assisted Language Learning, out of which we investigate two directions. Firstly, this can facilitate locating learning materials suitable for L2 learners from corpora, which are valuable and easily accessible examples of authentic language use. We propose a framework for selecting sentences suitable as exercise items which, besides linguistic complexity, encompasses a number of additional criteria such as well-formedness and independence from a larger textual context. An empirical evaluation of the system implemented using these criteria indicated its usefulness in an L2 instructional setting. Secondly, linguistic complexity analysis enables the automatic evaluation of L2 texts which, besides being helpful for preparing learning materials, can also be employed for assessing learners' writing. We show that models trained partly or entirely on reading texts can effectively predict the proficiency level of learner essays, especially if some learner errors are automatically corrected in a pre-processing step. Both the sentence selection and the L2 text evaluation systems have been made freely available on an online learning platform.

SAMMANFATTNING

Med allt fler intelligenta apparater och mobil teknologi i vår vardag blir det angeläget att undersöka hur tekniken kan bidra till att göra språkinlärningsprocessen effektivare och mer tilltalande. En grundläggande del i detta är förmågan att mäta den språkliga komplexiteten som elever kan hantera och producera på olika nivåer under deras utveckling.

I denna doktorsavhandling undersöker vi automatiska metoder för att modellera språklig komplexitet på olika inlärningsnivåer för svenska som andra och främmande språk (L2). Vi använder metoder från språkvetenskaplig databehandling för att extrahera olika språkliga särdrag och kombinerar dem med maskininlärningsmetoder. Vi studerar språklig komplexitet i två typer av L2-texter: sådana som experter (lärare) skriver för elever och sådana som produceras av eleverna själva. Vi utforskar dessutom denna typ av automatisk analys även för enstaka meningar.

Att automatiskt kunna bedöma färdighetsnivåer möjliggör ett antal intressanta tillämpningar för datorstödd språkinlärning, där vi har utforskat två spår. Å ena sidan kan detta underlätta framtagningen av korpusexempel som är värdefulla exempel på autentiskt språkbruk för L2-elever. Vi föreslår ett ramverk för att hitta korpusmeningar som kan återanvändas i övningar. Detta, förutom språklig komplexitet, omfattar ett antal ytterligare kriterier, såsom hur välformad meningen är och egenskapen av att vara oberoende av andra meningar från den ursprungliga kontexten. En empirisk utvärdering av meningsurvalsystemet som implementerades med dessa kriterier visade dess nytta för L2 inlärning. Å andra sidan, språklig komplexitetsanalys möjliggör också den automatiska utvärderingen av L2-texter som kan stödja förberedningen av L2 läromedel. Analysen kan också användas för att utvärdera elevers skriftliga produktion. Vi visar att maskininlärningsmodeller som helt eller delvis tränas på lästexter kan effektivt klassificera färdighetsnivån på elevuppsatser, speciellt om vissa L2-fel korrigeras automatiskt i ett förbehandlingssteg. Slutligen visar vi hur forskningsresultaten har integrerats i en fritt tillgänglig online-lärplattform.

ACKNOWLEDGEMENTS

There are many whom I would like to thank for having contributed to this work, in one way or another. They are more than I can individually list here, but I do hope that even those not explicitly named will find one or more categories in which they can feel included.

First and foremost, I would like to thank my supervisors. I am grateful to Lars Borin, for his valuable insights and guidance throughout these years. His keen interest in Hungarian language and culture has always made me feel a step closer to my country of birth. I am deeply thankful to my other supervisor, Elena Volodina, for introducing me to the field of Intelligent Computer-Assisted Language Learning. Her expertise, spirit of initiative, excellent networking skills and friendship have been invaluable.

I would also like to thank my examiner for the final seminar, Mats Wirén, for thoroughly revising an earlier version of this thesis and for providing insightful comments. I appreciate also the feedback received from the reviewers of my articles submitted during these years and the helpful comments of those who proof-read parts of this thesis, namely Jenny Kierkemann and Jenny Mattsson. I would also like to thank my cousin, Csaba Sajti, for designing the front cover illustration for this book and Sven Lindström, for his patience with small adjustments while preparing the cover design.

Special thanks to those teachers and students from the Centrum för Språkintroduktion who have taken the time and effort to participate in our user evaluation.

Moreover, I would like to express my gratitude to the generous providers of the funding that allowed me to attend numerous academic events: the Center for Language Technology, Språkbanken, Kungl. Vitterhetsakademien, Filosofiska fakulteternas gemensamma donationsnämnd, Adlerbert Scholarships, the INDUS network and the Horizon 2020 Framework Programme of the European Union which made my participation in the ENeL and the enetCollect COST actions possible. I have greatly benefitted from the attended events by being in contact with excellent researchers and by being able to follow the latest developments of my field.

I would also like to thank my colleagues from Språkbanken, the Department of Swedish, the whole NLP community in Gothenburg, in Sweden and from around the world for attending my presentations, providing useful comments and asking thought-provoking questions. I dearly treasure also the memory of our wonderful social events and I am thankful for being able to discover that, besides scientific curiosity, plenty of human warmth ties our research communities together. I am especially grateful to my co-authors for the opportunity to learn from them through inspiring brain-storming and experimenting sessions, not to mention our productive ice cream meetings.

I am thankful to those I have shared an office with throughout these years. They have created, together with many others in Språkbanken, a serene and welcoming environment that I will always fondly remember. I value also the freedom I received in Språkbanken for choosing the type of research to conduct. Being able to combine my two different backgrounds – language teaching and language technology – while maintaining social relevance in my research has been a continuous source of motivation.

I am grateful to my friends, near and far, work and non-work related – or between the two – for countless beautiful moments filled with laughter, inspiring discussions, boardgames, beach volleyball, traveling, good food and plenty more. I thank them also for being close in difficult times. I am thankful to my "adoptive families" in Sweden for sharing their home with me, providing me a safe place in a new land and a magic key to enter Swedish and, in general, Scandinavian culture.

I would like to deeply thank my family and relatives in Hungary, in particular my mother, Ildikó and my brother, Dani, as well as my acquired family in Italy, for their support and warm encouragement. They have always welcomed me with open arms and helped me recharge with energy during my time off. I am also thankful to those family members who sadly got to witness only part of this journey, my father and my grandparents on my mother's side.

Last, but by no means least, I am immensely grateful to Antonio, my life companion, best friend and occasional "supervisor" for being there and believing in me. I thank him also for helping me discover the incredible potential of being outside of one's comfort zone, which led me to embark, among others, on this journey.

Ildikó Pilán
Gothenburg, April 19, 2018

CONTENTS

Appendices

Part I

Introduction and overview of the thesis work

1 Introduction

Due to the rapid growth of international mobility for work, leisure or necessity in the past decades, the number of language learners world-wide has been steadily increasing (Castles, De Haas and Miller 2013). Effective communication skills in the language of the host country are a key for successful societal integration and they are crucial for accessing also the job market.

At the same time, numerous aspects of our everyday life are being enhanced by technology and the language learning domain is no exception. Early Computer-Assisted Language Learning (CALL) systems developed up to the 1990s were, however, often limited to offering manually created content in a digital format (Borin 2002a). Natural language processing[1] (NLP) techniques that enable a deep automatic analysis of written and spoken language have seen an unprecedented advance since those early systems. This gave rise to the combination of NLP and CALL in the 1990s, which became known as Intelligent CALL (ICALL).

ICALL has promising potentials for enhancing language teaching and learning practices in a variety of ways, such as predicting automatically at what language learning stage learners would be able to read or produce a certain text. While beginner learners typically know only a limited amount of words and simple structures to connect them, when they progress and become more proficient, they learn to master more complex and varied linguistic elements.

The present thesis focuses on the automatic analysis of linguistic complexity and explores how this analysis can be employed for the identification of suitable language learning materials and for the automatic evaluation of learner production. In this work, we operationalize the term *linguistic complexity* as the set of lexico-semantic, morphological and syntactic characteristics reflected in texts (or sentences) that determine the magnitude of the language skills and competences required to process or produce them. We use linguistic complexity analysis as a means of determining second and foreign language (L2) learning

[1] Alternative, but not entirely equivalent terms for this discipline are Computational Linguistics and Language Technology.

levels.[2] The scale of learning (*proficiency*) levels adopted in this work is the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). The CEFR offers a common ground for language learning and assessment and it proposes a six-point scale of proficiency levels (for a more detailed account of CEFR, see section 2.1).

In related literature, readability analysis, introduced in section 2.3.2, is often used as a synonym to proficiency level classification, especially in the case of data-driven approaches for the assessment of reading materials. A number of terms have been used in parallel in this context including *readability* (Branco et al. 2014; François and Fairon 2012), *difficulty* (Huang et al. 2011; Salesky and Shen 2014), *linguistic complexity* (Ströbel et al. 2016) and *CEFR level prediction* (Hancke 2013; Vajjala and Lõo 2014). This holds not only for previous work in the literature, but also for the publications included in this thesis. Parts II and III show, in fact, some variation in the use of this terminology. This is, in part, due to an evolving understanding of the phenomenon under investigation and, in part, a wish to establish a link with previous research as well as to adjust to different target audiences. Linguistic complexity analysis can be used for predicting both readability levels and proficiency (CEFR) levels. Although both readability and scales of proficiency levels also include a number of additional aspects, some criteria connected to linguistic complexity heavily underlies both and it is the one aspect that most NLP systems providing such analyses explicitly or implicitly capture.

Linguistic complexity has been explored across two different dimensions in this thesis: (i) the size of the linguistic context investigated and (ii) the type of learner skills involved when dealing with the texts. In the former case, we carried out experiments both at the text and at the sentence level. Regarding skill types, we distinguished between *receptive* skills, required when learners process passages produced by others, and *productive* skills, when learners produce the texts themselves.

The choice of focusing on automatic linguistic complexity analysis is motivated by a number of reasons. Firstly, it can constitute a valuable aid for teachers to carry out their tasks more efficiently and it can also become a powerful tool for self-directed learning. This type of analysis allows for the identification of additional reading material, the creation of automatic exercises and it facilitates the provision of feedback to learners. A sufficient amount of practice and repetition plays, in fact, a crucial role in L2 learning (DeKeyser 2007), not only when familiarizing with new vocabulary and grammar, but also for

[2] In this thesis, we will use the terms *second* and *foreign* language interchangeably since we we do not distinguish between these in our linguistic complexity analysis. The same applies to the terms *learning* and *acquisition*.

effectively remembering them (Settles and Meeder 2016). Digital collections of texts, i.e. *corpora*, are a rich source of diverse authentic examples whose positive effect on learners' progress has been shown, among others, in Cobb (1997) and Cresswell (2007).

1.1 Research questions and contributions

In this section we summarize the main research questions and contributions from parts II – IV related to learning material selection, learner text evaluation and feature selection across different L2 text types.

1.1.1 Learning material selection

One of the starting points of the automatic identification of L2 learning materials is the ability to assess whether the complexity of a linguistic unit (text or sentence) is appropriate for learners at different levels. A number of research questions arise in connection to this, which are investigated in chapter 7 and which include:

- How successfully can we automatically predict CEFR levels in Swedish using linguistic complexity features and machine learning techniques?
- Are traditional readability formulas useful for this task?
- Does the size of the linguistic input (text vs. sentences) influence performance?

One of the main contributions of this thesis in connection to these research questions is a supervised machine learning model for the automatic classification of proficiency levels in different types of L2 texts and sentences using linguistic complexity features. These models achieve a performance that compares well both to previously published results for other languages and to human annotators solving the same task. Two particular aspects of the models proposed are the use of: (i) weakly lexicalized features where word forms are represented by their CEFR level instead of their base form, and (ii) the inclusion of L2-relevant morphological features.

Being able to analyze linguistic complexity at the sentence level is useful, for instance, for the automatic generation of exercises. It enables the automatic identification of suitable sentences from various (even non L2-related) corpora. Besides linguistic complexity, however, a number of other factors need to be considered when selecting sentences from corpora for L2 exercises. A second set of research questions raised in chapters 8 and 9 are:

- What criteria should corpus example sentences satisfy to be useful for the generation of language learning exercises?
- How can we capture these criteria using NLP tools?
- How can we automatically select corpus examples that are independent from their textual context?
- How well does an automatic corpus sentence selection system perform in an educational setting?

Based on previous research and a qualitative analysis of empirical evidence from previous user evaluations, we propose a framework for the selection of sentences from corpora for L2 exercises. The framework aims at being generic enough to be useful for different types of L2 exercises and specific enough to satisfy certain needs relevant for the L2 context (e.g. CEFR level prediction). We implemented a hybrid system combining rule-based and machine learning techniques for selecting exercise item candidates based on the framework proposed. The rule-based nature of the system does not only offer direct user control over different linguistic characteristics of sentences, but it also allows for providing explicit and detailed information on the characteristics and quality of the sentences. To answer the fourth research question above, the framework and its implementation were evaluated with the help of a user study with L2 teachers and learners of Swedish. This indicated a promising practical applicability of our system in L2 teaching and learning.

1.1.2 Learner text evaluation

Since the lack of a sufficient amount of annotated data is a recurrent problem for different NLP tasks, we investigated also the potentials of transfer learning for automatic CEFR level prediction. The third set of research questions explored in this thesis connected to this topic include:

- How can we exploit coursebook texts to improve CEFR level classification for learner essays?
- Can a CEFR classification model be transfered across texts involving different L2 skills? More concretely, how well does a model predicting CEFR levels for reading comprehension texts perform when used to classify learner-written essays?

- Does correcting errors in the learner essays improve the usefulness of coursebook texts for the essay classification?

We show that reading texts can improve the classification of CEFR levels in learner essays either as an alternative source of data if errors are normalized in learner essays (chapter 10), or as the basis of lexical features (chapter 11).

1.1.3 Investigating feature importances

Identifying the optimal number and types of features to use in a machine learning task can boost performance and decrease computation time. This is especially important when models are planned to be integrated into NLP applications aiming at on-the-fly predictions. In chapter 12, we focus on this matter and report the results of feature selection experiments performed on three different datasets: one consisting of reading comprehension text from coursebook, one of learner written essays and a dataset of corpus example sentences with teacher-evaluated CEFR levels. These experiments aim at answering the following research questions:

- Which linguistic complexity features are most useful for determining proficiency levels in different L2 datasets?
- Are there features that are generally predictive regardless of input size and the type of skill considered?

We present, on the one hand, a subset of the most informative features for each of the three datasets and show that including only these features leads to an improved classification performance compared to using all of them. On the other hand, we identify some lexical, morphological and syntactic features that are good indicators of complexity across all three datasets.

1.1.4 Contributions related to web development and resource creation

The research carried out within this thesis work has been incorporated into a freely available online platform, *Lärka* (Volodina et al. 2014a), with the purpose of making it available to the general public. Both the sentence selection and the text evaluation systems are accessible through a graphical user interface and their functionalities can be re-used by other developers via the web services provided. The functionalities available in the two systems constitute part of the engineering contributions of this thesis. The graphical user interface has been implemented by others within the SweLL infrastructure project (Volodina et al. 2016a).[3]

Finally, additional contributions consisted of various forms of collaborations for the creation of L2 Swedish language resources, which were at the basis of the experiments presented and can be reused by other studies on L2 complexity in the future. This work included, on the one hand, participating in the preparation of a coursebook corpus described in section 3.1.3, as well as measuring inter-annotator agreement and compiling exploratory statistics about both this corpus and a learner essay corpus (section 3.1.4). A small dataset of sentences annotated with CEFR levels collected during a user evaluation (section 9.4) is also being made available. On the other hand, for the L2 word lists introduced in section 3.2.2, a number of post-processing steps were performed such as mapping semi-automatically to base forms some entries which were not lemmatized automatically.

1.2 Overview of publications

The following publications are included in this thesis:

1. Pilán, Ildikó, Sowmya Vajjala and Elena Volodina 2016. A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications (IJLCA) 7 (1): 143–159.* [Chapter 7]

2. Pilán, Ildikó 2016. Detecting Context Dependence in Exercise Item Candidates Selected from Corpora. In *Proceedings of the* 11^{th} *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 151–161. [Chapter 8]

3. Pilán, Ildikó, Elena Volodina and Lars Borin 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues (TAL) Journal, Special issue on NLP for learning and teaching* 57 (3): 67–91. [Chapter 9]

4. 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 the* 26^{th} *International Conference on Computational Linguistics (COLING)*, 2101–2111. [Chapter 10]

[3] `https://spraakbanken.gu.se/eng/swell_infra`

5. 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)*, 120–126. [Chapter 11]

6. Pilán, Ildikó and Elena Volodina. Investigating the importance of linguistic complexity features across different datasets related to language learning. Submitted. [Chapter 12]

In the case of the publications listed above, the first author is the main contributor in terms of both the ideas and the implementation of the presented research. This includes the design and the execution of the experiments described, data pre-processing and error analysis. Two exceptions to this are: (i) in Pilán, Vajjala and Volodina (2016) the experiment applying the sentence-level model to texts has been carried out by the other co-authors; (ii) in Pilán, Alfter and Volodina (2016), the other co-authors performed the mapping of frequency distributions to CEFR levels in the word list employed in the experiments. The text of these publications have been reformatted to ensure more homogeneity in the form of their presentation here. A short summary of the above publications constituting chapters 7 – 12 is provided in section 1.3.

A number of additional articles have been published during the same period which were not included in the thesis, but are listed in appendix A. Most of these articles describe collaborations for the creation of L2 Swedish language resources underlying the experiments. These include a coursebook corpus (Volodina et al. 2014b), a learner essay corpus (Volodina et al. 2016c) and frequency word lists based on these (François et al. 2016; Volodina et al. 2016b) which are described in chapter 3.

1.3 Structure of the thesis

This thesis is structured as follows. *Part I* presents an overview of the work carried out in the publications included in parts II – IV and summarizes their contributions. In this first chapter, we introduced the context of this work, motivated it and clarified some of the terminology used.

Chapter 2 provides an overview of the related literature. We first briefly introduce the L2 learning context and the CEFR in section 2.1. In section 2.2, we summarize previous work in ICALL. Section 2.3 is dedicated to different lines of work connected to linguistic complexity, namely readability, proficiency level prediction for receptive texts and learner text evaluation. Studies related to the selection of corpus example sentences to be used either as dictionary examples or as exercise items are outlined in section 2.4.

Chapter 3 describes the resources employed in our experiments. These include two L2 corpora, one consisting of reading texts and another composed of learner essays as well as three lexical resources containing information about word frequency and suggested CEFR levels.

Chapter 4 presents the core methods used in the included papers. We introduce a number of machine learning algorithms and measures to evaluate their performance as well as some domain adaptation methods.

Chapter 5 provides an overview of our research on linguistic complexity both for receptive and for productive texts. We introduce and motivate the feature set used and summarize our main results. Moreover, we investigate feature importances across different datasets. This is followed by a description of how research outcomes have been integrated into an online ICALL platform. We conclude this chapter with a discussions around the limitations of our studies.

Chapter 6 concludes part I and outlines future work.

Parts II and III include a number of selected peer-reviewed publications centered around the topic of linguistic complexity. *Part II* presents studies about receptive linguistic complexity for the identification of language learning material candidates.

In *chapter 7*, linguistic complexity for both sentences and texts is explored. We find that a traditional, count-based readability formula does not adequately reflect differences in complexity at various CEFR levels. We show how the same feature set capturing both lexical and grammatical aspects can classify the two types of data more reliably. We investigate also how homogeneous texts are in terms of the CEFR level of the sentences contained.

Chapter 8 investigates linguistic factors rendering sentences dependent on their larger textual context that includes both structural and lexical aspects such as referential expressions. An implementation of these aspects is also described and evaluated on different datasets.

Chapter 9 presents a framework and its implementation for selecting exercise item candidates from generic (not learner-specific) corpora. We describe a hybrid system based on both heuristics and machine learning that ensures a highly customizable sentence selection. The results of an empirical evaluation with language teachers and learners are also reported.

Part III includes two publications about assessing linguistic complexity in learner-written texts. Both chapters investigate how reading comprehension texts can be successfully exploited to overcome the problem of insufficient amount of learner-written texts when classifying proficiency levels.

Chapter 10 presents a number of attempts at how reading texts and learner essays can be combined for a more efficient classification of CEFR levels in the latter. We show that correcting learner errors improves classification performance considerably when only using information from reading texts.

In *chapter 11*, we investigate an alternative to how information from reading texts can be used for learner essay classification: using them for informing lexical features. We compare using a frequency word list based on web texts to a list based on L2 reading text frequencies and find that the latter boosts CEFR level classification accuracy.

Finally, in *part IV*, *chapter 12*, we conclude our investigations around linguistic complexity in the L2 context by reporting the results of feature selection experiments. We identify a subset of features which are informative (individually or shared) for three different datasets including reading comprehension text from coursebook, learner-written essays and a small dataset of corpus example sentences with teacher-evaluated CEFR levels.

2 Background

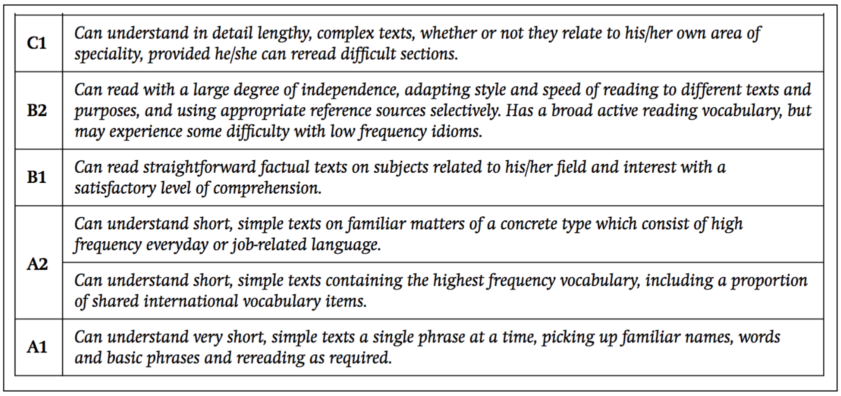
2.1 The second language learning context and the CEFR

First and second language acquisition present a number of differences, among others, in terms of learners' background knowledge and age (Beinborn, Zesch and Gurevych 2012). L2 learners already master at least one other language and they are often older compared to those acquiring their first language (L1). The mode of acquisition can also differ since in the case of an L2, there is often some form of structured instruction. These differences can influence the order in which linguistic elements are mastered compared to L1 acquisition, which needs to be taken into consideration when assessing L2 complexity.

In the Second Language Acquisition (SLA) literature, typically a distinction is made between the subconscious process of *acquiring* a language and the conscious process of *learning* it in an instructional setting (Krashen 1987). Since we do not distinguish between these during our analysis, as mentioned in the introduction, we will use these terms interchangeably.

An influential framework for L2 teaching is the Common European Framework of Reference for Languages, which aims at establishing international standards for L2 learning objectives and assessment (Little 2011; North 2007). It defines L2 skills and competences across six proficiency levels: A1, A2, B1, B2, C1, C2, where A1 is the lowest, beginner level, and C2 represents the highest level of near-native proficiency. In the past two decades since its publication, the majority of the European countries have adopted the CEFR guidelines and reorganized language teaching and testing practices to fit into this framework. However, the application of CEFR to language teaching and testing has often been perceived as non-straightforward and challenging. Instead of ready-made solutions, competences are described in terms of rather underspecified "can-do" statements (Little 2011; North 2007) that need to be adapted to a specific L2 learning context. An example of a "can-do" statement for overall reading competences (Council of Europe 2001: 61) is presented in figure 2.1.

This type of description leaves a lot of uncertainty as to how to interpret ex-



C1	*Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.*
B2	*Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.*
B1	*Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.*
A2	*Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.*
	Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
A1	*Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.*

Figure 2.1: CEFR scale example for overall reading skills.

pressions such as “simple texts” and “broad active reading vocabulary”. In fact, some previous studies measuring inter-annotator agreement between teaching professionals report a rather low degree of consensus when assessing CEFR levels. For example, a study aiming at developing an automatic assessment system for L2 Portuguese found only a slight agreement corresponding to a Fleiss kappa of 0.13 among five different language instructors assessing the complexity of L2 reading texts. Teachers agreed with a majority in only 67.27% of the cases on one of the five levels between A1-C1. Similarly, in Pilán, Volodina and Borin (2017), we report a majority agreement for L2 Swedish, sentence-level CEFR judgments of only 50% for exact level match. This indicates a need to further understand how CEFR levels are interpreted and applied in practice.

There have been initiatives to bring down the broad CEFR descriptors to more concrete *critical features*, i.e. linguistic elements to be mastered at different CEFR levels for individual languages (Salamoura and Saville 2010). These concrete content specifications, referred to as *Reference Level Descriptions*, are currently available for Croatian, Czech, English, German, French, Italian, Portuguese and Spanish and they are ongoing for a number of other languages.[4]

When assessing the suitability of a text for L2 learners, the CEFR document (Council of Europe 2001: 165) specifies the following set of aspects to consider:

> In evaluating a text for use with a particular learner or group of learners, factors such as linguistic complexity, text type, discourse structure, physical presentation, length of the text and its relevance for the learner(s), need to be considered.

[4] `https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions`

It is important to note that not only a number of text characteristics are mentioned such as "linguistic complexity" and "length" but also the learner-dependent factor of "relevance".

2.2 Intelligent Computer-Assisted Language Learning

In the early 2000s, Borin (2002b) finds that there is relatively little interaction between the fields of NLP and CALL, proven, among others, by the lack of CALL-related work in major NLP conferences. The past decade, however, has seen a steady growth of ICALL research and today there are a number of workshop series connected to the topic of combining NLP with language learning or with the broader domain of education. These workshops, which have become recurring events attracting an increasing audience, include the Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL), the Workshop on NLP Techniques for Educational Applications (NLP-TEA), the Workshop on Speech and Language Technology in Education (SLaTE) and the Workshop on Innovative Use of NLP for Building Educational Applications (BEA), which grew into a Special Interest Group within the Association for Computational Linguistics (ACL) in 2017. Furthermore, at the ACL 2016 conference, there was a session dedicated to *learner language*.

Although ICALL enjoys a better representation in NLP research nowadays, examples of its practical application in real-life settings, especially for prolonged periods, still remain relatively rare. According to Amaral and Meurers (2011), two ways in which practical applications of ICALL could boost language teaching are automatic feedback generation and handling more complex exercise types.

We can observe two major directions in the development of ICALL research: the analysis of learner texts and that of native language texts for re-use in L2 contexts (Meurers 2012). The former concerns ICALL tasks such as the analysis and the evaluation of learner essays and short answers and providing feedback on these. The study of L1 texts, on the other hand, includes the automatic generation of learning activities based on a targeted selection and an enhanced presentation of such texts. In the following two subsections, we briefly discuss some initiatives relevant for both of these directions.

2.2.1 Reading material selection

A prominent focus of previous ICALL research has been the automatic generation and evaluation of practice material targeting a range of skills and competences. A number of ICALL studies and systems focus on the retrieval of appropriate reading material for L2 English learners. These pedagogically aware search engines share features such as the assessment of texts for their difficulty level and the topic(s) they contain. In the Nordic context, one such initiative is the Squirrel project (Nilsson and Borin 2002), which aimed at creating a web browser useful for locating texts suitable for learners of Nordic languages. Based on an initial example text, the prototype system developed can retrieve similar texts from the web in terms of topic and LIX-based readability.

Similarly, the intelligent tutoring system, REAP (Heilman et al. 2008) assists L2 learners as well as teachers in reading and vocabulary practice. The online tool provides access to web texts that are pedagogically more relevant for a learner in terms of their difficulty and topic than traditional search engine results. The texts are enhanced with dictionary look-up for checking the definition of unknown vocabulary and with a text-to-speech component for listening to the pronunciation of words. Rather than an on-the-fly document retrieval, the system operates based on a pre-compiled annotated database of web pages. A similar system for English offering real-time readability classification for web texts, but rather than for L2 learners, for native speakers of English with low reading skill levels, is described in Miltsakaki and Troutt (2008). Moreover, the FLAIR system (Chinkina and Meurers 2016), besides assessing whether a text is suitable for an L2 English learner's proficiency level and interest in terms of topic, also allows for searches based on specific grammatical constructions. Furthermore, Text Inspector[5] provides CEFR-based lexical complexity information for English texts based on the English Profile project (Salamoura and Saville 2010).

2.2.2 Generation of learning activities

Several ICALL studies investigate *gap-filling* (cloze) exercise generation in which learners have to guess one or more target words omitted from the original version of a sentence or a text. The sentence forming the basis of this type of exercise is commonly referred to as a *seed sentence* (Sumita, Sugaya and Yamamoto 2005) or *carrier sentence* (Smith, Avinesh and Kilgarriff 2010) in the ICALL literature.

[5] `http://www.englishprofile.org/wordlists/text-inspector`

Automating the creation of gap-filling exercises has been explored in a number of studies with slight variations, one of the most popular alternatives being *multiple-choice* exercises. When solving a multiple-choice item, learners have to identify the missing correct solution from a number of options, typically all of which, except one, are *distractors*, that is, incorrect alternatives. A number of systems have been proposed for the fully or partially automatic generation of gap-filling items, mainly for English (Smith, Avinesh and Kilgarriff 2010; Sumita, Sugaya and Yamamoto 2005; Pino and Eskenazi 2009; Mitkov, Le An and Karamanis 2006). There are, however, also a few examples of systems for other languages, e.g. Basque (Arregik 2011) and Swedish (Volodina 2008).

A major issue when automatically generating this exercise type is the selection of appropriate distractors that are difficult enough to challenge learners, but whose level of ambiguity still allows for the identification of the correct alternative. Among the proposed solutions, we can find: information about co-occurrance with the collocate in a distributional thesaurus (Smith, Avinesh and Kilgarriff 2010), the amount of hits in a search engine (Sumita, Sugaya and Yamamoto 2005) and morphological, phonetic and orthographic confusability (Pino and Eskenazi 2009).

Besides multiple-choice exercises, the concept of *bundled gap-filling* has been recently introduced in the ICALL literature (Wojatzki, Melamud and Zesch 2016). Bundled gaps aim at reducing the problem of ambiguity of gap-fill exercises by presenting more than one seed sentence for the same missing target word. The additional sentences facilitate narrowing down the answer options to one correct candidate. The sentences grouped together into a bundle maximize the ratio between the probability of the target word and the other most likely word fitting into the sentences.

Rather than focusing on the the generation of gapped items, Beinborn, Zesch and Gurevych (2014a) investigate NLP approaches to determine their difficulty. The authors propose a model which takes into consideration not only the difficulty of identifying a solution, but also the readability of the excerpt of text in which the gaps appear.

Recently, a number of ICALL systems offering a variety of different activity types have emerged. One such system is WERTi (Meurers et al. 2010), a browser plug-in that enhances web pages for language learners to assist them in improving their grammatical competences. It offers color-highlighting for certain linguistic patterns that are typically difficult for L2 English learners (e.g. prepositions, determiners and phrasal verbs), and it also creates multiple-choice format exercises for practicing those based on the text found on the visited web page.

Language Muse (Burstein et al. 2012) is a system that aims at supporting teachers in generating classroom activities based on texts belonging to different subject areas. The texts provided by teachers are transformed into customizable activities to practice those lexical elements, syntactic structures and discourse relations that may be difficult for L2 English learners.

FeedBook (Rudzewitz et al. 2017) is an example of a paper-based L2 English workbook transformed into its web-based variant. Besides offering an electronic version of the activities, the system also assists teachers when providing *summative feedback* in the form of an overall score or *formative feeback* by correcting and annotating specific learner errors. Teachers' work is supported by automatic suggestions for errors and their types, as well as an alignment of student answers to a target answer with highlighted similarities and differences.

An online system that has gained a remarkable popularity the past years is Duolingo,[6] which combines language learning with crowdsourcing and a gamified design. The system was born as a platform for crowdsourcing translations while providing opportunities of additional practice to L2 learners at the same time (Garcia 2013; Settles and Meeder 2016). Today, Duolingo offers a number of activities to learners including not only translation, but also reading, listening and speaking exercises. Furthermore, it is possible to track one's progress, incentives are provided in the form of reward points and reminders are sent to users to ensure a continued practice.

2.2.3 Analysis of learner language

Throughout the language learning process, learners are required to produce different types of written responses which vary in size and quality depending on the specific task and learners' proficiency level.

A popular means to assessing L2 learning progress is requiring learners to compose an *essay*, a longer piece of text that, for example, narrates a story, describes someone (or something), or presents the writer's point of view. Such texts can be evaluated either in terms of a score (or grade) on the continuum between pass-fail (*essay scoring*) or a level indicating learning progress (*proficiency level classification*).

Automatic essay scoring (or grading) (AES) is a closely related task to the proficiency-level classification of L2 learner texts. Instead of proficiency levels, the goal is to predict numeric scores corresponding to grades or a binary distinction of pass vs. fail. Typically, besides the dimension of linguistic complexity, the relevance to a prompt can also have an impact on the assessment. AES has been an active research area since the 1990s, Burstein and Chodorow (2010) and Miltsakaki and Kukich (2004) provide an overview of such systems for English. E-rater (Burstein 2003) is a commercial essay scoring system that

[6] `https://www.duolingo.com/`

measures writing quality based on a variety of linguistic features. These include, for example, grammatical accuracy, the topical relevance of the vocabulary used (based on a comparison to previously graded essays) as well as features based on discourse analysis.

Annotated learner corpora for languages other than English have also become available in recent years, which enabled extending AES research also to other languages such as German (Zesch, Wojatzki and Scholten-Akoun 2015) and Swedish (Östling et al. 2013). The latter study addresses the automatic grading of Swedish upper secondary school (L1) essays on a four-point scale of grades. The authors found that the performance of their system which achieved 62% accuracy exceeded the extent to which two human assessors agreed on the same data (45.8%). Not only AES, but also proficiency level classification for L2 learner texts has been explored for some languages, these studies are discussed in section 2.3.4.

Besides evaluating longer written learner productions, grading short answers has also been an active research field within ICALL (e.g. Padó 2016; Horbach, Palmer and Pinkal 2013). Such short answers can be the result of reading comprehension questions. An additional dimension typically taken into consideration in such contexts, besides the accuracy of answers, is the relevance of an answer to a question. Padó (2016) investigates the usefulness of different types of features for short answer grading and concludes that lexical, syntactic and text similarity features are among the most efficient predictors. Burrows, Gurevych and Stein (2015) outline the history and trends within short answer grading and find a shift from rule-based methods towards statistical ones.

Regardless of their size, learner-produced texts are challenging to process automatically since, unlike the standard language texts used for training most NLP tools, they often contain errors. This is especially problematic for texts written by lower proficiency learners where the amount of such errors can have a substantial impact on the accuracy of automatic analyses. Both rule-based and statistical methods have been explored for the automatic detection and correction of errors, including finite state transducers (Antonsen 2012) and different hybrid systems proposed in connection with the CoNLL Shared Task on grammatical error correction for L2 English (Ng et al. 2014).

Yannakoudakis, Briscoe and Medlock (2011) present experiments for automatically predicting overall, human-assigned scores for texts written by L2 English test takers at upper-intermediate level. Error-rate features showing a high correlation with these scores were computed both based on the manual annotations in the L2 corpus used and the presence of a trigram in a language model trained on L1 and high proficiency L2 learner texts.

2.3 Linguistic complexity analysis in previous work

As mentioned in the introduction in section 1, linguistic complexity analysis can be used for determining both readability and L2 proficiency levels. Readability analysis and proficiency level classification focus on different types of language users and skills. The former targets reading skills of L1 speakers with low reading levels or cognitive impairment, while proficiency level analysis is employed to assess a variety of skills for L2 speakers. Nevertheless, part of the linguistic complexity features and the proposed approaches (e.g. machine learning) for these two tasks are shared. Thus, linguistic complexity analysis allows us to analyze different text types along similar dimensions.

2.3.1 Linguistic complexity

In cross-linguistic studies with a focus on typology, linguistic complexity is approached in absolute terms, describing complexity as a property of a linguistic system measured in e.g. the number of contrastive sounds (Moran and Blasi 2014). In this thesis, however, we investigate a *relative* type of linguistic complexity from a cognitive perspective, our focus being the ability of L2 learners to process while reading or produce in writing certain linguistic elements in writing at different stages of proficiency.

The effect of other languages known by learners, especially their mother tongue, is usually believed to have some influence on relative linguistic complexity. If the language being learned is genealogically related or geographically close to a language already known by learners, part of the grammatical and lexical peculiarities of the L2 are likely to be already familiar and, consequently, less complex for them (Moran and Blasi 2014). According to Brysbaert, Lagrou and Stevens (2017), however, L2 word processing seems to be more dependent on the characteristics of L2 words themselves rather than interference from L1.

Linguistic complexity plays an important role in efficiently processing and conveying information and, besides successful communication, it can influence performance on a number of different tasks. Tomanek et al. (2010), for example, showed that linguistic complexity has an impact on annotation accuracy of named entities.

2.3.2 Readability

The idea of quantitative readability measures arose in the 1940s when Dale and Chall (1949: 23) defined *readability* in the following way:

> The sum total 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.

This shows that the concept of readability encompasses both factors related to the properties of texts and the characteristics of readers themselves. The former category includes the complexity of morpho-syntactic structures and the semantics of the contained concepts, while readers' skills and their interests vary based on, among others, their experience, educational level and motivation. Thus, similarly to CEFR levels (see section 2.1), readability is influenced by, both more generic, textual factors and personal aspects.

Although the definition of readability and CEFR levels also includes dimensions connected to the reader, most approaches to the automatic classification of these (including the one presented in this thesis) aim primarily to account for the characteristics of the text. The other aspects usually remain unaddressed, which may be due to the lack of data to model different types of readers and language users.

A number of influential readability formulas have been proposed since the second half of the 20^{th} century ranging from simple count-based measures to sophisticated formulas relying on machine learning techniques. Early formulas were based on "surface" text properties such as sentence and word (*token*) length not requiring a deeper linguistic analysis. These formulas, often referred to as *traditional* measures today, mostly target L1 readers and assess the difficulty of texts either in terms of school grade levels or by making a binary distinction based on whether texts are suitable to L1 users with reading difficulties or not. One of the most popular readability formulas proposed for English is the *Flesch-Kincaid Grade Level* (FK) formula (Kincaid et al. 1975). This measure indicates a U.S. school grade level or the length of education (in years) necessary to understand a given text. The formula is computed as presented in (1) based on the number of syllables (N_{syll}), the number of words (N_w) and the number of sentences (N_{sent}).

$$FK = 0.39 \times \left(\frac{N_w}{N_{sent}}\right) + 11.8 \times \left(\frac{N_{syll}}{N_w}\right) - 15.59 \qquad (1)$$

A similar count-based measure suggested for Swedish is *LIX* (*Läsbarhetsindex* 'Readability index') computed as detailed in (2) according to Björnsson (1968). Instead of the number of syllables, the percentage of long words (*longw*) is taken into consideration, which are defined as tokens being longer than 6 characters. Punctuation marks are excluded when considering the number of tokens.

$$LIX = \frac{N_w}{N_{sent}} + \frac{N_{longw} \times 100}{N_w} \quad (2)$$

LIX provides a numeric score between 0 and 100 which can be interpreted according to the values presented in table 2.1 based on Björnsson (1968) in Heimann Mühlenbock (2013: 32). Volodina (2008) explores also a lexically enriched variant of LIX, not only for texts, but also for sentences.

LIX score	Difficulty	Text type
< 25	Very easy	Children's literature
25 – 30	Easy	Young Adults' literature
30 – 40	Standard	Fiction and daily news
40 – 50	Fairly difficult	Informative texts and non-fiction
50 – 60	Difficult	Specialist texts
> 60	Very difficult	Scientific texts

Table 2.1: The LIX scale.

Nominal ratio (NR, Hultman and Westman 1977) is another formula based on morphological information that aims at capturing information density. The simplest form of the measure is the ratio of nouns to verbs in a text. A more sophisticated variant consists of dividing the sum of nouns, prepositions and participles by the sum of pronouns, adverbs and verbs in the text. A higher number of nouns (ca. NR = 1) indicates higher information density and, consequently a higher complexity (Heimann Mühlenbock 2013: 46). To this category belong, for example news texts. Spoken language, on the other hand, typically exhibits a larger amount of verbs (NR = 0.25).

There are a number of online tools available for analyzing texts based on traditional count-based readability measures.[7] NLP-based readability analyzers, are, however, less common. One such system is Pylinguistics[8] for Portuguese.

With the advance of computational analyses of language, more complex, data-driven models have been proposed for a number of languages. They involve multiple dimensions of the text using a deeper computational analysis and often, machine learning methods. Such readability models, with a primary focus on native language users, have been explored for English (Collins-Thompson and Callan 2004; Schwarm and Ostendorf 2005; Miltsakaki and Troutt 2008; Feng et al. 2010; Vajjala and Meurers 2012: e.g.), Italian (Dell' Orletta, Montemagni

[7] E.g. `https://readable.io/` for English and `https://www.lix.se/` for Swedish texts

[8] `https://www.lume.ufrgs.br/handle/10183/147640`

and Venturi 2011), French (Collins-Thompson and Callan 2004), German (vor der Brück, Hartrumpf and Helbig 2008) and Swedish (Larsson 2006; Sjöholm 2012; Heimann Mühlenbock 2013; Falkenjack, Heimann Mühlenbock and Jönsson 2013). Predicting readability in these studies is usually approached as a text classification problem based on supervised machine learning methods relying on annotated corpora. The features that proved predictive in these studies include language models (Collins-Thompson and Callan 2004; Feng et al. 2010) and syntactic features (Schwarm and Ostendorf 2005). Besides investigating readability analysis at the text level, a few studies explore this task also at the sentence level (Dell' Orletta, Montemagni and Venturi 2011; Sjöholm 2012; Vajjala and Meurers 2014). Eye-tracking has been also employed for these purposes (Singh et al. 2016), where an indicator of sentence complexity is measured in terms of reading time.

Graesser et al. (2004) describe a multilevel text analysis tool, the *Coh-Metrix*. The model comprises over 200 different indicators which include aspects related to readability, discourse and cohesion. This tool relies on a large variety of resources, especially for the analysis of lexica in terms of, for example, abstractness, age of acquisition and imageability. Moreover, working memory load is determined based on, among others, the density of logical operators (*or*, *and*, *not*, and *if–then*) and syntactic characteristics such as the amount of noun phrase modifiers.

Heimann Mühlenbock (2013) proposed *SVIT*, a machine learning model for assessing readability in Swedish texts, based on the four dimensions connected to readability as outlined by Chall (1958: 40): (i) *vocabulary load* (e.g. word frequencies); (ii) *sentence structure* (e.g. length of dependency arcs); (iii) *idea density* (e.g. nominal and noun-pronoun ratio); and (iv) *human interest* (expressed as the amount of personal pronouns). An additional dimension consisting of count features was also included. The author employed text classification and showed that these features were more accurate in predicting text difficulty than LIX. The performance of the SVIT model on average was 78.8% accuracy (vs. the 40.5% of using LIX) for classifying easy-to-read vs. ordinary texts belonging to different text genres including children's and adults' fiction, news and information texts (Heimann Mühlenbock 2013: 122).

2.3.3 Proficiency level prediction for expert-written texts

Most traditional readability measures were designed for native language users and they typically aim at determining school grade levels or at making a binary distinction. In the L2 context, however, alternative scales of levels have been proposed which reflect progress in language proficiency. One such scale is the

CEFR, introduced in section 2.1.

In table 2.2 – repeated here for the reader's convenience from the publication in chapter 12 – we provide an overview of studies targeting L2 receptive complexity and compare the target language, the type and amount of training data as well as the methods used. We only include previous work here that shares the following characteristics: (i) texts rather than single sentences are the unit of analysis; (ii) receptive linguistic complexity is measured; and (iii) NLP tools are combined with machine learning algorithms. In table 2.2, studies are ordered alphabetically based on the target language of the linguistic complexity analysis. Under dataset size, we report the number of texts used (except for Heilman et al. 2007), where whole books were employed), followed by the number of tokens in parenthesis when available.

Although the majority of previous work targets L2 English, systems tailored to other languages have also been developed, e.g. for Arabic, Chinese, French and Russian. Two thirds of these machine learning based L2 complexity studies employ the CEFR scale. An alternative to the CEFR is the 7-point scale of the Interagency Language Roundtable (ILR), common in the United States and used in Salesky and Shen (2014). In other cases, the scale of choice remains unspecified (all other studies in table 2.2 which are not related to the CEFR).

In some cases, the corpus used for the experiments was collected from L2 coursebooks and exams, e.g. François and Fairon (2012); Karpov, Baranova and Vitugin (2014); Xia, Kochmar and Briscoe (2016). All the studies working with L2 data employed only instances that are a single coherent piece of texts, except for Heilman et al. (2007), where entire books were used including exercises and activity instructions. This can introduce some noise when modeling complexity given that it can be challenging for NLP tools to handle e.g. the analysis of gapped sentences. Other studies used authentic texts written primarily for L1 readers, which then were rated either by L2 teaching professionals (Salesky and Shen 2014; Sung et al. 2015) or by L2 learners (Zhang, Liu and Ni 2013). The amount of data varies considerably in the previous literature, which may depend on the availability of this type of material, copy-right issues and the annotation cost.

CEFR-based studies have been more commonly treated as a classification problem, while in other cases, regression was chosen. In the latter case, linguistic complexity corresponds to continuous (numeric) rather than discrete values. Opting for classification when using the CEFR levels seems preferable since these are not equally spaced in terms of the time required to reach them "because of the necessary broadening of the range of activities, skills and language involved" when moving higher up on the scale (Council of Europe 2001: 18). The highest (C2) level is omitted from some studies. This level represents a very high proficiency, and L2 material is not always available for this level, most

Study	Target language	CEFR	Dataset size in # texts	Text type	# levels	Method
Salesky and Shen (2014)	Arabic, Dari English, Pashto	No	4 × 1400	Non-L2	7	Regression
Sung et al. (2015)	Chinese	Yes	1578	L2	6	Classification
Heilman et al. (2007)	English	No	4 books (200,000)	L2	4	Regression
Huang et al. (2011)	English	No	187	Both	6	Regression
Xia et al. (2016)	English	Yes	331	L2	5 (A2-C2)	Both
Zhang et al. (2013)	English	No	15	Non-L2	1-10	Regression
François and Fairon (2012)	French	Yes	1852 (510,543)	L2	6	Classification
Branco et al. (2014)	Portuguese	Yes	110 (12,673)	L2	5 (A1-C1)	Regression
Curto et al. (2015)	Portuguese	Yes	237 (25,888)	L2	5 (A1-C1)	Classification
Karpov et al. (2014)	Russian	Yes	219	Both	4 (A1-B1, C2)	Classification
Reynolds (2016)	Russian	Yes	4689	Both	6	Classification

Table 2.2: An overview of studies on L2 receptive complexity.

likely because language users at this stage have little difficulty handling L1 material. When the task is regarded as classification, the most common choice of classifier has been SVMs (see section 4.2.3), but other algorithms have also been tested, for example, random forests (Reynolds 2016). Comparisons of different learning methods are explored in both Curto, Mamede and Baptista (2015) and Xia, Kochmar and Briscoe (2016).

A particular aspect distinguishing Xia, Kochmar and Briscoe (2016) from the rest of the studies mentioned in table 2.2 is the idea of using L1 data to improve the classification of L2 texts. Such transfer learning methods are introduced in section 4.4. For the sake of comparability, the information in table 2.2 describes only the experiments using the L2 data reported in this study.

A large number of features have been proposed and tested in this context. Count-based measures (e.g. sentence and token length, type-token ratio) and syntactic features such as dependency length have been confirmed to be determining factors in L2 complexity (Curto, Mamede and Baptista 2015; Reynolds 2016). Lexical information based on either n-gram models (Heilman et al. 2007) or frequency information from word lists (François and Fairon 2012; Reynolds 2016) and Google search results (Huang et al. 2011) has proven to be, however, one of the most predictive dimensions. Beinborn, Zesch and Gurevych (2014b) offer an in-depth investigation of the role of lexical features in L2 complexity and propose taking into consideration cognates. Heilman et al. (2007) find that these outperform grammatical features, which, although more important for L2 than L1 complexity, still remain less predictive for L2 English complexity than lexical features. Nevertheless, the authors mention that this may depend on the morphological richness of a language. Reynolds (2016), in fact, finds that morphological features are among the most influential ones for L2 Russian texts. Surface coherence features, measured in terms of the presence of connectives, were found not to affect linguistic complexity, at least in L2 English (Zhang, Liu and Ni 2013).

Most receptive L2 complexity models listed in table 2.2 target one language and part of the morpho-syntactic features build on the particularities of these languages. Salesky and Shen (2014), however, investigate a language independent approach. This work constitutes, thus, an example of a trade-off between the amount and the type of linguistic information used and their generalizability to a number of typologically rather different languages.

The state-of-the-art performance reported for the CEFR-based classification described in the studies included in table 2.2 ranges between 75% and 80% accuracy (Curto, Mamede and Baptista 2015; Sung et al. 2015; Xia, Kochmar and Briscoe 2016).

Besides the text-level analyses in table 2.2, studies targeting smaller units also appear in the literature. Linguistic complexity in single sentences from an L2 perspective has been explored in Karpov, Baranova and Vitugin (2014) and in Pilán, Volodina and Johansson (2014). Both studies are CEFR-related, but rather than classifying sentences into individual CEFR levels, a binary distinction is made aiming at categorizing sentences as below or at B1 level vs. above B1. Another approach to linguistic complexity analysis focusing on smaller units is proposed in Ströbel et al. (2016). The authors present Cocogen (Complexity Contour Generator), a system analyzing complexity locally, within specific sliding window sizes, which build up a distribution of linguistic complexity (complexity contour) in the text. The approach was tested by comparing complexity contours in expert-authored texts and written productions of highly proficient L2 English speakers.

2.3.4 Proficiency level prediction for learner texts

Similarly to L2 reading texts which demonstrate varying degrees of linguistic complexity at different CEFR levels, also texts produced by L2 learners manifest varying degrees of complexity at different stages of proficiency. Typically however, receptive linguistic complexity is somewhat higher than the productive counterpart for a learner at a given CEFR level (Barrot 2015).

Despite the availability of CEFR-level annotated corpora for several languages (e.g. Hancke and Meurers 2013; Nicholls 2003; Tenfjord, Meurer and Hofland 2006; Wisniewski et al. 2013), the number of projects aiming at automatizing this task has remained rather limited up to the time of writing. Previous studies include Hancke and Meurers (2013) for L2 German and Vajjala and Lõo (2014) for L2 Estonian. The most predictive features for L2 German include lexical and morphological features. Language-specific morphological features (e.g. amount of distinct cases used) are also among the most informative ones for L2 Estonian, as the authors showed, not only when classifying lower proficiency level texts, but also in later L2 development stages.

A fundamental difference between assessing receptive and productive texts is that, while receptive texts are expected to be relatively error free, the latter ones typically contain a varying amount of L2 errors. These errors can affect the automatic processing of learner-produced texts, and consequently, the estimation of feature values. Information about the nature and the amount of these errors has also been sometimes incorporated into models as features. Spelling errors are usually counted based on the output of a spell checker (Hancke and Meurers 2013; Tack et al. 2017). For the automatic annotation of other types of errors, however, off-the-shelf solutions are not readily available for most languages. Therefore, either this dimension is not included (Vajjala and Lõo 2014) or hand-crafted rules have been proposed for their detection (Tack et al.

2017). The reported state-of-the-art for CEFR-level classification in L2 learner texts in terms of accuracy are between 61% (Hancke and Meurers 2013) and 79% (Vajjala and Lõo 2014) accuracy. Pilán, Volodina and Zesch (2016) present domain adaptation experiments for classifying CEFR levels in L2 Swedish learner essays and report an F1 of 0.747 for texts without error correction.

Research on the automatic assessment of short answers to open-ended questions in terms of using CEFR has been investigated in Tack et al. (2017) for L2 English. The authors proposed an ensemble method consisting of integrating the votes of a number of traditional models (e.g. an SVM, a decision tree etc.) into a single prediction. This voting approach outperformed individual models and obtained an F_1 of .495 and an adjacent accuracy of .978 for a 5-level CEFR classification. Sentence and word length, lexical features and information about the age of acquisition of words were found especially predictive.

2.4 Sentence selection from corpora

2.4.1 Dictionary examples: GDEX

Corpus example sentences are valuable for illustrating authentic language, but they have to be carefully selected to ensure their appropriateness for the intended use (O'Keeffe, McCarthy and Carter 2007). When presented in isolation, a major issue arising is that some sentences may be *context dependent*, i.e. they contain expressions referring to a concept outside of the sentence appearing in the rest of the original textual context they are part of.

A well-known algorithm for the selection of corpus example sentences for illustrating the meaning and the usage of a lexical unit is *GDEX*, Good Dictionary Examples (Husák 2010; Kilgarriff et al. 2008). GDEX aims at mapping requirements of good dictionary examples such as typicality, informativity and intelligibility into measurable features. The linguistic criteria suggested include aspects such as sentence length, word frequency, pronouns, anaphors as well as proper sentence beginning and end (capital letter and punctuations). There are also preferences regarding the position of the search keywords which should ideally occur in the main clause and towards the end of the sentence to allow for a sufficient context for interpretation. GDEX has been integrated in the popular corpus management and query system, *Sketch Engine* (Kilgarriff et al. 2014).[9] Kilgarriff et al. (2008) underline the potentials of approaches like GDEX to support the use of corpora in the language learning classroom by allowing for selecting sentences that are more understandable and have a higher quality. GDEX configurations for languages other than English have been also explored,

[9] `https://www.sketchengine.co.uk/`

including Slovene (Kosem, Husák and McCarthy 2011) and Dutch (Tiberius and Kinable 2015). Parameter values for these languages were tuned based on good dictionary example data.

Besides GDEX, other similar algorithms for corpus example selection have also been proposed in previous literature. A method akin to GDEX for German is described in Didakowski, Lemnitzer and Geyken (2012). The algorithm not only ranks example sentences, but it also includes certain "hard criteria" which, if not met, result in the sentence being excluded from the final set of good example candidates. Boullosa et al. (2017) present a tool for the extraction of example sentences based on their sense. A particular feature of the system is that word senses automatically detected based on a topic model can be further refined through users' feedback.

Volodina, Johansson and Johansson Kokkinakis (2012) present algorithms for the selection of Swedish sentences for not only lexicographic purposes, but also for language learning exercises. Two algorithms were compared, one of which included an additional diversification component aiming at ensuring a more varied subset of good examples. Segler (2007) also describes example sentence selection from a language teaching perspective. Teachers' selection criteria was used to model with logistic regression the selection of sentences illustrating the meaning of words. The main dimensions examined included syntactic complexity and similarity between the original context of a word and an example sentence. Yet another attempt at the automatic selection of example sentences for language, more specifically, word learning is addressed in Tolmachev and Kurohashi (2017). The authors propose a flash card system where, at each iteration, a new example sentence is shown to the learner. The authors propose three sentence quality features: (i) centrality, i.e. grammatical and semantic representativeness; (ii) lexical difficulty based on frequencies and (iii) "goodness". This third aspect aims at ensuring that sentences are complete and well-formed, it remains, however, somewhat loosely defined. Besides these characteristics, the system also takes into consideration diversity among the selected sentences.

Not only rule-based methods, but also machine learning approaches have been explored for corpus example selection in recent years. A supervised machine learning approach to predicting the quality of corpus examples on a four-point scale is presented in Ljubešić and Peronja (2015). The features used for ranking examples include, among others, word and sentence length, word frequency and the amount of undesired elements such as non-alphabetical tokens, pronouns, proper named conjunctions. Geyken, Pölitz and Bartz (2015) investigated extending GDEX to polysemous entries. The authors develop an approach for mapping example sentences to dictionary senses using a classifier trained on sense-annotated data. Lemnitzer et al. (2015) describe a hybrid method for

German where a supervised machine learning classification approach is used to refine the results of a rule-based example selection method.

2.4.2 Sentence selection for ICALL purposes

Corpus examples in ICALL can be employed either as exercise items or as models for word meaning and usage. Corpus example sentence selection for exercise generation is a rather under-researched area. Some studies require sentences only to contain a lexical item or a linguistic pattern that is being focused on in the exercise, but additional aspects such as context dependence and linguistic complexity do not seem to be taken into consideration (Sumita, Sugaya and Yamamoto 2005; Mitkov, Le An and Karamanis 2006; Arregik 2011; Wojatzki, Melamud and Zesch 2016). Language Muse, the activity generator described in section 2.2 also belongs to this category.

Another solution proposed consists of using dictionary examples as seed sentences, e.g. from WordNet (Pino and Eskenazi 2009), since lexicographers have already ensured some aspects of quality for these. The amount of such sentences and the linguistic patterns they exhibit, however, is considerably more limited than what corpora in general, may offer. Moreover, information about L2 learning levels is typically not available in these resources.

A few studies use more refined selection criteria for identifying suitable exercise item candidates. Smith, Avinesh and Kilgarriff (2010) explored GDEX for selecting seed sentences indicating two key aspects in this context: the well-formedness of sentences and a sufficient amount of context in terms of sentence length. Lee and Luo (2016) describe an ICALL system for fill-in-the-blanks preposition learning exercises in which lexical difficulty is determined based on a graded vocabulary lists. In Pilán, Volodina and Johansson (2014) we presented a first version of an algorithm for candidate sentence selection for Swedish, comparing a rule-based and a hybrid method including also machine learning. Context dependence, not assessed by that version of the system, emerged as a key factor behind suboptimal sentences during a small-scale empirical evaluation.

3 SWEDISH RESOURCES FOR LINGUISTIC COMPLEXITY ANALYSIS

In this chapter, we describe the corpora and the lexical resources used to carry out the research and the experiments presented in this thesis. They constitute part of the research and infrastructure work carried out in *Språkbanken* 'Language Bank'.[10] Språkbanken, established in 1975, develops and maintains a large number of linguistic resources and language technology tools, primarily (but not exclusively) for Swedish.

3.1 Corpora

We begin by presenting two corpus-related language technology tools, which were used to access corpora and to annotate them. This is followed by a description of two Swedish L2 corpora: one building on coursebook texts and another based on L2 learner essays.

3.1.1 Korp: a corpus infrastructure

Korp[11] (Borin, Forsberg and Roxendal 2012) is a corpus infrastructure developed in Språkbanken. It provides access to a vast amount of Swedish corpora, whose size is in continuous increase. These include a diverse set of genres and text types such as literary fiction, newspapers, academic texts, blogs, social media texts and the Swedish Wikipedia. Korp also incorporates easy-to-read corpora such as LäsBaRT (Heimann Mühlenbock 2013) and the newspaper *Åtta sidor* 'Eight pages'.[12] Korp offers a number of functionalities for searching and extracting statistical information from these corpora. This includes *concordance* or *keyword in context* (KWIC) search which, given a query, e.g. an inflected

[10] `https://spraakbanken.gu.se/`

[11] `https://spraakbanken.gu.se/korp/`

[12] `http://8sidor.se/`

word form or a lemma, returns a list of sentences containing the item to look up.

3.1.2 Sparv: an annotation pipeline

The corpora in Korp are equipped with automatic linguistic annotations from the *Sparv*[13] pipeline (Borin et al. 2016). The available annotation types include lemmatization, part of speech tagging and dependency parsing. During the lemmatization process, the non-inflected form of each token is searched for in the SALDO lexicon (see section 3.2.3). Part of speech (POS) tagging is performed using the open source trigram tagger, HunPos (Halácsy, Kornai and Oravecz 2007). Dependency parsing is carried out based on the MaltParser (Nivre, Hall and Nilsson 2006). An example of Sparv annotations in XML format is presented in figure B.1 in appendix B.

3.1.3 COCTAILL: a corpus of L2 coursebooks

The COCTAILL (COrpus of CEFR-based Textbooks As Input for Learner Level modelling) corpus is a collection of L2 Swedish coursebooks for CEFR levels between A1 and C1 (Volodina et al. 2014b). It aims at supporting research related to language learning, among others, modeling receptive CEFR levels.

The corpus includes only coursebooks that have been suggested by at least two L2 teachers during the interviews conducted in connection with the corpus preparation. The digitized coursebooks were manually annotated for the different parts they contained with the help of an online editor developed specifically for this purpose. Within this annotation process, lessons (chapters) were identified in the coursebooks together with different subelements they contained: texts (including a genre and a topic), exercises (describing their type and the skills involved), instructions, lists and language examples (sentences aiming at illustrating the use of a lexical or grammatical pattern). Linguistic annotation has been added automatically using the Sparv pipeline.

The size of COCTAILL is 708,589 tokens in total distributed across five CEFR levels as shown in table 3.1 where the number of texts and tokens reported refers to reading comprehension texts only. Certain coursebooks contained lessons belonging to more than one CEFR level, therefore the total number of books in table 3.1 is not the sum of the preceding rows. The CEFR level of texts was derived from the lesson they contained.

[13] `https://spraakbanken.gu.se/sparv/`

CEFR level	**# books**	**# texts**	**# tokens (texts)**
A1	4	101	11,132
A2	4	232	37,259
B1	4	345	79,402
B2	4	314	101,583
C1	2	115	71,991
Total	**12**	**1,106**	**301,367**

Table 3.1: Reading comprehension texts per CEFR level from the COCTAILL corpus.

In the machine learning experiments presented in the subsequent chapters, two different types of elements have been used to create the datasets for modeling receptive CEFR levels. At the text level, only reading comprehension texts have been employed. (These are also referred to as *documents* in chapter 7.) At the sentence level, single sentences were collected from language examples and lists. An example reading comprehension text as well as the type of sentence-level data used can be found in appendix C.

The whole corpus is available in a scrambled version for both download and for browsing through the user interface of Korp.[14]

3.1.4 SweLL: a corpus of L2 learner essays

The SweLL (Swedish Learner Language) pilot corpus is a collection of essays written by L2 learners of Swedish which is comprised of three different subcorpora. Two of these consist of essays written within preparatory language courses and exams for university studies. The third sub-corpus includes essays written by the students of a school receiving newly arrived immigrants. These 144 essays were manually digitized and anonymized. The corpus contains in total 339 essays and ca. 144,000 tokens. Table 3.2 presents the number of essays and tokens per CEFR level in the corpus.

The essays were manually annotated for CEFR levels by language teachers except for two essays whose CEFR level remains unknown. The inter-annotator agreement for CEFR level assignment in terms of Krippendorff's α was 0.80 corresponding to a good annotation quality (Artstein and Poesio 2008).

The corpus includes also meta-data about the learners such as age, gender, L1, level of education and time spent residing in Sweden. Learners were

[14] `https://spraakbanken.gu.se/eng/resource/coctaill`

CEFR level	# essays	# tokens
A1	16	2,084
A2	83	18,349
B1	75	29,814
B2	74	32,691
C1	89	60,455
unknown	2	694
Total	**339**	**144,087**

Table 3.2: Distribution of essays per CEFR level in the SweLL corpus.

adults, 50% of them being between the age of 18 and 20. Most learners had English, Persian or German as L1, but native languages from larger immigrant communities (e.g. Somali, Arabic) are also represented.

The essay variables available for each text include, but are not limited to: the annotated CEFR level, date, topic(s), setting (exam, classroom or home) and information about the use of help materials (e.g. dictionary). The essays cover a range of topics, the most popular ones being health, personal identification and daily life. On average, the length of the essays is approximately 28 sentences, corresponding to an average of 425 tokens, with a substantial difference across CEFR levels.

The SweLL essays contain a varying number of learner errors, typically a decreasing amount as CEFR levels get higher. An indication of this tendency can be seen when inspecting the amount of non-lemmatized tokens per CEFR level (figure 3.1), that is, tokens whose inflected form could not be automatically matched to a dictionary entry during lemmatization.

Non-lemmatized tokens might also include correct out-of-vocabulary words such as infrequent compounds or proper names as in standard (non-L2) corpora, but part of them are misspelled or inexistent words. Learner errors have not been manually annotated up to the time of writing in this resource, which poses some challenges for its automatic analysis with tools trained on standard language. When working with this corpus in our experiments, for a better approximation of feature values, besides its original version, we used an automatically error-normalized variant of the essays (see section 10). Chapter C in the appendix shows an example of a learner essay both in its original and its error-normalized version using an off-the-shelf software.

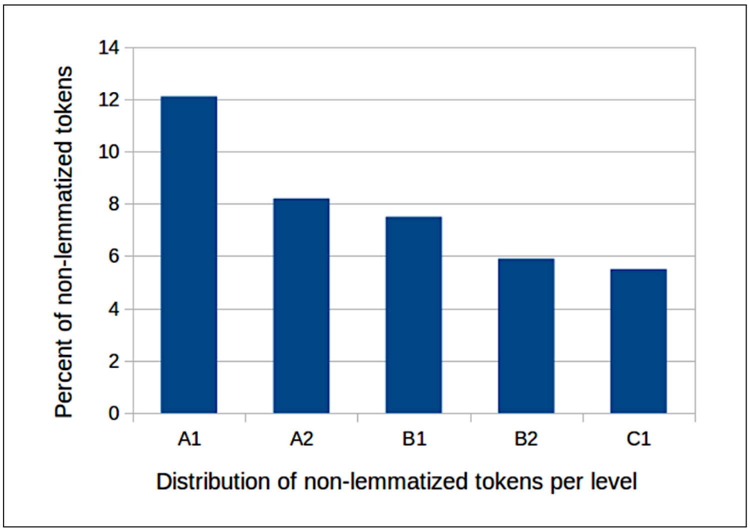


Figure 3.1: Distribution of non-lemmatized tokens per level.

3.1.5 A teacher-evaluated dataset of sentences

Besides the single sentences occurring in COCTAILL, we constructed a small dataset based on a user evaluation of the sentence selection system HitEx, developed as part of this thesis work. HitEx aims at identifying sentences from corpora suitable as exercise items. A detailed presentation of the HitEx system and its evaluation can be found in sections 5.2 and 9.4. The sentences included in this dataset are authentic examples from generic corpora that have been automatically assessed for their CEFR level and which have been filtered for their well-formedness, independence from the rest of their textual context and some additional lexical and structural criteria (e.g. abbreviations, interrogative form) using HitEx. Section E.2 in the appendix provides further information about the exact sentence selection criteria used.

Out of the original 330 sentences from the evaluation material, we only included in this dataset the subset of sentences: (i) that were found overall suitable (with an evaluation score $>=$ 2.5 out of 4); and (ii) where a majority of teachers agreed with the CEFR level assigned automatically by our system. This subset was complemented with 90 sentences for the otherwise insufficiently represented A1 level from the COCTAILL corpus (thus these are a subset of the sentence-level dataset from the study in chapter 7). The distribution of sentences per CEFR level in the dataset is presented in table 3.3. The total number of tokens in the dataset is 4,060. The dataset is available for download among the linguistic resources of Språkbanken.

CEFR level	**# sentences**
A1	98
A2	82
B1	58
B2	92
C1	45
Total	**375**

Table 3.3: Distribution of sentences per CEFR level.

3.2 Lexical resources

In the following, we describe the lexical resources utilized in our experiments. All these resources are freely available and have been, in part or entirely, developed in Språkbanken.

3.2.1 KELLY

KELLY is a frequency-based vocabulary list created during the EU project *KEywords for Language Learning for Young and adults alike*. The main purpose of the KELLY project was to provide a language learning resource available in nine different languages, namely Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish.

The underlying data used for estimating the frequencies consisted of a corpus of web texts, SweWaC, from the 2010s, comprising 114 million tokens. Both the corpus and the compilation of the statistical information about frequencies was performed using the SketchEngine corpus analysis tool (Kilgarriff et al. 2014).

The Swedish KELLY list includes the following main types of information: the (i) headword lemma and (ii) its part of speech (word class), an (iii) identification number based on the relative position in decreasing order of frequency, (iv) raw frequency, (v) normalized, word-per-million (WPM) frequency and (vi) a CEFR level. Grammatical information describing, for example, the gender of nouns, is also available for certain tokens. Table 3.4 shows an example entry from the KELLY list.

The Swedish KELLY list includes 8,425 entries, which have been equally divided into CEFR levels based on their frequency, with ca. 1400 headwords

Lemma	ID	Raw freq	WPM	CEFR	Source	Word class
häst	1319	7,920	69.47	A1	SweWAC	noun-en

Table 3.4: An example headword *häst* 'horse' from KELLY.

per level: the more frequent a headword was, the lower was the assigned CEFR level. In the final stage of its compilation, the list has undergone a proofreading process to ensure the quality of the resource. A small number of items (<1%) have been manually added to overcome data sparsity and to fill gaps in certain finite sets of basic (A1 level) vocabulary such as weekdays. 30% of headwords were added based on translations from KELLY lists for the partner languages.

In the machine learning experiments described in sections 7 – 11, KELLY has been utilized as a source of information for the lexical features both when calculating the distribution of tokens per CEFR levels and when computing average frequency.

The Swedish KELLY list, compared to SVALex and SweLLex described in the next section, has the advantage of stemming from a substantially larger corpus. The texts contained in the corpus, however, were not explicitly written for L2 learners, its pedagogical relevance in an L2 scenario would therefore need to be evaluated. An attempt in this direction is described in section 11, where we show that the CEFR level distributions of tokens in essays written by L2 learners according to KELLY are significantly flatter compared to using CEFR level suggestions from SVALex.

3.2.2 SVALex and SweLLex

The SVALex and SweLLex lists are not only intended for L2 learners, teachers and researchers, but they have also been created based on L2 data. These lists contain normalized frequency distributions of lemmas across CEFR levels. SVALex frequencies were estimated using the COCTAILL corpus described in section 3.1.3, that is, reading passages from L2 coursebooks. The CEFR level for the texts and the tokens contained in them, were derived from the CEFR level of the lesson (unit) in which the texts occurred. SweLL frequencies, on the other hand, were computed using the SweLL corpus consisting of L2 learner essays (see section 3.1.4).

Although the underlying corpora are different, both lists are based on the same methodology that has been proposed first for FLELex (François et al. 2014), a list of L2 French frequency distributions across CEFR levels. Each entry in SVALex and SweLLex is a *lemgram*, a combination of a lemma, its part of speech, and an index number, identifying a table of inflectional and compound forms. Besides single-word units, multi-word expressions were also included based on the annotations from the Sparv pipeline. The amount of multi-word expressions in SVALex is ca. 15% and in SweLLex ca. 8%.

An important aspect of these lists belonging to the *CEFRLex* family, a set of CEFR-linked frequency lists based on this methodology, is that distributions are not based on raw frequencies (RF). RF counts would be unreliable with *context-sensitive* words, i.e. lexical items that are over-represented in one (or only a few) texts, but which are under-represented across different texts and textbooks. This idea of representativeness is similar to TF-DF (term frequency - document frequency), and it serves the purpose of identifying lexical items that, since they occur more systematically across texts and coursebooks, can be considered more objectively as part of the vocabulary at a certain CEFR level.

Normalized frequency counts were based on Carroll, Davies and Richman (1971). First, a *dispersion index* (D) is calculated in the following way:

$$D_{w,K} = [\log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}]/\log(I) \tag{3}$$

In equation 3, K represents the number of difficulty levels (5 in the case of SVALex, 6 for SweLLex) and I stands for the amount of textbooks i at level $k \in K$. The value of D is computed based on p_i, the probability of word w in textbook i at level k.

To decrease the influence of context-specific lexical items in the per-CEFR distributions, raw frequency counts were combined with D as shown in equation 4, where N_k is the total number of tokens at level k. To compute U, a per-million normalized frequency value for word w, not only the dispersion index D is applied on RF, but also f_{min}. The value of f_{min} was computed as $\frac{1}{N}\sum_i f_i N_i$ where f_i is the frequency of a word in textbook i and N_i is the total number of tokens in textbook i.

$$U = \frac{1,000,000}{N_k}(RF * D_{w,K} + (1 - D_{w,K}) * f_{min}) \tag{4}$$

The lists are browsable also through an easy-to-use web interface,[15] where users can look up and compare the frequency distributions of lexical items both within and between the two Swedish resources. An example is shown in figure 3.2 where the receptive (SVALex-based) and productive (SweLLex-based) use of the verb *äta* ‘to eat’ is compared based on normalized frequency values.

[15] `http://cental.uclouvain.be/svalex/`

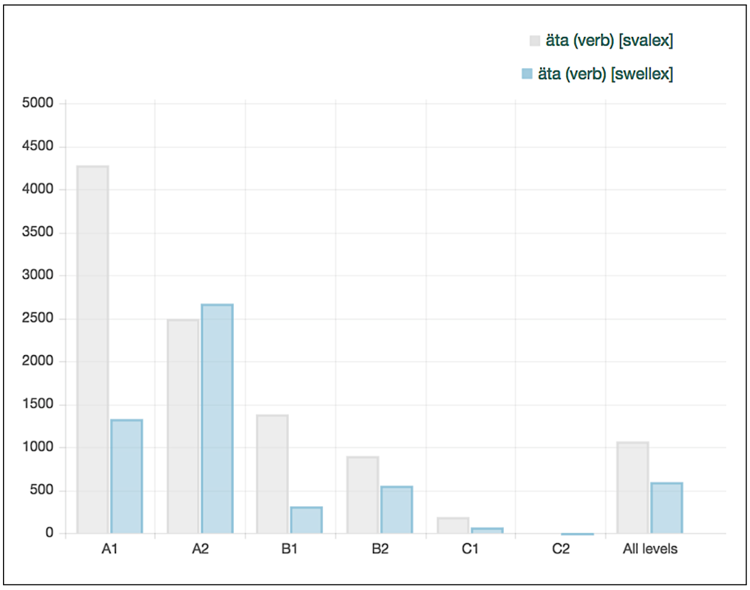


Figure 3.2: The browsable web interface of SVALex and SweLLex.

Further research on these lists described in Alfter et al. (2016) proposed a method for mapping the frequency distributions to a single CEFR level at which a word can be considered known receptively or productively. Instead of adopting the first CEFR level at which a word occurs or the level at which it is used most, the *significant onset of use (SOU)* is proposed as indicator of a level. The SOU value maps a word to the CEFR level at which the word is used considerably more often than at the preceding level. This measure compares the normalized frequency of a word at two adjacent levels k and $k-1$ and computes the difference. Frequencies are normalized so that they lie between 0 and 1 and, whenever the difference exceeds an empirically chosen threshold (SOU = 0.4), the word is mapped to level k, otherwise it is mapped to $k-1$.

This mapping approach was evaluated with the help of a vector space model. The results showed that the CEFR levels predicted based on the SOU mappings were often clustered close to each other in a vector space model created using the SweLL essay corpus with CEFR level information incorporated for each token in the representations.

The size of the resulting word lists in number of lemgrams with (MAPPED) and without (ORIG) the mapping to CEFR levels is presented in table 3.5. In ORIG, lemgrams appearing at different CEFR levels have been counted separately for each level of occurrence. Thus, these numbers represent the number of unique lemmas per CEFR level, but not across all CEFR levels.

	SVALex		**SweLLex**	
CEFR	ORIG	MAPPED	ORIG	MAPPED
A1	4,976	779	398	386
A2	6,995	1,950	1,327	1,056
B1	10,780	3,524	2,380	1,516
B2	7,349	4,659	2,396	959
C1	8,348	4,710	3,566	1,534
C2	7,433	-	145	-
Total	**15,681**	**15,622**	**6,965**	**5,451**

Table 3.5: Distribution of entries per CEFR level in SVALex and SweLLex.

3.2.3 SALDO

SALDO (Borin, Forsberg and Lönngren 2013) is a lexical-semantic resource for Swedish based on associations between word senses. It represents an alternative structure to the wide-spread resource, *WordNet*, (Fellbaum 1998) in which word senses are organized into *synsets*, i.e. sets of synonyms connected to each other by different lexical and semantic relations (e.g. hyperonymy). SALDO, besides representing a different theoretical choice, aims also at being easier to use computationally. Unlike WordNet, SALDO covers all parts of speech (POS) and it is based on association relations among hierarchically organized word senses. A further difference compared to WordNet is that terms with different POS can be associated with each other. The top node of the hierarchy is an artificial node, *PRIM*, whose children consist of 43 core senses. Senses become more and more peripheral as we move down in the hierarchy. Instead of textual definitions, each sense is defined in terms of another manually selected sense, a mandatory primary *descriptor* (PD), and one (or more) optional secondary sense descriptors. Each descriptor is a more central semantic neighbor of a given entry. Centrality is determined in terms of frequency, stylistic unmarkedness, morphological complexity and directional semantic relations (e.g. hyperonyms as descriptors of their hyponyms). Being a semantic neighbor entails a direct semantic and, occasionally, syntagmatic relationship between the terms. Due to the hierarchical structure of SALDO, each sense can be characterized by a *semantic depth*, i.e. a certain distance from the root node in terms of a list of ancestor senses, the average depth in the senses listed in the resource being 6 senses. Appendix D shows some example entries from SALDO.

4 Machine learning methods

Given the highly interdisciplinary nature of the thesis topic that might attract readers from different disciplines, in this chapter, we introduce the main statistical and machine learning methods underlying the papers listed in parts II and III. Clarification on some additional methods employed are detailed in the relevant publications.

4.1 Basic notions

In a machine learning scenario, a set of data points is used as example to make predictions about a certain property of unknown data (Witten et al. 2011). That is, given some input x, we would like to learn a function $f(x)$ that allows us to predict an output y, which can be either binary, real-valued (e.g. grade levels) or categorical (e.g. types of sentiments). In NLP, structured predictions are also common, in which case the output is, for example, a dependency tree (see figure 5.1). Each sample in a dataset is commonly referred to as an *instance*, and its properties (*features* or *attributes*) are typically encoded as a vector of numerical values.

As an example, let us consider predicting the readability of a sentence as a binary output $Y \in \{0,1\}$, where 0 is an easy-to-read sentence and 1 is a hard-to-read one. Let our instance be the sentence *I have a dream.* with an annotated label $y_1 = 0$. We could then represent this instance as a two-dimensional vector $x_1 = (4,1)$, where the first dimension corresponds to the number of tokens excluding punctuation and the second dimension stands for the number of verbs.

In some cases, hand-labeled (*annotated*) examples might be available to learn from. Based on the availability of labeled data, we can distinguish between three types of machine learning approaches: *supervised*, based on labeled instances, *unsupervised*, relying on unlabeled data points only (Hastie, Tibshirani and Friedman 2009), and *semi-supervised* approaches based on both labeled and unlabeled data (Søgaard 2013).

Instances are typically divided into *training set*, showed as examples to the learning algorithm during training time, and *test set*, used to measure performance by comparing predicted output to manual annotations in a supervised setup. The use of a *development set* is also common for hyperparameter tuning of a model and during the process of selecting the most predictive features.

In most cases, the natural language data used for training machine learning models contains a varying degree of *noise*, that is, erroneous values, and *bias*, stemming from preferences and prejudices that may underlie certain human decisions. Moreover, not all data points might follow the general trend in the data; such instances are *outliers*. When a model fits even to random irregularities in the data and its degree of generalizability to new, unseen data points decreases, we face a case of *overfitting* (Witten et al. 2011). On the other hand, a system may be *underfitting* and fail to capture patterns in the training data.

4.2 Learning algorithms

In this section, we briefly introduce three statistical models employed in the studies appearing in parts II and III. These algorithms are available in a number of machine learning tools, out of which WEKA[16] (Hall et al. 2009) and scikit-learn[17] (Pedregosa et al. 2011) have been used in the included publications.

4.2.1 Linear regression

Linear regression is a simple statistical method for predicting *numerical* outputs. The output is computed as the inner product of a vector of feature values representing a text (or a sentence) and a *weight vector* consisting of regression coefficients for each feature (Witten et al. 2011). These weights are estimated based on the training instances. This linear combination is presented in (5), where y_{pred} denotes the predicted output, w_1, w_2 and w_n are the weights and x_1, x_2 and x_n are feature values of instance x.

$$y_{pred} = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n \tag{5}$$

When choosing the weights, the goal is to minimize the sum of the squares of the differences between y_{pred} and the actual output label y in the annotated training data. For the i^{th} instance among n instances, this difference is computed thus as:

[16] `https://www.cs.waikato.ac.nz/ml/weka/`

[17] `http://scikit-learn.org`

$$\sum_{i=1}^{n}(y^i - y^i_{pred})^2 \tag{6}$$

Linear regression is computationally light and rather straightforward, which characteristics undoubtedly have contributed to its popularity among statistical applications. It does assume, however, linear dependencies in the data, which might not be the case in certain scenarios.

4.2.2 Logistic regression

Regression techniques can also be used to perform classification and to make binary predictions for an instance. A value of 1 can indicate then that an instance has a certain output y, 0 that it does not. Rather than predicting these values directly, the aim is to compute the probability of an instance of belonging to a class y (Witten et al. 2011). In other words, given a set of input feature vectors X, the conditional distribution of the classes Y has to be found. In logistic regression, the input values are also linearly combined, but the real-valued predictions are transformed into probabilities (p) with the help of the *logistic function* (7).

$$p = \frac{1}{1+e^{-(w_0+w_1x)}} \tag{7}$$

Unlike linear regression, which minimizes the sum of squared errors to find the coefficients, logistic regression selects the parameters that maximize the *likelihood* of observing the input values. Similarly to linear regression, the output of logistic regression is also more informative than a number of other algorithms, but unlike its linear counterpart, logistic regression can also capture non-linear relationships. Furthermore, probabilities provide useful insights into the system's confidence in the predictions.

4.2.3 Support vector machines

Support vector machines (SVMs, Vapnik 1998) have been successfully used in many NLP problems. SVMs tackle the problem of non-linearity by mapping the data, with the help of a *kernel function*, to a space with a higher number of dimensions where a linear boundary can be identified. SVMs aim at identifying a *hyperplane*, a boundary separating instances belonging to different classes. *Support vectors* are the data points determining the class boundary, the optimal *margin* (m), i.e. distance between these points and the separating hyperplane being the largest one (Witten et al. 2011: 223–225). Given an input vector x, a weight vector w and a bias term b, the equation defining the separating hyperplane (H_0) is: $w \cdot x + b = 0$. A two-dimensional representation of SVMs is provided in figure 4.1.

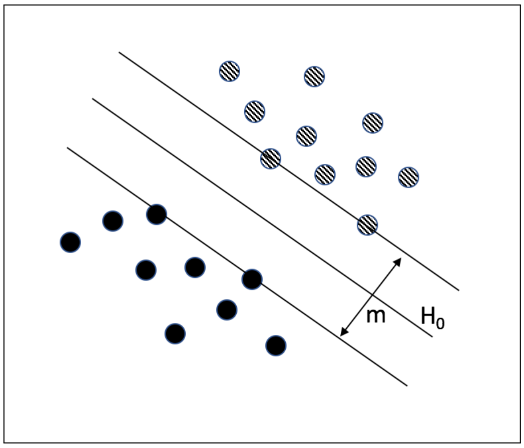


Figure 4.1: A separating hyperplane in SVMs.

Since some outliers may be present in the data, a *soft margin* allowing for some misclassifications can sometimes yield more appropriate models. An advantage of SVMs is that they are robust to overfitting. They are less sensitive to the ratio of the number of attributes and the number of instances, hence they are adequate also for smaller datasets. On the other hand, however, they are less transparent than linear regression models. Hämäläinen and Vinni (2006) compare different learning algorithms on smaller educational datasets and find that SVMs yield a high performance, in particular when the number of attributes is high.

4.3 Evaluation measures

A number of different measures can be used to evaluate the performance of supervised machine learning algorithms. When comparing the actual output (annotations) to the predicted output in a classification scenario, we can distinguish four different cases (Witten et al. 2011: 164). For example, given the task of determining the readability of a text, some easy-to-read instances will hopefully be classified as easy (*true positive*) and some hard-to-read instances will not be classified as easy (*true negative*). Our system, might, however, make some mistakes and classify some hard-to-read sentences as easy (*false positive*) and it may fail to identify all easy-to-read sentences, classifying them as being hard instead (*false negative*). This example is illustrated also in table 4.1, where the columns represent predictions and rows the actual classes.

		Predicted	
		Easy	Hard
Actual	Easy	True Positive (TP)	False Negative (FN)
	Hard	False Positive (FP)	True Negative (TN)

Table 4.1: Comparing predictions and actual classes.

Given these four cases, we can compute a number of measures (Witten et al. 2011: 175). The *accuracy* of our system is the sum of true and false positives divided by the total number of predictions. *Precision* can be calculated as $TP/(TP+FP)$, i.e. dividing the number of easy texts classified as easy, by the number of all instances labeled as easy (regardless of whether they were correct) in our example. *Recall*, on the other hand, equals to $TP/(TP+FN)$, the number of texts classified as easy divided by the total number of texts annotated as easy in our dataset.

Precision and recall can also be combined into a single measure, namely the *F score*, which is popular in the NLP literature. A common version of this measure is the F_1 score (8), where precision and recall are weighted equally.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

When evaluating performance, it is common to analyze the *confusion matrix* over the predictions made. A confusion matrix is a tabular representation of predicted and actual values per class. Table 4.1 can be seen as a example of a confusion matrix for a 2-class setup. Confusion matrices provide information about which classes are frequently misclassified as certain other classes.

In the case of numeric predictions, measures providing information about the extent of errors are very informative. Such models, e.g. linear regression, are often evaluated in terms of mean-squared error (MSE), see the formula in (6). To obtain values that are of the same dimension as the predicted values, often the square root of MSE (RMSE) is reported instead (Witten et al. 2011: 180–182).

The classification of CEFR levels has been treated both as a numeric and as a multi-class prediction task, the latter being more wide-spread (see table 2.2). A measure that has commonly been adopted when describing the performance of CEFR level classification systems is *adjacent accuracy* (or within 1-level accuracy), as in e.g. François and Fairon (2012) and Vajjala and Lõo (2014).

When computing adjacent accuracy, misclassifications occurring with neighboring classes are considered correct. Thus, even when linguistic complexity is treated as a classification task in an L2 setting, the categorical labels of CEFR levels are often treated implicitly as ordered numeric values.

Another similar measure is *quadratic weighted kappa* (QWK), proposed and used in the context of automated essay scoring (e.g. Zesch, Wojatzki and Scholten-Akoun 2015) and the Automated Student Assessment Prize (ASAP) of the Hewlett Foundation.[18] The measure expresses the extent to which predicted outputs match the actual values from the annotations. Instead of considering every error equally severe (of value 1), a set of weights are assigned to all predictions computed according to equation 9, where the dividend is the difference between the predicted (y_{pred}) and the actual output (y), and N_y is the total number of categories.

$$QWK = 1 - \frac{(y - y_{pred})^2}{(N_y - 1)^2} \tag{9}$$

In the case of CEFR level classification, the use of this measure requires the transformation of the categorical labels A1, A2 etc. into consecutive numerical values 1, 2 etc. For example, let N_y be 5 (1, 2, 3, 4 and 5 corresponding to the CEFR levels A1-C1) and let the predicted class (y_{pred}) be B1, but y be A2. Then, we would compute QWK as $1 - (2 - 3)^2/(5 - 1)^2 = 0.9375$. The complete set of QWK values for a 5 class problem would be 1, 0.937, 0.750, 0.437 and 0 for a difference of 0 (when $y_{pred} = y$) to 5 classes respectively. Thus, higher values of QWK indicate a higher degree of agreement between predicted and actual classes.

When evaluating a machine learning model, one can either measure performance on a held-out test set, or use *n-fold cross-validation*. Cross-validation consists of n train-test iterations, where a different n^{th} portion of the data is set aside for testing, while the rest is used for training each time (Witten et al. 2011: 152–154). Common numbers for n include 10 and 5 depending on the size of the dataset. The folds can be *stratified*, that is, created in a way that they maintain the same distribution of classes as the one present in the original dataset.

Machine learning and NLP sytems in general can also be evaluated with the help of human assessors or target user groups, which is especially valuable when only a small amount of (or no) annotated data is available (Paroubek, Chaudiron and Hirschman 2007). An example of such an evaluation is presented in section 9.4.

[18] `https://www.kaggle.com/c/asap-aes#evaluation`

4.4 Domain adaptation

Supervised machine learning methods require annotated data which, however, is not always readily available for the target language or task. The main reason lying behind this data sparsity problem is that the creation of annotated corpora is expensive both in terms of time and effort. Sometimes, a sufficient amount of annotated data may not be at one's disposal from exactly the same domain, but it may be available from a related domain. In this case, *transfer learning* methods can be applied. These allow for learning a predictive function for the addressed (*target*) domain by using knowledge from an additional, related (*source*) domain.

Pan and Yang (2010) present a survey and a categorization of transfer learning methods. The rationale behind these methods is humans' ability of solving tasks faster if similar experiences are available. As an example, the authors mention learning to play the piano, when one knows already how to play the organ. The difference between datasets can consist, for example, of the language, the text genre, the features used or the data distribution. A type of transfer learning method, used also in this thesis, is *domain adaptation*, where the set of outputs and the feature space are shared between the datasets, but the underlying probability distributions are different.

There are a number of different ways in which source data can improve predictions in a target domain using domain adaptation (Daumé III and Marcu 2006; Pan and Yang 2010). These either concern the type and amount of instances included or manipulate the feature space. With regard to the choice of training instances, in the simplest case, suitable also when no annotated target domain instances are available, a model can be trained using the source domain data only and this classifier can be then used to make predictions about the target data.

If some annotated data is available from the target domain, this can be combined with the source domain instances at training time. The two types of data can be combined having equal weight, alternatively, target domain instances might be assigned a higher weight to receive an increased influence in the predictions made. Furthermore, instead of including all available source domain instances, opting for a selected subset of these can have a positive effect on training performance since it eliminates too dissimilar instances which could act as noise for the target task. A classifier trained on a small amount of target domain instances can be used to classify source domain data and source domain instances can be selected based on whether their label matches the predictions of this model. A similar method is presented in Jiang and Zhai (2007).

When it comes to manipulating the feature space, target domain instances can receive an additional dimension based on a classifier trained using only source data. That is, once a model is trained on source domain instances, the predictions of this system are added as extra feature for each target instance. Then, a new model is trained on the target domain data equipped with this additional feature.

A popular domain adaptation technique is EASYADAPT (Daumé III 2007). This method is also based on augmenting the feature space by including three versions of each feature: besides maintaining the original one, a source-specific and a target-specific version is added. We can formalize this as the application of the mapping function $\phi^S(x) = \langle x, x, 0 \rangle$ to each feature vector x in the source domain and $\phi^T(x) = \langle x, 0, x \rangle$ in the target domain, where 0 is a zero vector of length $|x|$. This method has been successfully applied in a number of tasks including also ICALL-related topics such as essay scoring (Phandi, Chai and Ng 2015) and CEFR-level classification of reading texts (Xia, Kochmar and Briscoe 2016).

5 Proficiency level prediction for ICALL purposes

In this chapter, we provide an overview of the main findings and contributions arising from this thesis work. We introduce and motivate our features for linguistic complexity analysis, as well as summarize and discuss the results of the publications from parts II – IV.

The groups of studies included in this thesis vary along two dimensions when it comes to the automatic prediction of proficiency levels: (i) the size of the linguistic unit taken into consideration, namely sentences vs. texts, and (ii) the type of data, that is, expert-written texts as opposed to learner-produced texts. In connection to these, we investigate two directions of the applicability of proficiency level prediction within the field of ICALL: the identification of suitable language learning materials from authentic sources and the evaluation of learner-produced texts.

Section 5.2 addresses the question of how well the feature set described in section 5.1 can predict L2 receptive linguistic complexity in terms of CEFR levels. Here, we also explore the difference between text- and sentence-level linguistic complexity analysis. Then, focusing on L2 applications of sentence complexity, we provide an overview of our framework for the automatic detection of exercise item candidates outlining the aspects that are relevant for identifying good candidates. Moreover, we investigate the extent to which the selected sentences correspond to L2 teachers' expectations and L2 learners' knowledge at different CEFR levels.

Section 5.3 focuses on linguistic complexity in learner produced texts. Since our learner essay dataset is rather small, we investigate whether coursebook texts can be useful as additional training data to improve CEFR-level prediction. Furthermore, we explore the effects of correcting learner errors on this domain transfer. Finally, we study whether lexical complexity features can be better estimated based on frequencies from a coursebook rather than from a web text corpus.

After analyzing the importance of single features in section 5.4, we conclude this chapter with a short description of how the research outcomes have been integrated into an online learning platform in section 5.5.

5.1 A flexible feature set for linguistic complexity analysis

In this section, we present and motivate the features underlying all linguistic complexity experiments. The feature set is “flexible” in the sense that it can be applied to different types of L2 data and units of analysis (e.g. texts or sentences). Consequently, it does not incorporate text-level features (e.g. discourse-related aspects) or learner language specific ones (e.g. L2 error features). The feature set, however, does aim at covering the basic constraints for complexity measures outlined in Menn and Duffield (2014: 283–285): it identifies the presence of potentially difficult structures and it takes into consideration aspects of frequency as well as the availability of competing interpretations causing ambiguity.

Our feature set is comprised of 61 features in total. Previous literature presents different alternatives for how to group these features. We present them based on the type of NLP tools and resources used, dividing them into five sub-categories: *count-based*, *lexical*, *morphological*, *syntactic* and *semantic* as shown in table 5.1. We, then, provide theoretical motivation for these, including cognitive aspects and a second language learning perspective.

Name	Type	Name	Type
Sentence length	Count	Modal V to V	Morph
Avg token length	Count	Particle IncSc	Morph
Extra-long token	Count	3SG pronoun IncSc	Morph
Nr characters	Count	Punctuation IncSc	Morph
LIX	Count	Subjunction IncSc	Morph
Bilog TTR	Count	PR to N	Morph
Square root TTR	Count	PR to PP	Morph
Avg KELLY log freq	Lexical	S-V IncSc	Morph
A1 lemma IncSc	Lexical	S-V to V	Morph
A2 lemma IncSc	Lexical	ADJ IncSc	Morph
B1 lemma IncSc	Lexical	ADJ variation	Morph
B2 lemma IncSc	Lexical	ADV IncSc	Morph
C1 lemma IncSc	Lexical	ADV variation	Morph
C2 lemma IncSc	Lexical	N IncSc	Morph
Difficult W IncSc	Lexical	N variation	Morph
Difficult N&V IncSc	Lexical	V IncSc	Morph
OOV IncSc	Lexical	V variation	Morph
No lemma IncSc	Lexical	Function W IncSc	Morph
Avg. DepArc length	Syntactic	Neuter N IncSc	Morph
DepArc Len $>$ 5	Syntactic	CJ + SJ IncSc	Morph
Max length DepArc	Syntactic	Past PC to V	Morph
Right DepArc Ratio	Syntactic	Present PC to V	Morph
Left DepArc Ratio	Syntactic	Past V to V	Morph
Modifier variation	Syntactic	Supine V to V	Morph
Pre-modifier IncSc	Syntactic	Present V to V	Morph
Post-modifier IncSc	Syntactic	Nominal ratio	Morph
Subordinate IncSc	Syntactic	N to V	Morph
Relative clause IncSc	Syntactic	Lex T to non-lex T	Morph
PP complement IncSc	Syntactic	Lex T to Nr T	Morph
Avg senses per token	Semantic	Relative structure IncSc	Morph
N senses per N	Semantic		

Table 5.1: The feature set proposed for linguistic complexity analysis of L2 Swedish.

5.1.1 Count-based features

Our feature set includes seven indicators that are based on simple counts, commonly used in traditional readability measures (e.g. Dale and Chall 1949). We consider sentence length in terms of the number of tokens not including punctuation as well as in terms of the number of characters. Sentence length can indicate syntactic difficulty and it can be a sign of e.g. multiple clauses or larger noun phrases. Average token (T) length is computed based on the number of characters. Extra-long words, i.e. tokens longer than 13 characters, are also counted as in Heimann Mühlenbock (2013) since compounding, frequent in Swedish, can result in particularly long words. Compounds can be more challenging to process since the boundaries between the parts might not be obvious, and their meaning is not necessarily *compositional*, a sum of the meaning of the parts. LIX, a traditional Swedish readability formula (see Section 9.2) combines the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters (Björnsson 1968). Type-token ratio (TTR), the ratio of unique tokens to all tokens, is an indicator of lexical richness (Graesser et al. 2004). Using a more varied set of vocabulary items in a text increases its lexical complexity as the links between words referring to similar concepts need to be recognized. We use bi-logarithmic and a square root TTR to decrease the effect of text and sentence length as in Vajjala and Meurers (2012).

5.1.2 Word-list based lexical features

Besides richness, the frequency of words also influences lexical complexity. According to the *lexical entrenchment hypothesis* (Diependaele, Lemhöfer and Brysbaert 2013) low frequency words are especially demanding to understand or produce for beginner learners because their lexical representations for these are weaker. The more frequent a word is, the higher the likelihood that different learners encounter it early in their L2 development process and this repeated exposure has positive effects on the ease of their processing (Diependaele, Lemhöfer and Brysbaert 2013; Graesser et al. 2004). When L2 learners' vocabulary size at higher proficiency levels becomes comparable to that of L1 speakers, this frequency effect disappears as also Brysbaert, Lagrou and Stevens (2017) confirm based on lexical decision time experiments. Similarly, when measuring productive linguistic complexity, higher proficiency level learners' writings are typically characterized by the increased use of words with low frequency and a higher lexical diversity overall (Crossley and McNamara 2011).

As source of information for lexical frequencies, we use the KELLY list presented in section 3.2.1. The frequencies in this word list are calculated from a corpus of web texts which makes this resource independent from the texts and sentences of the datasets used in our experiments. The use of log frequencies is motivated by the fact that these better reflect reading times since the effect of high frequency function words and other common, but less frequent words, is balanced out (Graesser et al. 2004).

Since the amount of L2 Swedish data at our disposal was rather small, we do not employ n-grams as features. Instead, we propose weakly lexicalized features to increase the generalizability of our models on unseen data. We represent each token by its corresponding CEFR level. To extract this information, we use the CEFR level suggested for each lemma in the KELLY list. When considering the distribution of tokens per CEFR level, instead of absolute counts, we use a normalized value to reduce the influence of sentence length. We compute *incidence scores* (INCSC) in a similar fashion to Graesser et al. (2004), by dividing 1000 with the total number of tokens (N_t) and multiply that with the count of a certain category of tokens (N_c) in the text or sentence as shown in (10).

$$\textsc{IncSc} = \frac{1000}{N_t} \times N_c \tag{10}$$

We also compute the INCSC of *difficult* tokens, that is, tokens above a certain reference level, i.e. the level of an L2 learner writing a text or whom the text would be presented to as reading material. This value is also computed separately for nouns and verbs, since these are crucial when conveying meaning.

Moreover, we consider the INCSC of tokens not present in KELLY, i.e. out-of-vocabulary words (*OOV* INCSC) and the INCSC of tokens for which the lemmatizer could not identify a corresponding lemma in SALDO (*No lemma* INCSC). The relevance of OOV words for readability were suggested already in traditional formulas (Dale and Chall 1949). They can, among other, be uncommon simple or compound words, neologisms or they may indicate erroneous tokens, especially in texts written by L2 learners.

5.1.3 Morphological features

Morphological features include not only INCSC of different morpho-syntactic categories, but also variational scores, i.e. the ratio of a category to the ratio of *lexical* tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). The information is based on HunPos tagger (Halácsy, Kornai and Oravecz 2007).

A number of language specific features were included here, part of which was inspired by L2 learning materials (Fasth and Kannermark 1997), others by feature sets targeting L2 complexity (François and Fairon 2012) or readability (Heimann Mühlenbock 2013; Vajjala and Meurers 2012). We include, among others, the ratio of different verb forms to verbs which are typically introduced at varying stages of L2 learning. *S-verbs* (*S-VB*) are a special group of Swedish verbs ending in *-s* that indicate either reciprocity, a passive construction or are *deponent* verbs, i.e. verbs active in meaning, but passive in form. L2 grammar books usually dedicate some attention to these due to their peculiarity in terms of morphology and semantics (Fasth and Kannermark 1997). Neuter gender nouns are also considered since they can indicate the abstractness of a concept which is often a sign of higher linguistic complexity (Graesser et al. 2004). Among relative structures we count relative adverbs, determiners, pronouns and possessives. *Nominal ratio* (Hultman and Westman 1977) corresponds to the ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR), adverbs, and verbs. Its simplified version is the ratio of nouns to verbs, and it is meant to measure the information load of a text or to reveal its genre (e.g. spoken vs. news text). A higher value corresponds to higher degrees of complexity and a more elaborate genre.

Besides lexical categories, we incorporated also information about different function words. We compute IncSc for punctuation marks as well as sub- and conjunctions (SJ, CJ) whose presence in larger quantities can indicate a more complex syntactic structure. Particles also increase linguistic complexity since, similarly to English, when they form a particle verb, they can change the meaning of verbs considerably (Heimann Mühlenbock 2013). A higher density of pronouns indicates a more complex text (Graesser et al. 2004). The IncSc of the third person singular (3SG) pronoun inspired by Zhang, Liu and Ni (2013) is also included since this is often used referentially, which can further increase the difficulty of processing due to potential ambiguities in what is being referred to.

5.1.4 Syntactic features

Syntactic features were computed based on the output of the MaltParser (Nivre et al. 2007). An example output for the sentence *Stora delar av staden består av grönområden.* ‘Large parts of the city consist of green areas.’ is shown in figure 5.1. The abbreviations on the arcs are dependency relation tags and the ones above the words in squares are parts of speech. (See appendix B for a complete list of the tag sets and for the annotations in XML format).

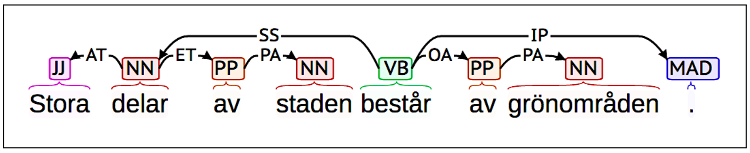


Figure 5.1: Example Sparv annotation with POS tags and dependency relations.

Syntactic aspects are related to readers’ working memory load when processing sentences which can be increased by ambiguity or embedded constituents

(Gibson 1998; Graesser et al. 2004). The CEFR document also points out that “particularly complex syntax consumes attentional resources that might otherwise be available for dealing with content” (Council of Europe 2001: 165). We compute the average length (depth) of dependency arcs (*DepArc*) and consider also their direction since right dependencies tend to be harder to process (Sjöholm 2012).

To measure embedded constituents, we consider relative clauses in clefts,[19] as well as pre- and post-modifiers (e.g. adjectives and prepositional phrases) and prepositional complements as in Heimann Mühlenbock (2013). Complex nominal phrases were found to correlate well, among others, with L2 English proficiency levels in learner texts (Lu 2011). Moreover, subordinates, commonly used in previous research on linguistic complexity (e.g. Heimann Mühlenbock 2013; Lu 2011; Schwarm and Ostendorf 2005) are also counted.

5.1.5 Semantic features

The amount of semantic features remains rather limited, since word-sense disambiguation has only been recently integrated into the Sparv pipeline. The two features included thus quantify available word senses per lemma based on the SALDO lexicon (cf. section 3.2.3). Both the average number of senses per token and the average number of noun senses per noun are considered. Polysemous words can be demanding for readers as they need to be disambiguated for a full understanding of the sentence (Graesser, McNamara and Kulikowich 2011).

5.1.6 Additional possible features

Besides the features outlined above, also the lack of culture-related knowledge can be a factor influencing L2 complexity. Moreover, unlike in L1 acquisition, in L2 learning a number of different learner variables also play role, which include, for example, learners’ age and their knowledge of other languages (Beinborn, Zesch and Gurevych 2012). We, however, do not address these dimensions in the current stage due to a lack of a sufficient amount of relevant data for modeling these.

[19] Sentences that begin with a constituent receiving particular focus, followed by a relative clause. E.g.: *It is John (whom) Jack is waiting for.*

5.2 Summary of the studies for learning material selection

In this section, we summarize our results on linguistic complexity analysis for receptive texts and sentences for the purposes of identifying learning material of a suitable difficulty for learners at different CEFR levels. Furthermore, we outline the main features of HitEx, a system for corpus example selection for language learning and summarize the outcome of its evaluation with users.

5.2.1 Receptive linguistic complexity analysis

The feature set described in section 5.1 has been employed for classifying receptive and productive L2 complexity at both the text and the sentence level. Table 5.2 summarizes our results for receptive L2 complexity reported in Pilán, Vajjala and Volodina (2016) and in chapter 7 in this thesis, where the WEKA implementation of a logistic regression classifier was used in a stratified 10-fold cross-validation setup. The size of the datasets was 867 instances for texts and 1874 for sentences, unevenly distributed across CEFR levels A1-C1. Both texts and individually occurring sentences were collected from the COCTAILL corpus (cf. section 3.1.3).

	Acc (%)		F_1		**RMSE**	
	Text	Sent	Text	Sent	Text	Sent
MAJORITY	33.2	40.2	0.17	0.23	0.52	0.49
LIX	34.9	41.4	0.22	0.3	0.38	0.38
ALL	**81.3**	63.4	**0.81**	0.63	0.27	0.31

Table 5.2: Classification results for receptive linguistic complexity.

As the results show, the traditional readability measure, LIX, used as a single feature cannot properly distinguish between texts belonging to different CEFR levels, its performance is approximately on par with a baseline of assigning the majority class label to all instances (MAJORITY). With the complete feature set (ALL), however, a significantly better performance can be achieved. Our system correctly classified 8 out of 10 texts and 6 out of 10 sentences, which is comparable both to previously reported results (cf. section 2.3.3) and to human annotators' performance, which was between 50% and 67% for exact CEFR-level agreement respectively for Swedish sentences and texts in other languages (cf. section 2.1 and section 9.4.3.1). Even though the number of instances was higher, classifying CEFR levels at the sentence level proved to be a harder task, potentially due to the availability of only a short linguistic context.

When studying the contribution of different feature groups, we found that, while lexical features are highly predictive at the text level, and can, alone, almost reproduce the results of the full feature set, at the sentence level, the combination of different feature groups was more effective. Moreover, by applying our sentence classifier to texts, we showed that texts at a certain CEFR level contain a substantial amount of sentences of other, typically lower levels (see figure 7.1).

5.2.2 HitEx: a corpus example selection system

Being able to automatically analyze linguistic complexity enables the identification of materials suitable for L2 learners at different levels. Since the re-use of whole texts might be precluded for reasons of copyright and given that L2 exercises are often sentence-based, analyzing complexity at the level of single sentences, is especially relevant in the L2 learning context. However, to make corpora useful for L2 learning purposes, besides complexity, there are a number of additional aspects that need to be taken into consideration.

In Pilán, Volodina and Borin (2017) – see chapter 9 – we proposed a framework for selecting candidate sentences for language learning exercises, called *HitEx* (*Hitta Exempel* 'Find Examples'). HitEx was inspired by previous work on exercise generation and dictionary example selection. This related work and the qualitative analysis of a small relevant set of data revealed that not only need the selected sentences be of an appropriate degree of complexity, but they should also be well-formed and independent from the surrounding context. We defined *well-formedness* in terms of having a dependency root, being non-elliptic (containing a subject and a finite verb), being complete (starting with a capital letter, ending with a sentence-final punctuation mark) and containing a low amount of non-lemmatized and non-alphabetical tokens. We detected *context-dependence* based on referring expressions of different kind such as pronominal and adverbial anaphora and those sentences containing structural connectives, where the first clause referred to remains outside of the sentence boundaries.

We complemented the aspects of well-formedness and context-dependence with a number of additional structural and lexical criteria. The additional *structural* properties that we proposed include, among others, whether the sentence is interrogative and whether it contains negative formulations, direct speech or modal verbs. *Lexical* aspects look at the presence, the frequency and the CEFR level of words in different lexical resources, including a list of sensitive, pedagogically potentially inappropriate words. Proper names and abbreviations are also counted here and the typicality of a sentence in terms of word co-occurrence can also be measured. For a comprehensive list of the sentence selection criteria proposed, see table 9.1. Such a comprehensive framework specifically tailored to finding exercise items in corpora was lacking in previous literature, as we argued also in the overview of previous approaches provided in section 2.4.2.

Since an explicit control and detailed information about the suitability of sentences may be desirable in certain application scenarios, and a sufficient amount of annotated data was not available, all HitEx criteria are rule-based, except for the aspect of L2 complexity. It is possible to either filter or rank sentences based on the criteria mentioned, all criteria having an equal weight in the latter case. The resulting subset of corpus sentences maximizes the fit to users' selection criteria and are arranged in a descending order of goodness by the implemented system.

5.2.3 A user evaluation of HitEx

To investigate the pedagogical usefulness and appropriateness of the HitEx sentence selection system, we carried out an evaluation with L2 learners and teachers. Teachers rated 330 sentences selected with HitEx from generic Swedish corpora on a four-point scale according to three aspects: linguistic complexity, context independence and overall suitability, the latter aiming at measuring the interplay among all criteria. They found that the system satisfied all three criteria to a large extent with average evaluation scores per criteria between 3.05 and 3.18. Interestingly, teachers agreed with both the systems and with each other on a CEFR level in only half of the cases. Almost all disagreements remained, however, within one CEFR level distance (cf. table 9.6). We measured also the inter-rater agreement in terms of Krippendorff's alpha among teachers, which was, on average, 0.62 across all CEFR levels. This would indicate that CEFR level assignment at the sentence level is a rather challenging task. Finally, both linguistic complexity and context independence positively correlated with overall suitability with a Spearman correlation of $\rho = 0.34$ and $\rho = 0.53$ respectively.

A subset of the teacher rated sentences were given to L2 learners between A1–B1 levels in the form of semi-automatically constructed *wordbank* exercises (see figure E.5 in the appendix for an example). Learners were required to identify for a set of five gapped sentences the missing word from a list of six candidate words containing a distractor, i.e. an option not fitting into any of the sentences. We found that the exercises on average matched well the ideal item difficulty (IID) score (cf. equation 13 in section 9.4.3.2) computed taking into consideration the probability of correct answers by chance. According to the IID score, 64.5% of students should have answered correctly the items on average. As table 5.3 shows, the actual item difficulty corresponding to the portion of student answering correctly in practice was 62%. A1 level items proved to be somewhat hard, while at A2 and B1 level sentences proved to be appropriately challenging, potentially even slightly easier than desired at B1 level (see also table 9.2).

	A1	**A2**	**B1**	**Average**	**Ideal**
Item Difficulty	0.505	0.665	0.695	**0.621**	**0.645**

Table 5.3: Item difficulty in the evaluation data with L2 learners.

5.3 Overview of the experiments on learner texts

Not only reading materials provided to L2 learners at different proficiency levels vary in linguistic complexity, but also the texts that learners themselves produce. We, therefore, investigate the usefulness of the features presented in section 5.1 for predicting CEFR levels also in essays written by L2 learners of Swedish. The collection of this type of data is, however, time consuming and poses, among others, problems of digitization and anonymization of sensitive personal information. Data sparsity thus is rather common in this domain.

Typically, L2 learners are able to cope with a higher degree of complexity than what they produce (Barrot 2015). Thus, even though both reading texts in coursebooks and essays can be divided into proficiency levels, the underlying distribution of the complexity feature values are likely to be somewhat different. A compelling research question arising is, then, whether these differences could be still small enough to allow for making generalizations across these two types of data, a question which has not been investigated in previous literature to our knowledge.

Since our reading text dataset was larger, we used that to improve CEFR level predictions for the smaller amount of learner-written texts available. We tested three different approaches: (i) training an essay classifier only on the coursebook data; (ii) combining coursebook texts with part of the essays during training; and (iii) using coursebook texts for estimating the values of the lexical complexity features.

One of the main differences between the two types of texts is the presence of errors in the learner productions such as spelling, grammatical errors (e.g. number or gender agreement between nouns and adjectives) or lexical-semantic ones (e.g. wrong words combined in multiword expressions). The second research question we investigated in this context is thus whether normalizing (at least part of) these errors could improve the cross-domain classification results. Since we did not have access to data annotated with L2 Swedish errors to train a supervised system tailored for this task, we opted for the open source spell-checker, LanguageTool[20] (Naber 2003). LanguageTool can identify not only spelling, but also a few types of grammatical errors such as gender disagreement, and it suggests a number of correction candidates. To find the candidate that is the closest possible to the word intended by a learner, we ranked these correction suggestions based on word co-occurrence information, considering as optimal candidate the one maximizing co-occurrence with neighboring words within the same sentence. For a more detailed presentation of this error correction approach, see section 10.5.2.

We summarize the results of our experiments on the classification of CEFR levels in learner essays in table 5.4, aggregating information from tables 10.4 and 11.3. Out of the domain adaptation setups and learning algorithms, we only include the two best performing ones. These are: (i) the +FEATURE setup where the predictions of a classifier trained on coursebook texts has been incorporated as an extra feature for a classifier trained on the essay data; and (ii) WEIGHTED, where instances both from coursebook texts and from essays were combined in the training set, the essays receiving a higher weight. (A more detailed description of the different domain adaptation setups tested is presented in section 10.5.1.) The experiments were carried out using an SVM classifier (Sequential minimal optimization from WEKA).

	ORIGINAL		NORMALIZED	
	F_1	κ^2	F_1	κ^2
IN-DOMAIN (web text list)	0.721	0.886	0.720	0.872
IN-DOMAIN (L2 list)	**0.808**	**0.922**	-	-
SOURCE-ONLY	0.438	0.713	0.620	0.807
+FEATURE	0.709	0.879	**0.802**	0.864
WEIGHTED	**0.747**	**0.890**	0.779	**0.915**

Table 5.4: Domain adaptation results for learner essay classification with and without error normalization.

[20] www.languagetool.org

We found that training on the limited amount of learner essays (IN-DOMAIN) achieved acceptable performance levels compared to the reported state-of-the-art results (cf. section 2.3.4) with $F_1 = 0.721$. Changing the original web text based list, KELLY (section 3.2.1), with an L2 list compiled using the coursebook texts, SVALex (section 3.2.2), for computing the lexical features yielded an additional performance gain of +0.08 F_1. Using only the coursebook texts as training data (SOURCE-ONLY) resulted in a notably lower performance, which underlines the assumption that linguistic complexity in these two types of texts within the same CEFR levels differs to a substantial extent. Correcting errors in the learner essays, however, improved domain adaptation results significantly, potentially because feature values could be more accurately estimated, and hence, the similarity between the two domains also became more evident. When combining coursebook data with some essay instances, the most successful approaches proved to be: (i) using as additional feature the predictions of a coursebook based classifier (+FEATURE) and (ii) a weighted combination of the two kinds of instances (WEIGHTED). (See sections 4.4 and 10.5.1 for additional details about these and other approaches). These, performed approximately at the same level as using an L2 list with only the limited amount of essays as training data. Morphological features transferred best in a SOURCE-ONLY setup and lexical features were among the most predictive ones both within- and cross-domain. We found, moreover, that a coursebook-based classifier shows a tendency to consider essays to be of a lower level than they are according to their annotations (64% of all misclassifications).

To summarize, we showed, that coursebook texts can be employed to classify CEFR levels in learner-written texts if used as additional training data, as sole training data, if L2 learner errors are normalized, or as a word list informing lexical features.

5.4 Investigating the importance of linguistic complexity features

Typically, out of the features proposed for a specific machine learning task some contribute more than others to a good prediction performance. Eliminating redundant features and retaining only the *k* number of best ones can result in simpler and improved models that are not only faster, but might also generalize better on unseen data (Witten et al. 2011: 308). In this section, we summarize our experiments from chapter 12 investigating the importance of individual features for predicting proficiency levels in different datasets. These include the texts-level datasets used in the experiments summarized in sections 5.2 and 5.3: the reading comprehension texts (TEXT-R) and the learner essays (TEXT-E) in their original form, without error-normalization. For the experiments at the sentence level, the dataset constructed from the HitEx evaluation data has been used, which we described in section 3.1.5.

These experiments were also based on the feature set described in section 5.1, but the values for the lexical features were computed using the two L2 lists, SVALex and SweLLex rather than KELLY (see chapter 3 for more details about these resources). SVALex was used for the learner essays and the sentence dataset (SENT). Since both the reading text data and SVALex are based on the COCTAILL corpus, the SweLLex list was used instead for TEXT-R. Information from the KELLY list has been preserved only in the feature based on average log frequencies.

We used as development set (DEV) 85% of each dataset with stratified 5-fold cross-validation. We evaluated the generalizability of the selected subset of features on the remaining 15% of the data (TEST). As learning algorithm for these feature selected models, we used an SVM classifier implemented in scikit-learn, namely *LinearSVC*.[21]

We employed a *univariate feature selection* method, also available in scikit-learn, to identify the most informative features scored with *analysis of variance* (ANOVA). ANOVA is a statistical test that can be used to measure how strong the relationship is between each feature and the output classes, CEFR levels in our case. It relies on *F-tests*, which can be used to score features based on significant differences in their per-class mean values. Our feature selection method identified fourteen features that were informative across all three datasets. These are presented in table 5.5.[22]

The count-based measure of square root TTR was one of the most predictive features across all three dataset. It seems thus that a varied way of expression, through e.g. the use of synonyms, is a good indicator of linguistic complexity in the L2 context. Moreover, the proportion of difficult lexica and the amount of tokens at the extremes of the CEFR scale (A1 and C1 levels) were also useful predictors. Both lexical variation and frequency were among the three best indicators of essay quality also for L2 English in Crossley and McNamara (2011). Lexical variation in terms of TTR as well as verb variation were also found highly predictive for L2 Estonian learner texts (Vajjala and Lõo 2014).

Out of our syntactic features, the ones relative to the length of dependency arcs and the INCSC of participles were among the *k*-best for all datasets. Participles are, in fact, typically introduced explicitly to L2 learners at higher

[21] We used scikit-learn for training our feature selected models since this package is implemented in the same programming langauge (Python) as the other functionalities of our online learning platform and was hence more convenient to integrate compared to WEKA, employed for the previously published experiment results.

[22] See section F.2 in the appendix for a more complete list of selected features and their ANOVA F-values.

Feature name	**Rank**		
	Sent	Text-R	Text-E
Square root TTR	2	7	9
A1 lemma IncSc	3	3	2
Punctuation IncSc	4	11	12
Relative clause	6	>24	8
Difficult N&V IncSc	7	1	1
Avg. DepArc length	8	10	14
Max length DepArc	9	6	13
Present PC to V	13	18	17
Past PC to V	14	>24	18
Particle IncSc	15	>24	16
V variation	16	15	10
Difficult W IncSc	17	2	4
C1 lemma IncSc	19	>24	5
N to V	21	>24	20

Table 5.5: The *k*-best features (and their rank) shared across all three datasets.

CEFR levels (Fasth and Kannermark 1997). The amount of punctuation marks and particles were also strong indicators of complexity. The former can, for example, signal clause boundaries and hence more complex sentences. Particles, on the other hand, can be challenging for language learners, since they alter the meaning of verbs (e.g. *hålla* 'hold', *hålla med* 'agree', *hålla på* 'be busy with something').

The results of the feature-selected models in terms of accuracy and F_1 are summarized in table 5.6.

Data	**Features**	**Sent**		**Text-R**		**Text-E**	
		Acc	F_1	Acc	F_1	Acc	F_1
Dev	All	0.62	0.61	0.68	0.68	0.73	0.71
Dev	*k*-best	0.73	0.71	0.70	0.70	0.81	0.81
Test	*k*-best	0.81	0.79	0.73	0.73	0.84	0.82
Number of *k*-best		21		54		24	

Table 5.6: Accuracy with feature selection across datasets.

Using the selected subset of (*k*-BEST) features yielded an improvement over the models trained with the complete set of features (ALL) for all three datasets. For TEXT-R the amount of *k*-best features selected was more than twice as many as the ones for the other two datasets. Nevertheless, the classification performance on this dataset with the selected features remained below the results for the other two datasets. The scores in table 5.6 for TEXT-R and TEXT-E with ALL show a drop in performance compared to the results reported in tables 5.2 (for Text) and 5.4 (for IN-DOMAIN with L2 list). This could be due to the 15% less training data which has been set aside for testing in the feature selection experiments and, possibly, because of the change in the algorithm used. Moreover, for TEXT-R, the small size of the SweLLex list could also explain the decrease in performance, therefore returning to the use of the KELLY list for this dataset might be preferable in the future.

For the small dataset of teacher-evaluated corpus examples, SENT, feature selection proved to be especially beneficial. The selected subset of features yielded an improvement of 0.11 percentage point of accuracy compared to using all features with the development set and also with respect to the dataset of coursebook sentences (see table 5.2).

Figures F.1 – F.3 in the appendix show how classification accuracy changes with the gradual increase of the number of features included in the model based on the *k*-best ranks from table 5.6. Confusion matrices for TEST are also available in section F.4.

5.5 Integration of research outcomes into an ICALL platform

We integrated the outcomes of the research described in this thesis, summarized in sections 5.1 – 5.3, into a freely available online platform, *Lärka*,[23] to ensure their accessibility to the general public. As mentioned in section 1.1, the implementation of the functionalities behind HitEx (section 5.5.2) and TextEval (section 5.5.3) described in this thesis constitute part of the engineering contribution of this thesis work, while the graphical user interface has been realized by others.

5.5.1 Lärka

Lärka (LÄR språket via KorpusAnalys 'Learn the language via Corpus Analysis') is a web-based ICALL platform that builds on corpora, lexical resources and NLP tools. This makes the platform a flexible and rich source of additional

[23] `https://spraakbanken.gu.se/larkalabb/`

L2 practice materials for Swedish. Thanks to Lärka's *service-oriented architecture* (SOA), most functionalities are also available as *web services*. Web services are "a software system designed to support interoperable machine-to-machine interaction over a network". [24] These modular services are independent from programming languages and platforms which ensures an easy re-use of different functionalities by other applications (Srinivasan and Treadwell 2005).

Initially, Lärka mainly aimed at offering automatically generated exercises for learners of L2 Swedish and students of Swedish linguistics. The original Lärka exercises were based on an earlier system called ITG from the *IT-based Collaborative Learning in Grammar* e-learning project (Borin and Saxena 2004). Later it has developed into a research infrastructure, which today also includes an L2 corpus annotation interface, as well as two modules based on this thesis work: the sentence selection module, HitEx, and the linguistic complexity analysis tool, TextEval. These, on the one hand, provide self-learning opportunities for L2 learners, and on the other, aid those teaching them either by supporting learner text evaluation or by facilitating the preparation of learning materials. This is beneficial as it reduces time and effort and it provides the opportunity to involve authentic language use in the language learning process.

The automatic generation of exercises is based on real-life language examples from corpora. Exercise generation is aimed at two groups of learners: students of (Swedish) linguistics and learners of Swedish as a second language (L2). The currently available exercises have a multiple-choice format. Each exercise item consists of a sentence containing either a highlighted word or a gap, as well as a list of some answer alternatives out of which one is the correct answer and the others are distractors. The integration of the updated version of HitEx (described in chapter 9) for the selection of appropriate corpus examples used to automatically generate exercises for language learners is currently in progress. (The earlier version of the system relied on Pilán, Volodina and Johansson (2014) for selecting sentences for these exercises.)

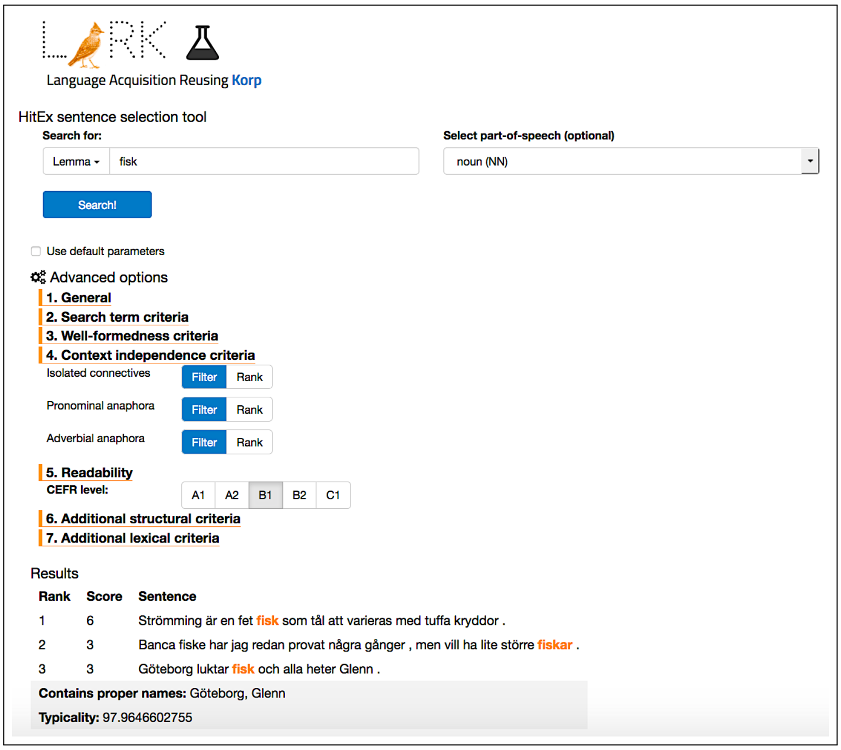
5.5.2 HitEx

As we described in section 5.2, the main purpose of HitEx[25] is to identify sentences from generic corpora which are suitable as exercise items for L2 learners. Figure 5.2, repeated here for the reader's convenience from the publication in chapter 9, illustrates the graphical user interface of the system. The suitability of the sentences is determined based on a number of criteria that reflect different

[24] `https://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice`

[25] `https://spraakbanken.gu.se/larkalabb/hitex`

linguistic characteristics of the sentences. Through the interface, it is possible to perform a customized sentence search according to these aspects. (For the complete set of advanced search options, see appendix E.1).



HitEx sentence selection tool

Search for: Lemma fisk

Select part-of-speech (optional): noun (NN)

□ Use default parameters

Advanced options

1. General

2. Search term criteria

3. Well-formedness criteria

4. Context independence criteria

Isolated connectives: Filter

Pronominal anaphora: Filter

Adverbial anaphora: Filter

5. Readability

CEFR level: B1

6. Additional structural criteria

7. Additional lexical criteria

Results

Rank	Score	Sentence
1	6	Strömming är en fet **fisk** som tål att varieras med tuffa kryddor .
2	3	Banca fiske har jag redan provat några gånger , men vill ha lite större **fiskar** .
3	3	Göteborg luktar **fisk** och alla heter Glenn .

Contains proper names: Göteborg, Glenn

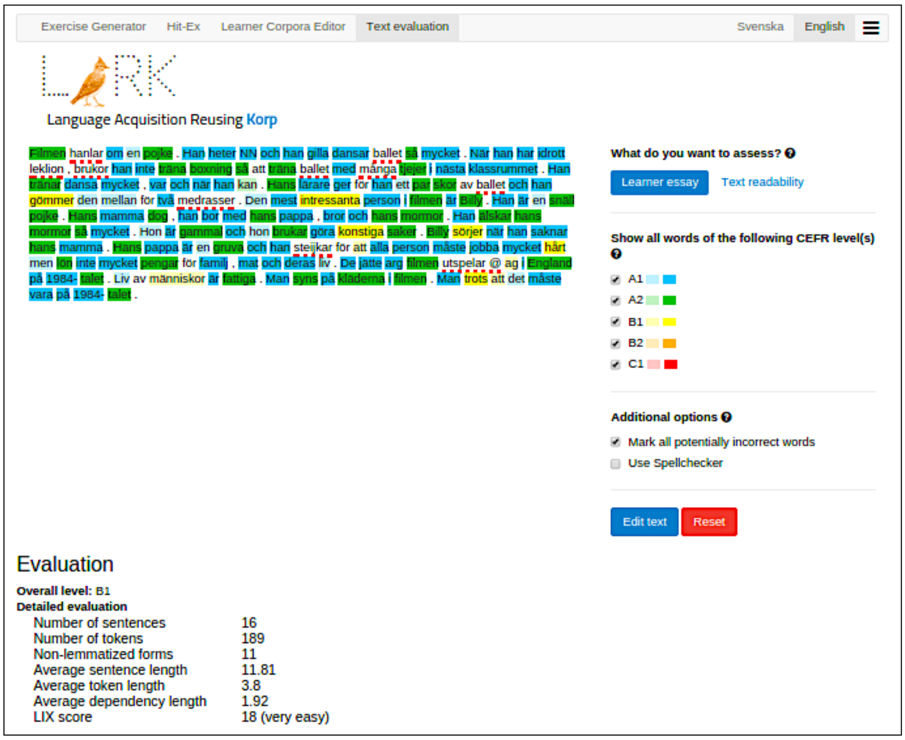
Typicality: 97.9646602755

Figure 5.2: The HitEx user interface with *fisk* ‘fish’ as search term.

One can search for inflected or uninflected forms of a word, but more complex queries (e.g. any plural noun preceded by an adjective or specific word senses) are also possible through the use of CQP (corpus query protocol) expressions. The result returned to the user consists of a list of sentences ranked based on how well they satisfy the specified search criteria. Detailed information about linguistic properties corresponding to the selection criteria is also available for each sentence (see figure 5.2). Besides its applicability to the language learning domain, HitEx can be also useful for lexicographers for finding dictionary examples that illustrate the meaning and usage of lexical items.

5.5.3 TextEval

The machine learning based linguistic complexity analysis both for receptive and productive L2 Swedish has been integrated into Lärka under the module *TextEval*.[26] This offers an interface where a Swedish text can be typed or pasted into a text box and it can be automatically analyzed for its degree of complexity. This interface is shown in figure 5.3 in this section, as well as in the article in which it has been originally published (cf. figure 11.1).



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 . Den mest intressanta person i filmen är Billy . Han är en snäll pojke . Hans mamma dog , han bor med hans pappa , bror och hans mormor . Han älskar hans mormor så mycket . Hon är gammal och hon brukar göra konstiga saker . Billy sörjer när han saknar hans mamma . Hans pappa är en gruva och han steijkar för att alla person måste jobba mycket hårt men lön inte mycket pengar för familj , mat och deras liv . De jätte arg filmen utspelar @ ag i England på 1984- talet . Liv av människor är fattiga . Man syns på kläderna i filmen . Man trots att det måste vara på 1984- talet .

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

Overall level: B1

Detailed evaluation

Number of sentences	16
Number of tokens	189
Non-lemmatized forms	11
Average sentence length	11.81
Average token length	3.8
Average dependency length	1.92
LIX score	18 (very easy)

Figure 5.3: The interface for linguistic complexity analysis

TextEval calls the backend of the Sparv annotation pipeline to perform lemmatization, POS tagging and syntactic parsing. Then, a machine learning based analysis using the methods described in the previous sections returns an overall CEFR level for the text. Given the somewhat 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. Furthermore, a number of linguistic indicators relevant for measuring text complexity are also computed. These include:

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

- text length (in number of sentences)
- token length (in number of characters)
- average sentence length
- average token length
- amount of non-lemmatized tokens
- average length of dependency arcs
- LIX score
- nominal ratio
- pronoun to noun ratio

In addition, it is possible to mark with *color-enhanced highlighting* the CEFR level of tokens, which provides users with a straightforward visual feedback about the lexical complexity of a text. This information is retrieved from the two L2 Swedish word lists, SVALex based on expert-produced texts and SweLLex representing learners' productive vocabulary. As described in section 3.2.2, the information about CEFR levels in these lists originates from frequency distributions in L2 texts. For each CEFR level, a darker and a lighter shade of the same color represent productive and receptive vocabulary respectively at a given level. TextEval can be useful either for learners who would like to evaluate their own texts or for teachers and coursebook writers who need to assess the suitability of L2 learning material candidates in terms of their linguistic complexity.

5.6 Limitations

There are a number of limitations to the work presented in this thesis which we would like to summarize in this section.

To start with, the size of the L2 Swedish datasets is relatively small, which may effect to some extent the generalizability of the machine learning based results presented and the subset of predictive features identified. It is also worth noting that the CEFR levels for the reading comprehension texts have been derived from the CEFR levels of the lessons in which they occurred in, but no distinction was made between texts appearing in the first or the last lessons of a certain CEFR level. Texts *within* a CEFR level might also differ somewhat in terms of complexity, and not retaining information about the order of occurrence within a level might contribute to the dataset being harder to classify. Experiments using regression and information about the order of texts within levels might yield better results for this dataset.

Moreover, learner-related variables (e.g. age, other languages known) are important factors determining to what extent individual learners can handle different degrees of linguistic complexity. These aspects, however, are not currently incorporated in our models due to an insufficient amount of relevant data.

When it comes to the analysis of learner texts containing L2 errors, NLP tools trained on standard language are likely to tend towards making an increased amount of tagging errors. This affects to some degree the estimation of feature values in the machine learning experiments relying on this data.

As for the lexical resources used, since the corpora from where frequencies were estimated for the KELLY list consist of materials from the web, these might not fully reflect the type of linguistic content that L2 learners are typically exposed to. Therefore, the suggested CEFR levels for some lemmas might not always correspond to those in L2 curricula. The two other word lists employed, SVALex and SweLLex are based on L2 data, however, their size remains rather limited. This means not only that, when looking up words during text analysis, many may appear out-of-vocabulary, but the information at the basis of the mappings to CEFR levels (the significant onset of use, cf. section 3.2.2) may also be less representative in some cases. The suggested CEFR levels in all three lists would require additional empirical and pedagogical validation.

Concerning the selection of sentences from corpora, the set of criteria proposed for HitEx is not an exhaustive one. The need for additional criteria may arise depending on the specific end use, e.g. different exercise formats. Rather than equal weights, giving a different degree of importance to the criteria may improve the quality of the sentence selection. These weights could be estimated once data for the target tasks is available.

Finally, additional, larger scale usability studies would need to be conducted to establish the usefulness of TextEval for text analysis and HitEx for automatic exercise generation and example sentence selection in real-life pedagogical settings.

6 Conclusion

6.1 Summary

In this thesis, we have offered an in-depth, data-driven analysis of aspects influencing linguistic complexity in the context of Swedish as a second language. We explored machine learning methods for the classification of CEFR levels using linguistic complexity features both at the text and at the considerable less-researched, smaller unit of sentences. We proposed a CEFR classification system for Swedish that compares well to both human performance and previous work on other languages. Moreover, we proposed a sentence selection framework specifically for identifying sentences in corpora that are suitable as exercise items, which previously has been a somewhat neglected building block in the process of completely automatizing exercise generation.

Furthermore, we have investigated the relationship between the degree of linguistic complexity learners can deal with and produce at different CEFR levels by transferring a classifier trained on reading comprehension texts to the task of classifying learner-written essays. Knowledge transfer in a machine learning context between these types of texts has not been explored previously. Our results indicated that these two text types show some difference in terms of linguistic complexity, but we, nevertheless, identified a number of ways in which combining these data sources can lead to an improved classification performance. Such transfer learning techniques are valuable especially for less resourced languages since they reduce the need for annotated data. Moreover, they can offer interesting pedagogical insights into the relationship between productive and receptive L2 skills.

We also pinpointed a subset of linguistic complexity features that were among the best predictors of proficiency levels across productive and receptive texts and sentences. These include information about lexical frequency and variation as well as morpho-syntactic dimensions such as the amount and type of participles and the length of syntactic dependencies.

6.2 Future directions

There are various directions in which the work presented in this thesis could be extended and improved. The experiments could be repeated with larger datasets to confirm their generalizability once additional L2 data is available. The ongoing SweLL research infrastructure project (Volodina et al. 2016a) aims at collecting and annotating approximately 600 essays written by adult learners of L2 Swedish between 2017 – 2019. Learner errors are also planned to be annotated and categorized in the resource. This would open the way to investigating further the importance of error features for linguistic complexity in L2 productive texts and the effect of error-correction on the domain transfer between these learner texts and coursebook texts. Additional manually annotated data would be especially beneficial for continuing the exploration of applying CEFR levels to sentences, which appeared to be a hard task both for humans and automated systems.

Besides experiments with larger amount of learner text data, additional transfer learning methods mentioned in Pan and Yang (2010) could also be investigated. Exploring this task for other languages could provide valuable insights into whether there are approaches and features that generalize well in this context.

The feature set proposed could be extended in a number of ways. For example, a list of cognate words could be easily incorporated into the system if such a lexical resource was available. The relevance of cognates for linguistic complexity has been pointed out in e.g. Beinborn, Zesch and Gurevych (2014b). Moreover, annotation for word senses has now been made available in the Sparv annotation pipeline, which would enable research on additional semantic features based on resources such as SALDO and the Swedish FrameNet (Borin et al. 2010).

Data-driven approaches to the selection of candidates for exercise items could also be explored. The COCTAILL corpus contains a number of exercise items, most of which, however, would require additional annotation to be correctly analyzable by NLP tools and to be usable as dataset instances since they contain gaps, shuffled word order etc. Furthermore, a more intuitive and easy-to-interpret scoring for the selected sentences could also be explored. Finally, the usefulness of HitEx could be investigated through logging and observing learners' results on exercises automatically generated based on the selected sentences.

6.3 Significance

This thesis work encompassed different stages of research and development of NLP applications which ranged from the corpus preparation phase, through the extraction of relevant features to the implementation of web applications and their evaluation with the intended user groups. We have actively worked on making our research results available not only to research communities, but also to the general public in the form of prototype online tools. These, besides showcasing the potentials of NLP for language learning may aid L2 Swedish teachers and learners in their daily practices.

The research carried out in this thesis can assist teaching professionals in locating high quality learning materials in larger quantities at an increased speed. With the help of our text evaluation system, they can also receive indications about linguistic complexity in learner-written texts. Moreover, the machine learning based linguistic complexity analysis method proposed may contribute to more objective applications of the CEFR scale.

Apart from teachers, L2 learners can also benefit from the research results presented as these create self-learning opportunities by laying the ground for automatic activity generation and holistic feedback on learner texts. Providing technology-enhanced opportunities for autonomous learning is particularly valuable in an era of international mobility and migration coupled with the widespread use of digital devices.

Furthermore, the availability of web services offers the possibility to those building ICALL applications to easily integrate our algorithms in their own systems. Although our implementation is based on Swedish NLP resources, most linguistic aspects considered are applicable to a wide variety of languages and could be easily transfered to systems targeting those.

To conclude, in this thesis, we presented concrete steps within research and development focusing on language learning applications of NLP that are aware of proficiency levels, aiming thus at increasing their usefulness in the language teaching practice.

Part II

Studies on learning material selection

7 Linguistic complexity for texts and sentences

This chapter is a postprint version of the following publication:

Pilán, Ildikó, Sowmya Vajjala and Elena Volodina 2016. A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications (IJLCA) 7 (1): 143–159.*

Abstract. Corpora and web texts can become a rich language learning resource if we have a means of assessing whether they are linguistically appropriate for learners at a given proficiency level. In this paper, we aim at addressing this issue by presenting the first approach for predicting linguistic complexity for Swedish second language learning material on a 5-point scale. After showing that the traditional Swedish readability measure, Läsbarhetsindex (LIX), is not suitable for this task, we propose a supervised machine learning model, based on a range of linguistic features, that can reliably classify texts according to their difficulty level. Our model obtained an accuracy of 81.3% and an F-score of 0.8, which is comparable to the state of the art in English and is considerably higher than previously reported results for other languages. We further studied the utility of our features with single sentences instead of full texts since sentences are a common linguistic unit in language learning exercises. We trained a separate model on sentence-level data with five classes, which yielded 63.4% accuracy. Although this is lower than the document level performance, we achieved an adjacent accuracy of 92%. Furthermore, we found that using a combination of different features, compared to using lexical features alone, resulted in 7% improvement in classification accuracy at the sentence level, whereas at the document level, lexical features were more dominant. Our models are intended for use in a freely accessible web-based language learning platform for the automatic generation of exercises.

7.1 Introduction

Linguistic information provided by natural language processing (NLP) tools has good potential for turning the continuously growing amount of digital text into interactive and personalized language learning material. Our work aims at overcoming one of the fundamental obstacles in this domain of research, namely how to assess the linguistic complexity of texts and sentences from the perspective of second and foreign language (L2) learners.

There are a number of readability models relying on NLP tools to predict the difficulty (readability) level of a text (Collins-Thompson and Callan 2004; Schwarm and Ostendorf 2005; Graesser, McNamara and Kulikowich 2011; Vajjala and Meurers 2012; Heimann Mühlenbock 2013; Collins-Thompson 2014). The linguistic features explored so far for this task incorporate information, among others, from part-of-speech (POS) taggers and dependency parsers. Cognitively motivated features have also been proposed, for example, in the Coh-Metrix (Graesser, McNamara and Kulikowich 2011). Although the majority of previous work focuses primarily on document-level analysis, a finer-grained, sentence-level readability has received increasing interest in recent years (Vajjala and Meurers 2014; Dell'Orletta et al. 2014; Pilán, Volodina and Johansson 2014).

The previously mentioned studies target mainly native language (L1) readers including people with low literacy levels or mild cognitive disabilities. Our focus, however, is on building a model for predicting the proficiency level of texts and sentences used in L2 teaching materials. This aspect has been explored for English (Heilman et al. 2007; Huang et al. 2011; Zhang, Liu and Ni 2013; Salesky and Shen 2014), French (François and Fairon 2012), Portuguese (Branco et al. 2014) and, without the use of NLP, for Dutch (Velleman and van der Geest 2014).

Readability for the Swedish language has a rather long tradition. One of the most popular, easy-to-compute formulas is LIX (*Läsbarthetsindex*, 'Readability index') proposed in Björnsson (1968). This measure combines the average number of words per sentence in the text with the percentage of long words, i.e. tokens consisting of more than six characters. Besides traditional formulas, supervised machine learning approaches have also been tested. Swedish document-level readability with a native speaker focus is described in Heimann Mühlenbock (2013) and Falkenjack, Heimann Mühlenbock and Jönsson (2013). For L2 Swedish, only a binary sentence-level model exists (Pilán, Volodina and Johansson 2014), but comprehensive and highly accurate document- and sentence-level models for multiple proficiency levels have not been developed before.

In this paper, we present a machine learning model trained on course books currently in use in L2 Swedish classrooms. Our goal was to predict linguistic complexity of material written by teachers and course book writers for learners, rather than assessing learner-produced texts. We adopted the scale from the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) which contains guidelines for the creation of teaching material and the assessment of L2 proficiency. CEFR proposes six levels of language proficiency: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Since sentences are a common unit in language exercises, but remain less explored in the readability literature, we also investigate the applicability of our approach to sentences, performing a 5-way classification (levels A1-C1). Our document-level model achieves a state-of-the-art performance (F-score of 0.8), however, there is room for improvement in sentence-level predictions. We plan to make our results available through the online intelligent computer-assisted language learning platform Lärka[27], both as corpus-based exercises for teachers and learners of L2 Swedish and as web-services for researchers and developers.

In the following sections, we first describe our datasets (section 7.2) and features (section 7.3), then we present the details and the results of our experiments in section 7.4. Finally, section 7.5 concludes our work and outlines further directions of research within this area.

7.2 Datasets

Our dataset is a subset of COCTAILL, a corpus of course books covering five CEFR levels (A1-C1) (Volodina et al. 2014b). This corpus consists of twelve books (from four different publishers) whose usability and level have been confirmed by Swedish L2 teachers. The course books have been annotated both content-wise (e.g. exercises, lists) and linguistically (e.g. with POS and dependency tags) (Volodina et al. 2014b). We collected a total of 867 texts (reading passages) from this corpus. We excluded texts that are primarily based on dialogues from the current experiments due to their specific linguistic structure, with the aim of scaling down differences connected to text genres rather than linguistic complexity. We plan to study the readability of dialogues and compare them to non-dialogue texts in the future.

Besides reading passages, i.e. texts, the COCTAILL corpus contains a number of sentences independent from each other, i.e. not forming a coherent text, in the form of lists of sentences and *language examples*. This latter category consists of sentences illustrating the use of specific grammatical patterns or lexical items. Collecting these sentences, we built a sentence-level dataset consisting of 1874 instances. The information encoded in the content-level annotation of

[27] http://spraakbanken.gu.se/larka/

COCTAILL (XML tags *list*, *language_example* and the attribute *unit*) enabled us to include only complete sentences and exclude sentences containing gaps and units larger or smaller than a sentence (e.g. texts, phrases, single words etc.). The CEFR level of both sentences and texts has been derived from the CEFR level of the lesson (chapter) they appeared in. In table 7.1, columns 2-5 give an overview of the distribution of texts across levels and their mean length in sentences.[28] The distribution of sentences per level is presented in the last two columns of table 7.1. COCTAILL contained a somewhat more limited amount of B2 and C1 level sentences in the form of lists and language examples, possibly because learners handle larger linguistic units with more ease at higher proficiency levels.

	Document level				**Sentence level**	
CEFR	**Books**	**Publ.**	**Texts**	**Mean nr. sent**	**Books**	**Sentences**
A1	4	3	49	14.0	4	505
A2	4	3	157	13.8	4	754
B1	5	3	258	17.9	4	408
B2	4	3	288	26.6	3	124
C1	2	2	115	42.1	1	83
Total	12	4	**867**	-	4	**1874**

Table 7.1: The distribution of items per CEFR level in the datasets.

7.3 Features

We developed our features based on information both from previous literature (Heilman et al. 2007; Vajjala and Meurers 2012; François and Fairon 2012; Heimann Mühlenbock 2013; Pilán, Volodina and Johansson 2014) and a grammar book for Swedish L2 learners (Fasth and Kannermark 1997). The set of features can be divided in the following five subgroups: length-based, lexical, morphological, syntactic and semantic features (table 7.2).

Length-based (LEN): These features include sentence length in number of tokens (#1) and characters (#4), extra-long words (longer than thirteen characters) and the traditional Swedish readability formula, LIX (see section 7.1). For the sentence-level analysis, instead of the ratio of number of tokens to

[28] The number of different books and publishers is reported per each level, some books spanning more levels.

the number of sentences in the text, we considered the number of tokens in one sentence.

Lexical (LEX): Similar to Pilán, Volodina and Johansson (2014), we used information from the Kelly list (Volodina and Kokkinakis 2012), a lexical resource providing a CEFR level and frequencies per lemma based on a corpus of web texts. Thus, this word list is entirely independent from our dataset. Instead of percentages, we used *incidence scores* (INCSC) per 1000 words to reduce the influence of sentence length on feature values. The INCSC of a category was computed as 1000 divided by the number of tokens in the text or sentence multiplied by the count of the category in the sentence. We calculated the INCSC of words belonging to each CEFR level (#6 - #11). In features #12 and #13 we considered *difficult* all tokens whose level was above the CEFR level of the text or sentence. We computed also the INCSC of tokens not present in the Kelly list (#14), tokens for which the lemmatizer did not find a corresponding lemma form (# 15), as well as average log frequencies (#16).

Morphological (MORPH): We included the variation (the ratio of a category to the ratio of lexical tokens - i.e. nouns, verbs, adjectives and adverbs) and the INCSC of all lexical categories together with the INCSC of punctuations, particles, sub- and conjunctions (#34, #51). Some additional features, using insights from L2 teaching material (Fasth and Kannermark 1997), captured fine-grained inflectional information such as the INCSC of neuter gender nouns and the ratio of different verb forms to all verbs (#52 - #56). Instead of simple type-token ratio (TTR) we used a bilogarithmic and a square root TTR as in Vajjala and Meurers (2012). Moreover, nominal ratio (Heimann Mühlenbock 2013), the ratio of pronouns to prepositions (François and Fairon 2012), and two *lexical density* features were also included: the ratio of lexical words to all non-lexical categories (#48) and to all tokens (#49). Relative structures (#57) consisted of relative adverbs, determiners, pronouns and possessives.

Syntactic (SYNT): Some of these features were based on the length (depth) and the direction of dependency arcs[29] (#17 - #21). We complemented this, among others, with the INCSC of relative clauses in clefts[30] (#26), and the INCSC of pre-and postmodifiers (e.g. adjectives and prepositional phrases) (Heimann Mühlenbock 2013).

Semantic (SEM): Features based on information from SALDO (Borin, Forsberg and Lönngren 2013), a Swedish lexical-semantic resource. We used the average number of senses per token as in Pilán, Volodina and Johansson (2014) and included also the average number of noun senses per nouns. Once reliable

[29] The tags were obtained with the MaltParser (Nivre et al. 2007).

[30] Sentences that begin with a constituent receiving particular focus, followed by a relative clause. E.g.: It is John (whom) Jack is waiting for.

Nr.	Feature Name	Nr.	Feature Name
	Length-based		*Morphological*
1	Sentence length	30	Modal verbs to verbs
2	Average token length	31	Particle IncSc
3	Extra-long words	32	3SG pronoun IncSc
4	Number of characters	33	Punctuation IncSc
5	LIX	34	Subjunction IncSc
	Lexical	35	S-verb IncSc
6	A1 lemma IncSc	36	S-verbs to verbs
7	A2 lemma IncSc	37	Adjective IncSc
8	B1 lemma IncSc	38	Adjective variation
9	B2 lemma IncSc	39	Adverb IncSc
10	C1 lemma IncSc	40	Adverb variation
11	C2 lemma IncSc	41	Noun IncSc
12	Difficult word IncSc	42	Noun variation
13	Difficult noun and verb IncSc	43	Verb IncSc
14	Out-of-Kelly IncSc	44	Verb variation
15	Missing lemma form IncSc	45	Nominal ratio
16	Avg. Kelly log frequency	46	Nouns to verbs
	Syntactic	47	Function word IncSc
17	Average dependency length	48	Lexical to non-lexical words
18	Dependency arcs longer than 5	49	Lexical words to all tokens
19	Longest dependency from root node	50	Neuter gender noun IncSc
20	Ratio of right dependency arcs	51	Con- and subjunction IncSc
21	Ratio of left dependency arcs	52	Past participles to verbs
22	Modifier variation	53	Present participles to verbs
23	Pre-modifier IncSc	54	Past verbs to verbs
24	Post-modifier IncSc	55	Present verbs to verbs
25	Subordinate IncSc	56	Supine verbs to verbs
26	Relative clause IncSc	57	Relative structure IncSc
27	Prepositional complement IncSc	58	Bilog type-token ratio
	Semantic	59	Square root type-token ratio
28	Avg. nr. of senses per token	60	Pronouns to nouns
29	Noun senses per noun	61	Pronouns to prepositions

Table 7.2: The complete feature set.

word-sense disambiguation methods become available for Swedish, additional features based on word senses could be taken into consideration here.

The complete set of 61 features is presented in table 7.2. Throughout this paper we will refer to the machine learning models using this set of features, unless otherwise specified. Features for both document- and sentence-level analyses were extracted for each sentence, the values being averaged over all sentences in the text in the document-level experiments to ensure comparability.

7.4 Experiments and results

7.4.1 Experimental setup

We explored different classification algorithms for this task using the machine learning toolkit WEKA (Hall et al. 2009). These included: (1) a multinomial logistic regression model with ridge estimator, (2) a multilayer perceptron, (3) a support vector machine learner, Sequential Minimal Optimization (SMO), and (4) a decision tree (J48). For each of these, the default parameter settings have been used as implemented in WEKA.

We considered classification accuracy, F-score and Root Mean Squared Error (RMSE) as evaluation measures for our approach. We also included a confusion matrix, as we deal with a dataset that is unbalanced across CEFR levels. The scores were obtained by performing a ten-fold Cross-Validation (CV).

7.4.2 Document-level experiments

We trained document-level classification models, comparing the performance between different subgroups of features. We had two baselines: a majority classifier (MAJORITY), with B2 as majority class, and the LIX readability score. Table 7.3 shows the type of subgroup (*Type*), the number of features (*Nr*) and three evaluation metrics using logistic regression.

Type	Nr	Acc (%)	F	RMSE
MAJORITY	-	33.2	0.17	0.52
LIX	1	34.9	0.22	0.38
LEX	11	**80.3**	**0.80**	0.24
ALL	61	**81.3**	**0.81**	0.27

Table 7.3: Document-level classification results.

Not only was accuracy very low with LIX, but this measure also classified 91.6% of the instances as B2 level. Length-based, semantic and syntactic features in isolation showed similar or only slightly better performance than the baselines, therefore we excluded them from table 7.3. Lexical features, however, had a strong discriminatory power without an increase in bias towards the majority classes. Using this subset of features only, we achieved approximately the same performance (0.8 F) as with the complete set of features, ALL (0.81 F). This suggests that lexical information alone can successfully distinguish the CEFR level of course book texts at the document level. Using the complete feature set we obtained 81% accuracy and 97% *adjacent accuracy* (when misclassifications to adjacent classes are considered correct). The same scores with lexical features (LEX) only were 80.3% (accuracy) and 98% (adjacent accuracy).

Accuracy scores using other learning algorithms were significantly lower (see table 7.4), therefore, we report only the results of the logistic regression classifier in the subsequent sections.

Type	**Nr**	**Perceptron**	**SMO**	**J48**
LEX	11	**77.4**	42.1	55
ALL	61	62.2	52.7	50.5

Table 7.4: Accuracy scores (in %) for other learning algorithms.

Instead of classification, some readability studies (e.g. Huang et al. (2011); Branco et al. (2014)) employed linear regression for this task. For a better comparability, we applied also a linear regression model to our data which yielded a correlation of 0.8 and an RMSE of 0.65.

To make sure that our system was not biased towards the majority classes B1 and B2, we inspected the confusion matrix (table 7.5) after classification using ALL. We can observe from table 7.5 that the system performs better at A1 and C1 levels, where confusion occurred only with adjacent classes. Similar to the findings in (François and Fairon 2012) for French, classes in the middle of the scale were harder to distinguish. Most misclassifications in our material occurred at A2 level (23%) followed by B1 and B2 level, (20% and 17% respectively).

To establish the external validity of our approach, we tested it on a subset of LÄSBART (Heimann Mühlenbock 2013), a corpus of Swedish easy-to-read (ETR) texts previously employed for Swedish L1 readability studies (Heimann Mühlenbock 2013; Falkenjack, Heimann Mühlenbock and Jönsson 2013). We used 18 fiction texts written for children between ages nine to twelve, half of which belonged to the ETR category and the rest were unsimplified. Our model generalized well to unseen data, it classified all ETR texts as B1 and all ordinary texts as C1 level, thus correctly identifying in all cases the relative difference in complexity between the documents of the two categories.

Predictions						
A1	**A2**	**B1**	**B2**	**C1**		
37	12	0	0	0	**A1**	L
12	121	18	5	1	**A2**	a
4	11	206	24	13	**B1**	b
0	5	21	238	24	**B2**	e
0	0	0	12	103	**C1**	l

Table 7.5: Confusion matrix for feature set ALL at document level.

Although a direct comparison with other studies is difficult because of the target language, the nature of the datasets and the number of classes used, in terms of absolute numbers, our model achieves comparable performance with the state-of-the-art systems for English (Heilman et al. 2007; Salesky and Shen 2014). Other studies for non-English languages using CEFR levels include: François and Fairon (2012), reporting 49.1% accuracy for a French system distinguishing six classes; and Branco et al. (2014) achieving 29.7% accuracy on a smaller Portuguese dataset with five levels.

7.4.3 Sentence-level experiments

After building good classification models at document level, we explored the usability of our approach at the sentence level. Sentences are particularly useful in Computer-Assisted Language Learning (CALL) applications, among others, for generating sentence-based multiple choice exercises, e.g. Volodina et al. (2014a), or vocabulary examples (Segler 2007). Furthermore, multi-class readability classification of sentence-level material intended for second language learners has not been previously investigated in the literature.

With the same methodology (section 7.4.1) and feature set (section 7.3) used at the document level, we trained and tested classification models based on the sentence-level data (see section 7.2). The results are shown in table 7.6.

Although the majority baseline in the case of sentences was 7% higher than the one for texts (table 7.3), the classification accuracy for sentences using all features was only 63.4%. This is a considerable drop (-18%) in performance compared to the document level (81.3% accuracy). It is possible that the features did not capture differences between the sentences because the amount of context is more limited on the fine-grained level. It is interesting to note that, although

Type	**Nr**	**Acc (%)**	**F**	**RMSE**
Majority	-	40.2	0.23	0.49
LIX	1	41.4	0.3	0.38
Lex	11	56.8	0.53	0.33
All	61	**63.4**	**0.63**	0.31

Table 7.6: Sentence-level classification results.

there was no substantial performance difference between Lex and All at a document level, the model with all the features performed 7% better at sentence level.

Most misclassifications occurred, however, within a distance of one class only, thus the adjacent accuracy of the sentence-level model was still high, 92% (see table 7.7). Predictions were noticeably more accurate for classes A1, A2 and B1 which had a larger number of instances.

Predictions						
A1	**A2**	**B1**	**B2**	**C1**		
371	123	9	2	0	**A1**	L
120	541	78	11	4	**A2**	a
27	136	212	23	10	**B1**	b
8	34	39	30	13	**B2**	e
0	18	21	9	35	**C1**	l

Table 7.7: Confusion matrix for feature set All at sentence level.

In the next step, we applied the sentence-level model on the document-level data to explore how homogeneous texts were in terms of the CEFR level of the sentences they contained. Figure 7.1 shows that texts at each CEFR level contain a substantial amount of sentences of the same level of the whole text, but they also include sentences classified as belonging to other CEFR levels.

Finally, as in the case of the document-level analysis, we tested our sentence-level model also on an independent dataset (SenRead), a small corpus of sentences with gold-standard CEFR annotation. This data was created during a user-based evaluation study (Pilán, Volodina and Johansson 2013) and it consists of 196 sentences from generic corpora, i.e. originally not L2 learner-focused corpora, rated as being suitable at B1 or being at a level higher than B1.

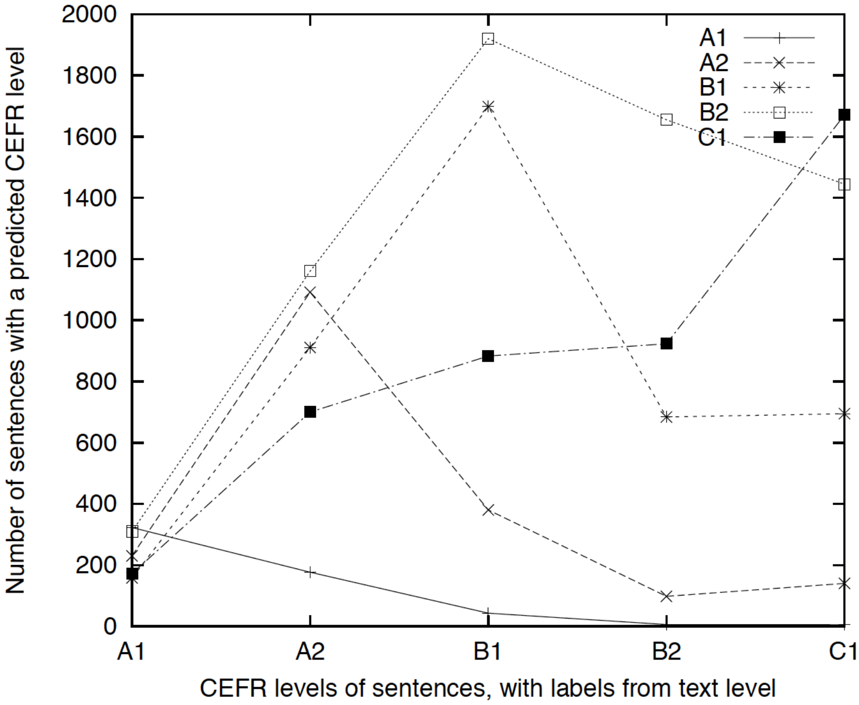


Figure 7.1: Distribution of sentences per CEFR level in the document-level data.

We used this corpus along with the judgments of the three participating teachers. Since SENREAD had only two categories - $<= B1$ and $> B1$, we combined the model's predictions into two classes - A1, A2, B1 were considered as <=B1 and B2, C1 were considered as >B1. The majority baseline for the dataset was 65%, <=B1 being the class with most instances. The model trained on COCTAILL sentences predicted with 73% accuracy teachers' judgments, an 8% improvement over the majority baseline. There was a considerable difference between the precision score of the two classes, which was 85.4% for <=B1, and only 48.5% for >B1.

Previously published results on sentence-level data include Vajjala and Meurers (2014), who report 66% accuracy for a binary classification task for English and Dell'Orletta et al. (2014) who obtained an accuracy between 78.9% and 83.7% for Italian binary class data using different kinds of datasets. Neither of these studies, however, had a non-native speaker focus. Pilán, Volodina and Johansson (2014) report 71% accuracy for Swedish binary sentence-level classification from an L2 point of view. Both the adjacent accuracy of our sentence-level model (92%) and the accuracy score obtained with that model on SENREAD (73%) improve on that score. It is also worth mentioning that the labels in the dataset from Pilán, Volodina and Johansson (2014) were based on the assumption that all sentences in a text belong to the same difficulty level which, being an approximation, introduced some noise in that data.

Although more analysis would be needed to refine the sentence-level model, our current results indicate that a rich feature set that considers multiple linguistic dimensions may result in an improved performance. In the future, the dataset could be expanded with more gold-standard sentences, which may improve accuracy. Furthermore, an interesting direction to pursue would be to verify whether providing finer-grained readability judgments is a more challenging task also for human raters.

7.5 Conclusion and future work

We proposed an approach to assess the proficiency (CEFR) level of Swedish L2 course book texts based on a variety of features. Our document-level model, the first for L2 Swedish, achieved an F-score of 0.8, hence, it can reliably distinguish between proficiency levels. Compared to the wide-spread readability measure for Swedish, LIX, we achieved a substantial gain in terms of both accuracy and F-score (46% and 0.6 higher respectively). The accuracy of the sentence-level model remained lower than that of the text-level model, nevertheless, using the complete feature set the system performed 23% and 22% above the majority baseline and LIX respectively. Misclassifications of more than one level did not occur in more than 8% of sentences, thus, in terms of adjacent accuracy, our sentence-level model improved on previous results for L2 Swedish readability.

Most notably, we have found that taking into consideration multiple linguistic dimensions when assessing linguistic complexity is especially useful for sentence-level analysis. In our experiments, using only word-frequency features was almost as predictive as a combination of all features for the document level, but the latter made more accurate predictions for sentences, resulting in a 7% difference in accuracy. Besides L2 course book materials, we tested both our document- and sentence-level models also on unseen data with promising results.

In the future, a more detailed investigation is needed to understand the performance drop between document and sentence level. Acquiring more sentence-level annotated data and exploring new features relying on lexical-semantic resources for Swedish would be interesting directions to pursue. Furthermore, we intend to test the utility of this approach in a real-world web application involving language learners and teachers.

8 DETECTING CONTEXT DEPENDENCE IN CORPUS EXAMPLES

This chapter is a postprint version of the following publication:

Pilán, Ildikó 2016. Detecting Context Dependence in Exercise Item Candidates Selected from Corpora. In *Proceedings of the* 11^{th} *Workshop on Innovative Use of NLP for Building Educational Applications*, 151–161.

Abstract. We explore the factors influencing the dependence of single sentences on their larger textual context in order to automatically identify candidate sentences for language learning exercises from corpora which are presentable in isolation. An in-depth investigation of this question has not been previously carried out. Understanding this aspect can contribute to a more efficient selection of candidate sentences which, besides reducing the time required for item writing, can also ensure a higher degree of variability and authenticity. We present a set of relevant aspects collected based on the qualitative analysis of a smaller set of context-dependent corpus example sentences. Furthermore, we implemented a rule-based algorithm using these criteria which achieved an average precision of 0.76 for the identification of different issues related to context dependence. The method has also been evaluated empirically where 80% of the sentences in which our system did not detect context-dependent elements were also considered context-independent by human raters.

8.1 Introduction

Extracting single sentences from corpora with the use of natural language processing (NLP) tools can be useful for a number of purposes including the detection of candidate sentences for automatic exercise generation. Such sentences are also known as *seed sentences* (Sumita, Sugaya and Yamamoto 2005) or *carrier sentences* (Smith, Avinesh and Kilgarriff 2010) in the Intelligent Computer-Assisted Language Learning (ICALL) literature. Interest for the use of corpora in language learning has arisen already in the 1980s, since the increasing amount of digital text available enables learning through authentic language use (O'Keeffe, McCarthy and Carter 2007). However, since sentences in a text form a coherent discourse, it might be the case that for the interpretation of the meaning of certain expressions in a sentence, previously mentioned information, i.e. a *context*, is required (Poesio, Ponzetto and Versley 2011). Corpus sentences whose meaning is hard to interpret are less optimal to be used as exercise items (Kilgarriff et al. 2008), however, having access to a larger linguistic context is not possible due to copy-right issues sometimes (Volodina, Johansson and Johansson Kokkinakis 2012).

In the following, we explore how we can automatically assess whether a sentence previously belonging to a text can also be used as a stand-alone sentence based on the linguistic information it contains. We consider a sentence *context-dependent* if it is not meaningful in isolation due to: (i) the presence of expressions referring to textual content that is external to the sentence, or (ii) the absence of one or more elements which could only be inferred from the surrounding sentences.

Understanding the main factors giving rise to context dependence can improve the trade-off between discarding (or penalizing) sub-optimal candidates and maximizing the variety of examples and thus, their authenticity. Such a system may not only facilitate teaching professionals' work, but it can also aid the NLP community in a number of ways, e.g. evaluating automatic single-sentence summaries, detecting ill-formed sentences in machine translation output or identifying dictionary examples.

Although context dependence has been taken into consideration to some extent in previous work, we offer an in-depth investigation of this research problem. The theoretical contribution of our work is a set of criteria relevant for assessing context dependence of single sentences based on a qualitative analysis of human evaluators' comments. This is complemented with a practical contribution in the form of a rule-based system implemented using the proposed criteria which can reliably categorize corpus examples based on context dependence both when evaluated using relevant datasets and according to human raters' judgments. The current implementation of the system has been tested on Swedish data, but the criteria can be easily applied to other languages as well.

8.2 Background

8.2.1 Corpus examples combined with NLP for language learning

In a language learning scenario, corpus example sentences can be useful both as exercise items and as vocabulary examples. Previous work on exercise item generation has adopted different strategies for carrier sentence selection. In some cases, sentences are mainly required to contain a lexical item or a linguistic pattern that constitutes the target of the exercise, but context dependence is not explicitly addressed (Sumita, Sugaya and Yamamoto 2005; Arregik 2011). Another alternative has been using dictionary examples as carrier sentences, e.g. from WordNet (Pino and Eskenazi 2009). Such sentences are inherently context-independent, however, they pose some limitations on the linguistic aspects to target in the exercises. In Pilán, Volodina and Johansson (2014) we presented and compared two algorithms for carrier sentence selection for Swedish, using both rule-based and machine learning methods. Context dependence, which had not been specifically targeted in that phase, emerged as a key issue for sub-optimal candidate sentences during an empirical evaluation.

Identifying corpus examples for illustrating lexical items is the main purpose of the GDEX (Good Dictionary Examples) algorithm (Husák 2010; Kilgarriff et al. 2008) which has also inspired a Swedish algorithm for sentence selection (Volodina, Johansson and Johansson Kokkinakis 2012). GDEX incorporates a number of linguistic criteria (e.g. sentence length, vocabulary frequency) based on which example candidates are ranked. Some of these are related to context dependence (e.g. incompleteness of sentences, presence of personal pronouns), but they are somewhat coarser-grained criteria not focusing on syntactic aspects. A system using GDEX for carrier sentence selection is described in Smith, Avinesh and Kilgarriff (2010) who underline the importance of the well-formedness of a sentence and who determine a sufficient amount of context in terms of sentence length. Segler (2007) focuses on vocabulary example identification for language learners. Teachers' sentence selection criteria has been modeled with logistic regression, the main dimensions examined being syntactic complexity and similarity between the original context of a word and an example sentence.

8.2.2 Linguistic aspects influencing context dependence

The relationship between sentences in a text can be expressed either explicitly or implicitly, i.e. with or without specific linguistic elements requiring extra-sentential information (Mitkov 2014). The explicit forms include words and phrases that imply structural discourse relations or are anaphoric (Webber et al.

2003). In a text, the way sentences are interconnected can convey an additional relational meaning besides the one which we can infer from the content of each sentence separately. Examples of such elements include *structural connectives*: conjunctions, subjunctions and “paired” conjunctions (Webber et al. 2003).

Another form of reference to previously mentioned information is *anaphora*. The phenomenon of anaphora consists of a word or phrase (*anaphor*) referring back to a previously mentioned entity (*antecedent*). Mitkov (2014) outlines a number of different anaphora categories based on their form and location, the most common being pronominal anaphora which has also been the focus of recent research within NLP (Poesio, Ponzetto and Versley 2011; Ng 2010; Nilsson 2010). A number of resources available today have noun phrase coreference annotation, such as the dataset from the SemEval-2010 Task (Recasens et al. 2010) and SUC-CORE for Swedish (Nilsson Björkenstam 2013).

Besides the anaphora categories described in Mitkov (2014), Webber et al. (2003) argue that adverbial connectives (*discourse connectives*), e.g. *istället* ‘instead’, also behave anaphorically, among others because they function more similarly to anaphoric pronouns than to structural connectives. A valuable resource for developing automatic methods for handling discourse relations is the Penn Discourse Treebank (Prasad et al. 2008) containing annotations for both implicit and explicit discourse connectives. Using this resource Pitler and Nenkova (2009) present an approach based on syntactic features for distinguishing between discourse and non-discourse usage of explicit discourse connectives (e.g. *once* as a temporal connective corresponding to "as soon as" vs. the adverb meaning "formerly"). Another phenomenon connected to context dependence is *gapping* where the second mention of a linguistic element is omitted from a sentence (Poesio, Ponzetto and Versley 2011).

8.3 Datasets

Instead of creating a corpus specifically tailored for this task with gold standard labels assigned by human annotators, which can be a rather time- and resource-intensive endeavor, we explored how different types of existing data sources which contained inherently context-(in)dependent sentences could be used for our purposes.

Language learning coursebooks contain not only texts, but also single sentences in the form of exercise items, lists and language examples illustrating a lexical or a grammatical pattern. We collected sentences belonging to these two latter categories from COCTAILL (Volodina et al. 2014b), a corpus of coursebooks for learners of Swedish as a second language. Most exercises contained gaps which might have misled the automatic linguistic annotation, therefore they have not been included in our dataset.

Dictionaries contain example sentences illustrating the meaning and the usage of an entry. One of the characteristics of such sentences is the absence of referring expressions which would require a larger context to be understood (Kilgarriff et al. 2008), therefore they can be considered suitable representatives of context-independent sentences. We collected instances of good dictionary example sentences from two Swedish lexical resources: SALDO (Borin, Forsberg and Lönngren 2013) and the Swedish FrameNet (SweFN) (Heppin and Gronostaj 2012). These sentences were manually selected by lexicographers from a variety of corpora.

Sentences explicitly considered dependent on a larger context are less available due to their lack of usefulness in most application scenarios. Two previous evaluations of corpus example selection for Swedish are described in Volodina, Johansson and Johansson Kokkinakis (2012) and Pilán, Volodina and Johansson (2013), we will refer to these as EVAL1 and EVAL2 respectively. In the former case, evaluators including both lexicographers and language teachers had to provide a score for the appropriateness of about 1800 corpus examples on a three-point scale. In EVAL2, about 200 corpus examples selected with two different approaches were rated by a similar group of experts based on their understandability (readability) for language learners, as well as their appropriateness as exercise items and as good dictionary examples. The data from both evaluations contained human raters' comments explicitly mentioning that certain sentences were context-dependent. We gathered these instances to create a negative sample. Since comments were optional, and context dependence was not the focus of these evaluations, the amount of sentences collected remained rather small, 92 in total. It is worth noting that this data contains spontaneously occurring mentions based on raters' intuition, rather than being labeled following a description of the phenomenon of context dependence as it would be customary in an annotation task.

The sentences from all data sources mentioned above constituted our development set. The amount of sentences per data source is presented in table 8.1, where CIND indicates positive, i.e. context-independent samples, and CDEP the negative, context-dependent ones. The suffix -LL stands for sentences collected from language learning materials while -D represents dictionary examples.

Source	Code	Nr. sent	Total
COCTAILL	CIND-LL	1739	
SALDO	CIND-D	4305	**8729**
SweFN	CIND-D	2685	
EVAL1	CDEP	22	
EVAL2	CDEP	70	**92**

Table 8.1: Number of sentences per source.

8.4 Methodology

As the first step in developing the algorithm, we aimed at understanding the presence or absence of which linguistic elements make sentences dependent on a larger context by analyzing our negative sample. Although the number of instances in the context-independent category was considerably higher, certain linguistic characteristics of such sentences could have been connected to aspects not relevant to our task. Negative sentences on the other hand, although modest in number, were explicit examples of the target phenomenon. Information about the cultural context may also be relevant for this task, however, we only concentrated on linguistic factors which can be effectively captured with NLP tools.

We aimed at covering a wide range of potential application scenarios, therefore we developed a method that was independent of: (i) information from surrounding sentences and (ii) the exact intended use for the selected sentences. The first choice was motivated by the fact that, even though most previous related methods (see section 8.2.2) rely on information from neighboring sentences as well, sometimes a larger context might not be available either due to the nature of the task (e.g. output of single-sentence summarization systems) or copy-right issues. Secondly, for a more generalizable approach, we aimed at assessing sentences based on whether their information content can be treated as an autonomous unit rather than according to whether they provide the appropriate amount and type of context to, for example, be solved as exercise items of a certain type. This way the method could serve as a generic basis to be tailored to specific applications which may pose additional requirements on the sentences.

Being that the amount of negative samples was rather restricted, we opted for the qualitative method of *thematic analysis* (Boyatzis 1998; Braun and Clarke 2006) aiming at discovering *themes*, i.e. categories, in our negative sample. Once we collected a set of context-dependent sentences, we started coding our data, in other words, manually labeling the instances with *codes*, a word or a phrase shortly describing the type of element that inhibited the interpretation of the sentence in isolation (for some examples see table 8.2 on the next page). In the subsequent phases, we grouped together codes into themes, i.e. broader categories, according to their thematic similarity in a mixed deductive-inductive fashion. We started out with an initial pool of themes inspired by phenomena proposed in previous literature relevant to context dependence. Some of the codes, however, could not be placed in any of these themes. For part of these we have found a theme candidate in the literature after the pattern emerged during the code grouping phase. In other cases, in absence of an existing category matching some instances of the CDEP data, we created our own theme labels.

Besides thematic analysis, we carried out also a quantitative analysis based on the distribution of part of speech tags in both our positive and negative sample in order to identify potential differences that could support and complement the information emerged in the themes.

In the following step, we implemented a rule-based algorithm for handling context dependence using the findings from the qualitative and quantitative analyses. Since most emerged aspects could be translated into rather easily detectable linguistic clues, and a sufficiently large dataset annotated with the different context-dependent phenomena was not available for Swedish, we opted for a heuristic-based system. We applied the algorithm and observed its performance on our development data. Our primary focus was on evaluating how precisely are context-dependent elements identified in CDEP, but we complemented this also with observing the percentage of false positives for context dependence in our positive sample.

Finally, in order to test candidate selection empirically, a new set of sentences has been retrieved from different corpora. These sentences were then first given to our system for assessment, then the subset of candidates not containing context dependent elements were given to evaluators for an external validation.

Theme	ID	Nr	Example code	Example CDep sentence
Incomplete sentence	IncompSent	12	incorrect sent. tokenization	*” **p**iper hon och alla skrattar .* ‘” **s**he whines and everyone laughs.’
Implicit anaphora	ImpAnaphora	11	omitted verb	*Till jul skulle hon **[X]**.* ‘For Christmas she should have **[X]**.’
Pronominal anaphora	PNAnaphora	23	pronoun as subject	*Eller också sitter **den** i taket.* ‘Or **it** sits on the roof.’
Adverbial anaphora 1 (Temporal and locative)	AdvAnaphora1	12	locative adverb	***Då** ska folk kunna lämna området .* ‘**Then** people can leave the area.’
Adverbial anaphora 2 (Discourse connectives)	AdvAnaphora2	22	adv. anaphora	*Vissa gånger sover hon inte **heller**.* ‘Sometimes she does not sleep **either**.’
Structural connectives	StructConn	17	coordinating conjunction	***Men** de pratade inte på samma ställe.* ‘**But** they did not talk at the same place.’
Answers to closed ended questions	CEQAnswer	11	yes/no answer	***Ja,** men det är ju jul.* ‘**Yes,** but it is of course Christmas.’
Context-depend properties of concepts	CDPC	8	unusual noun-noun comb.	*Du lämnar **planen**, **tolvan**!* ‘You leave the **field**, **twelve**!’

Table 8.2: Thematic analysis results.

8.5 Data analysis results

8.5.1 Qualitative results based on thematic analysis

The list of themes collected during our qualitative analysis is presented in table 8.2. For each theme, we provide an identifier (*ID*), the number of occurrence in the CDEP dataset (*Nr*)[31] together with an example code and an example sentence.[32]

The total number of codes emerged from the data was 22, which we mapped to 8 themes. Some of the themes were related to the categories mentioned in previous literature which we described in section 8.2. These included pronominal anaphora (Mitkov 2014), adverbial anaphora (Webber et al. 2003), connectives (Miltsakaki et al. 2004). Incomplete sentences (Didakowski, Lemnitzer and Geyken 2012) contained incorrectly tokenized sentences, titles and headings. Moreover, we distinguished three themes among different anaphoric expressions: pronominal anaphora, adverbial anaphora (with temporal and locative adverbs) and discourse connectives, i.e. adverbials expressing logical relations. Under the implicit anaphora theme we grouped different forms of gapping.

Two themes that emerged from the data during the thematic analysis were answers to closed ended questions and context-dependent properties of concepts. In the case of the former category, answers were mostly of the yes/no type. As for the latter theme, our data showed that the unexpectedness of the context of a word (especially if this is short, such as a sentence) can also play a role in whether a sentence is interpretable in isolation. Previous literature (Barsalou 1982) defines this phenomenon as "context-dependent properties of concepts". While the "core meanings" of words are activated "independent of contextual relevance", context-dependent properties are "only activated by relevant contexts in which the word appears" (Barsalou 1982: p. 82). In (11) we provide an example of both context-independent and context-dependent properties of the noun *tak* 'roof', from the EVAL2 data.

(11) (a) *Troligen berodde olyckan på all snö som låg på taket.*
'The accident probably depended on all the snow that covered the roof.'

(b) *Fler än hundra levande kunde dras fram under taket .*
'More than a hundred [people] were pulled out from under the roof alive.'

[31] Occasionally sentences included more than one theme.

[32] Tokens relevant to each theme are in bold and [X] indicates the position of an omitted element.

Sentence (11b) was considered context-dependent by human raters, while (11a) was not. Being covered in snow (11a) appears a more easily interpretable property of roof without a larger context than having something being pulled out from under it. The context that activates the context-dependent property of roof in (11b) is that the roof had collapsed, which, however, is missing from the sentence.

Finally, for 7 sentences in our CDEP data, no clear elements causing context dependence could have been clearly identified, these are omitted from table 8.2, but they have been preserved in the experiments.

8.5.2 Quantitative comparison of positive and negative samples

Besides carrying out a thematic analysis, we compared our positive and negative samples also based on quantitative linguistic information in search of additional evidence for the emerged themes and to detect further aspects that could be potentially worth targeting. Overall part of speech (POS) frequency counts showed some major differences between the CDEP and CINDEP sentences. There was a tendency towards a nominal content in context-independent sentences, where 21.6% of all POS tags were nouns. However, this value was 9% lower for context-dependent sentences, which would suggest a preference for a higher density of concepts in context-independent sentences. Pronouns, on the other hand, were more frequent in context-dependent sentences (12.6% in total) than in context-independent ones (7% less frequent).

The qualitative analysis revealed that elements responsible for context dependence commonly occurred at the beginning of the sentence. Therefore, we compared the percentage of POS categories for this position in the two groups of sentences. Context-independent sentences showed a strong tendency towards having a noun in sentence-initial position, almost one fourth of the sentences fit into this category. On the other hand, only 3% of the positive examples started with a conjunction, but 16% of context-depend items belonged to this group.

8.6 An algorithm for the assessment of context dependence

Inspired by the results of the thematic analysis and the quantitative comparison described above, we implemented a heuristics-based system for the automatic detection of context dependence in single sentences. For retrieving example sentences the system uses the concordancing API of Korp (Borin, Forsberg and Roxendal 2012), a corpus-query system giving access to a large amount of Swedish corpora. All corpora were annotated for different linguistic aspects including POS tags and dependency relation tags which served as a basis for the implementation. The system scores each sentence based on the amount of phenomena detected that match an implemented context dependence theme. Users can decide whether to *filter*, i.e. discard sentences that contain any element indicating context dependence. Alternatively, sentences can be *ranked* according to the amount of context-dependent issues detected: sentences without any such elements are ranked highest, followed by instances minimizing these aspects. All themes have an equal weight of 1 when computing the final ranking score, except for pronominal anaphora in which case, if pronouns have antecedent candidates, the weight is reduced to 0.5. In what follows, we provide a detailed description of the implementation of the themes listed in table 8.2.

Incomplete sentence. To detect incomplete sentences the algorithm scans instances for the presence of an identified dependency root, the absence of which is considered to cause context dependence. Moreover, orthographic clues denoting sentence beginning and end are inspected. Sentence beginnings are checked for the presence of a capital letter optionally preceded by a parenthesis, quotation mark or a dash, frequent in dialogues. Sentences beginning with a digit are also permitted. Sentence end is checked for the presence of major sentence delimiters (e.g. period, exclamation mark).

Implicit anaphora. Candidate sentences are checked for gapping, in other words, omitted elements. Our system categorizes as gapped (elliptic) a sentence which either lacks a finite verb or a subject. Finite verbs are all verbs that are not infinite, supine or participle. Modal verbs are considered finite in case they form a verb group with another verb. Subjects include also logical subjects, and in the case of a verb in imperative mode, no subject is required.

Explicit pronominal anaphora. We considered in this category the third person singular pronouns *den* ‘it’ (common gender) and *det* ‘it’ (neuter gender) as well as demonstrative pronouns (e.g. *denna* ‘this’, *sådan* ‘such’ etc.). We did not include here the animate third person pronouns *han* ‘he’ and *hon* ‘she’ since corpus-based evidence suggests that these are often used in isolated sentences in coursebooks (Scherrer and Lindemalm 2007) as well as in conversation (Mitkov 2014). Similarly to the English pronoun *it*, the Swedish equivalent *det* can also be used non-anaphorically in expositions, clefts and expressions describing a local situation, such as time and weather (Holmes and Hinchliffe 2003; Li et al. 2009; Gundel, Hedberg and Zacharski 2005) as the examples in (12) show.

(12) (a) *det* with weather-related verbs
Det regnar.
'It is raining.'

(b) Cleft
Det är sommaren (som) jag älskar.
'It is the summer (that) I like.'

(c) Exposition
Det är viktigt att du kommer.
'It is important that you come.'

Our system treats as non-anaphoric the pronoun *det* if it is expletive (pleonastic) syntactically according to the output of the dependency parser which covers expositions and clefts. To handle cases like (2a), weather-related verbs have been collected from lexical resources. The list currently comprises 14 items. First, verbs related to the class *Weather* in the Simple+ lexicon (Kokkinakis, Toporowska-Gronostaj and Warmenius 2000) have been collected. Then for each of these, the child nodes from the SALDO lexicon have been added. Finally, the list has been complemented with a few manual additions.

For potentially anaphoric pronouns, the system tries to identify antecedent candidates in a similar way to the robust pronoun resolution algorithm proposed in Mitkov (1998). We count proper names and nouns occurring with the same gender and number to the left of the anaphora. This is complemented with an infinitive marker headed by a verb as potential candidate for *det*. Since certain types of information useful for antecedent disambiguation were not available through our annotation pipeline or lexical resources for Swedish (e.g. gender for named entities, animacy), the final step for scoring and choosing candidates is not applied in this initial version of the algorithm. Lastly, pronouns followed by a relative clause introduced by *som* 'which' were considered non-anaphoric.

Explicit adverbial anaphora. Adverbs emerged as an undesirable category during both EVAL1 and EVAL2. However, a deeper analysis of our development data revealed that not all adverbs have equal weight when determining the suitability of a sentence. Some are more anaphoric then others. We collected a list of anaphoric adverbs based on Teleman, Hellberg and Andersson (1999). Certain time and place adverbials, also referred to as demonstrative pronominal adverbs (Webber et al. 2003) are used anaphorically (e.g. *där* 'there', *då* 'then'). Sentences containing these adverbs are considered context-independent only when: (i) they are the head of an adverbial of the same type that further specifies them, e.g. *där på landet* 'there on the countryside'; (ii) they appear with a determiner, which in Swedish builds up a demonstrative pronoun, e.g. *det där* *huset* ‘that house’.

Discourse connectives. Discourse connectives, i.e. adverbs expressing logical relations, fall usually into the syntactic category of conjunctional adverbials in the dependency parser output. Several conjunctional adverbials appear in the context-dependent sentences from EVAL1 and EVAL2. Our system categorizes a sentence containing a conjuctional adverb context-independent when a sentence contains: (i) at least 2 coordinate clauses; (ii) coordination or subordination at the same dependency depth or a level higher, that is, a sibling node that is either a conjunction or a subjunction.

Structural connectives. Sentences with conjunctions as dependency roots are considered context-dependent unless they are paired conjunctions with both elements included (e.g. *antingen ... eller* ‘either ... or’). Conjunctions in sentence initial position are also treated as an indication of context dependence except when there are at least two clauses or conjuncts in the sentence.

Answers to closed ended questions. To identify sentences that are answers to closed ended questions, the algorithm tries to match POS-tag patterns of sentence-initial interjections (e.g. *ja* ‘yes’, *nej* ‘no’) and adverbs surrounded by minor delimiters (e.g. dash), the initial delimiter being optional in the case of interjections.

Context-dependent properties of concepts. Apart from the theme implementations described above, we are currently investigating the usefulness of word co-occurrence information for this theme. The corpus query tool Korp for instance offers an API providing mutual information scores. The intuition behind this idea is that the frequency of words appearing together is positively correlated with the unexpectedness of the association between them.

8.7 Performance on the datasets

We evaluated our system both on the hand-coded negative example sentences collected from EVAL1 and EVAL2 (CDEP) and the positive samples comprised of the good dictionary examples (CINDEP-D) and the coursebook sentences (CINDEP-LL). The performance when predicting different aspects of context dependence is presented in table 8.3.

We focused on maximizing precision, i.e. on correctly identifying as many themes as possible in the hand-coded CDEP sentences, recall values were of lower importance since we aimed at avoiding every context-dependent sentence

Theme	Precision	Recall	F1
INCOMPSENT	0.75	0.5	0.6
IMPANAPHORA	0.33	0.36	0.35
PNANAPHORA	0.75	0.78	0.77
ADVANAPHORA1	0.91	0.83	0.87
ADVANAPHORA2	0.87	0.59	0.70
STRUCTCONN	0.7	0.82	0.76
CEQANSWER	1.0	0.55	0.71
Average	**0.76**	0.63	0.60

Table 8.3: Theme prediction performance in CDEP sentences.

rather than retrieving them all. Most themes were correctly identified, all themes except one was predicted with a precision of at least 0.7 and above. The only theme that yielded a lower result was that of implicit anaphoras. The error analysis revealed that these cases were mostly connected to an incorrect dependency parse of the sentences, mainly subjects tagged as objects in sentences with an inverted (predicate-subject) word order.

As mentioned previously, we strived for minimizing sub-optimal sentences in terms of context dependence, while trying to avoid being excessively selective to maintain a varied set of examples. To assess performance with respect to this latter aspect, we inspected also the percentage of sentences identified as context-dependent in dictionary examples (CIND-D) and coursebook sentences (CIND-LL). The percentage of predicted themes per dataset is shown in table 8.4 where *Total* stands for the percentage of sentences with at least one predicted theme.

Theme	CIND-D	CIND-LL
INCOMPSENT	2.37	3.39
IMPANAPHORA	4.61	5.80
PNANAPHORA	9.39	11.0
ADVANAPHORA1	3.59	2.93
ADVANAPHORA2	9.95	3.74
STRUCTCONN	3.70	0.92
CEQANSWER	0.37	2.59
Total	**33.35**	**26.74**

Table 8.4: Percentage of sentences with a predicted theme in the CIND datasets.

We can observe that even though all sentences are expected to be context-independent, our system labeled as context-dependent about three out of ten good dictionary examples and coursebook sentences. The error analysis revealed that some of these instances did indeed contain context-dependent elements, e.g. the conjunction *men* ‘but’ in sentence-initial position. In CIND-LL in the case of some sentences containing anaphoric pronouns an image provided the missing context in the coursebook, thus not all predicted cases were actual false positives, but rather, they indicated some noise in the data. As for dictionary examples, the presence of such sentences may also suggest that the criterion of context dependence can vary somewhat depending on the type of lexicon or lexicographers’ individual decisions.

Some sentences exhibited more than one phenomenon connected to context dependence. Multiple themes were predicted in 30.43% of the CDEP sentences, but only 6.54% and 7.25 of the CIND-D and CIND-LL sentences respectively.

8.8 User-based evaluation results

The algorithm was tested also empirically during an evaluation of automatic candidate sentence selection for the purposes of learning Swedish as a second language. The evaluation data consisted of 338[33] sentences retrieved from a variety of modern Swedish corpora and classified as not containing context dependence themes according to our algorithm (with the exception of 4 control sentences that were context-dependent). These were all unseen sentences not present in the datasets described in section 8.3. In the evaluation setup, all implemented themes were used as filters, i.e. sentences containing any recognized element connected to context dependence, described in section 8.6, were discarded. Besides context dependence, the evaluated system incorporated also other selection criteria (e.g. readability), but for reasons of relevance and space these aspects and the associated results are not discussed here.

The selected sentences were given for evaluation to 5 language teachers who assessed the suitability of these sentences based on 3 criteria: (i) their degree of being independent of context, (ii) their CEFR[34] level and (iii) their overall suitability for language learners. Teachers were required to assess this latter aspect without a specific exercise type in mind, but considering a learner reading the sentence instead. Sentences were divided into two subsets, each being rated by at least 2 evaluators. Teachers had to assign a score between 1 to

[33] We excluded 8 sentences with incomplete evaluator scores during the calculation of the results.

[34] The Common European Framework of Reference for Languages (CEFR) is a scale describing proficiency levels for second language learning (Council of Europe 2001).

4 to each sentence according to the scale definition in table 8.5.

The sentence...	
1	*... doesn't satisfy the criterion.*
2	*... satisfies the criterion to a smaller extent.*
3	*... satisfies the criterion to a larger extent.*
4	*... satisfies the criterion entirely.*

Table 8.5: Evaluation scale.

The results were promising, the average score over all evaluators and sentences for context independence was 3.05, and for overall suitability 3.23. For context-independence, 61% of the sentences received score 3 or 4 (completely satisfying the criterion) from at least half of the evaluators, and 80% of the sentences received an average score higher than 2.5. This latter improves significantly on the percentage of context-dependent sentences that we reported previously in Pilán, Volodina and Johansson (2013), where about 36% of all selected sentences were explicitly considered context-dependent by evaluators.

Furthermore, we computed the Spearman correlation coefficient for teachers' scores of overall suitability and context dependence to gain insight into how strongly associated these two aspects were according to our evaluation data. The correlation over all sentences was ρ=0.53, which indicates that not being context-dependent is positively associated with overall suitability. Therefore, context dependence is worth targeting when selecting carrier sentences. Additional details about this evaluation are described in Chapter 9.

8.9 Conclusion and future work

We described a number of criteria that influence context dependence in corpus examples when presented in isolation. Based on the thematic analysis of a set of context-dependent sentences, we implemented a rule-based algorithm for the automatic assessment of this aspect which has been evaluated not only on our datasets but also with the help of language teachers with very positive results.

About 76% of themes were correctly identified in context-dependent sentences, while the amount of false positives in the context-independent data was maintained rather low. Approximately 80% of candidate sentences selected with a system incorporating the presented algorithm were deemed context-independent in our user-based evaluation. The results also showed a positive correlation between sentences being context-independent and overall suitable for language learners.

In the future, we are planning to explore the extension of the algorithm to other languages as well as to experiment with machine learning approaches for this task using, among others, the resources mentioned in this paper.

9 CANDIDATE SENTENCE SELECTION FOR LANGUAGE LEARNING EXERCISES

This chapter is a postprint version of the following publication:

Pilán, Ildikó, Elena Volodina and Lars Borin 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues (TAL) Journal, Special issue on NLP for learning and teaching* 57 (3): 67–91.

Abstract. We present a framework and its implementation relying on natural language processing methods, which aims at the identification of exercise item candidates from corpora. The hybrid system combining heuristics and machine learning methods includes a number of relevant selection criteria. We focus on two fundamental aspects: linguistic complexity and the dependence of the extracted sentences on their original context. Previous work on exercise generation addressed these two criteria only to a limited extent, and a refined overall candidate sentence selection framework appears also to be lacking. In addition to a detailed description of the system, we present the results of an empirical evaluation conducted with language teachers and learners which indicate the usefulness of the system for educational purposes. We have integrated our system into a freely available online learning platform.

9.1 Introduction

Several tasks related to foreign and second language (L2) learning can be partly or entirely automatized with the help of natural language processing (NLP) tools. One such task is exercise generation, whose automation offers both self-directed learning opportunities and support for teaching professionals' practice. The pedagogical relevance and practical usefulness of such solutions, however, would need to be further improved before these systems can become widely used in language instruction. During our work, we aimed at maintaining a pedagogical angle, on the one hand, by incorporating statistical information from existing hand-written teaching materials into our selection criteria and, on the other hand, by evaluating the performance of our system with L2 teachers and learners.

Practice plays an important role in L2 learning for the development of both receptive and productive skills (DeKeyser 2007). Corpora as potential practice material are readily available in large quantities, however, their use in L2 teaching has been both supported and opposed in previous years, O'Keeffe, McCarthy and Carter (2007) present an overview of this debate. Corpora offer a large amount of diverse examples at a low cost, and their use has been shown to have a positive effect on learners' progress (Cobb 1997; Cresswell 2007). Moreover, corpora are evidence of real-life language use which, however, might be hard for learners to process (Kilgarriff 2009). Non-authentic, teacher-constructed materials have also been subject to criticism. While this approach benefits from teachers' expert knowledge, these materials are "based on intuition about how we use language, rather than actual evidence of use" (O'Keeffe, McCarthy and Carter 2007: p. 21). We aim at bringing together intuition and evidence about language use by employing insights from coursebooks for selecting examples from real-life corpora (e.g. news texts, novels).

Recent years have seen a number of efforts in the NLP community to automatically generate exercise items e.g. Arregik (2011); Smith, Avinesh and Kilgarriff (2010); Sumita, Sugaya and Yamamoto (2005). Most of these, however, tend to neglect what criteria sentences should fulfil in order to be suitable as exercise items and, instead, build on either a predefined set of manually selected sentences, or require merely a certain linguistic pattern (e.g. a particular word) to be present in the sentence (see section 9.2.2). When selecting sentences from corpora, however, there are a number of additional aspects that sentences need to adhere to in order to be usable and understandable in isolation. These have been previously explored mostly in a lexicographic context (Kilgarriff et al. 2008), but they are also relevant for language teaching (Kilgarriff 2009). Two fundamental questions in this respect are: (i) Can the sentence function in isolation, outside its larger textual context? (ii) Is the complexity of the linguistic content of the sentence suitable for the intended L2 learner(s)? We will refer to the former as *context independence* and to the latter as *L2 complexity*.

Language learners pass through different learning stages (levels) reflecting the development and improvement of their competences. A scale of such levels is *CEFR*, the Common European Framework of Reference for Languages (Council of Europe 2001). The CEFR defines proficiency levels on a six-point scale: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (mastery). A subset of language learners' competences are *linguistic competences*, which include, among others, lexical, grammatical and semantic competences. When assessing L2 complexity, we concentrate on linguistic competences required for reading comprehension since these can be matched to linguistic elements observable in language samples written for learners at different CEFR levels.

Both context independence and L2 complexity emerged as a main reason for discarding candidate sentences in previous evaluations (Arregik 2011; Pilán, Volodina and Johansson 2013), but thorough methods targeting these aspects have not been proposed up to date to our knowledge. Our approach, building on previous attempts at selecting sentences, contributes to previous research by offering a comprehensive set of criteria and by performing a more sophisticated selection in terms of the two fundamental aspects just mentioned, context independence and L2 complexity. We propose a hybrid system with both rule-based and machine learning driven components that encompasses a wide range of aspects. Incorporating rules makes the system customizable to users' needs and thus relevant for a wide range of application scenarios including vocabulary and grammatical exercises of different formats, as well as vocabulary examples. An evaluation with teachers and students indicates that our system identifies sentences that are, in general, of a suitable level of difficulty for learners. The algorithm is available to the general public free of charge both as a customizable sentence selection interface and as a web service. The development of automatically generated exercises using the selected sentences is also in progress. Our target language is Swedish, a language for which the number of L2 learners has grown rapidly over recent years (Statistics Sweden 2016). Although the current implementation is based on resources and tools for Swedish, the methods described can serve as an example for future implementations of exercise item candidate selection systems for other languages.

This paper is structured as follows. In section 9.2, we provide an overview of the related literature. Then, in section 9.3, we describe our sentence selection framework in detail together with its implementation. Finally, in section 9.4, we present and discuss the results of a user-based evaluation of the system.

9.2 Related work

In this section, we provide an overview of the related literature which includes sentence selection strategies for both vocabulary examples and exercise items as well as studies on readability and CEFR level prediction.

9.2.1 Sentence selection for vocabulary examples

GDEX, Good Dictionary Examples (Husák 2010; Kilgarriff et al. 2008) is an algorithm for selecting sentences from corpora for the purposes of illustrating the meaning and the usage of a lexical unit. It incorporates a number of linguistic criteria (e.g. sentence length, vocabulary frequency, anaphoric pronouns) based on which example candidates are ranked. Some of these are related to context dependence (e.g. incompleteness of sentences, presence of personal pronouns), but they are somewhat coarse-grained criteria without a focus on syntactic aspects.

Besides English, the algorithm has also been successfully implemented for other languages. Kosem, Husák and McCarthy (2011) and Tiberius and Kinable (2015) explore GDEX configurations for Slovene and Dutch respectively, aiming at identifying the optimal parameter settings for these languages for lexicographic projects. Didakowski, Lemnitzer and Geyken (2012) propose an example selection algorithm similar to GDEX for German. A fundamental difference of this method compared to the ranking mechanism of GDEX is having "hard criteria" which, if not met, result in sentences being excluded. GDEX has also inspired a Swedish algorithm for sentence selection (Volodina, Johansson and Johansson Kokkinakis 2012) and it has been employed also for generating gap-fill exercises (Smith, Avinesh and Kilgarriff 2010). Furthermore, a number of machine learning approaches have been explored for these purposes in recent years (Geyken, Pölitz and Bartz 2015; Lemnitzer et al. 2015; Ljubešić and Peronja 2015). Example sentence selection for illustrating lexical items has also been addressed from a language teaching perspective by Segler (2007), where a set of selection criteria used by teachers was modelled with logistic regression. The main dimensions examined include syntactic complexity and similarity between the original context of a word and an example sentence.

9.2.2 Sentence selection for exercise item generation

In a language-learning scenario, corpus example sentences can be useful both as exercise items and as vocabulary examples. Sentences used in exercises are also known as *seed sentences* (Sumita, Sugaya and Yamamoto 2005) or *carrier sentences* (Smith, Avinesh and Kilgarriff 2010) in the Intelligent, i.e. NLP-enhanced, Computer-Assisted Language Learning (ICALL) literature.

Previous work on exercise item generation has taken into consideration a rather limited amount of aspects when selecting seed sentences. In some cases, sentences are only required to contain a lexical item or a linguistic pattern that constitutes the target of the exercise, but context dependence and L2 complexity are not explicitly addressed (Sumita, Sugaya and Yamamoto 2005; Mitkov, Le An and Karamanis 2006; Arregik 2011; Wojatzki, Melamud and Zesch 2016). LanguageMuse (Burstein et al. 2012), a system supporting teachers in generating activities for learners of English, also belongs to this category. The sentences are selected from texts provided by teachers, the criteria of selection being the presence of a specific linguistic element that constitutes the target of the exercise: a lexical entity, a syntactic structure or a discourse relation.

Another alternative has been using dictionary examples as seed sentences, e.g. from WordNet (Pino and Eskenazi 2009). Such sentences are inherently context-independent, however, they impose some limitations on which linguistic aspects can be targeted in the exercises and they are not adjusted to finer-grained L2 learning levels. A system using GDEX for seed sentence selection is described in Smith, Avinesh and Kilgarriff (2010), who underline the importance of the well-formedness of a sentence and determine a sufficient amount of context in terms of sentence length. Lee and Luo (2016) describe an ICALL system for fill-in-the-blanks preposition learning exercises, where seed sentences are checked for their lexical difficulty based on the level of the words according to a graded vocabulary lists. Pilán, Volodina and Johansson (2014) present and compare two algorithms for candidate sentence selection for Swedish, using both rule-based and machine learning methods. Context dependence, which has not been specifically targeted in their system, emerged as a key issue underlying suboptimal candidate sentences during an empirical evaluation.

9.2.3 Readability and proficiency level classification

The degree of complexity in the linguistic content of sentences and texts is one of the aspects underlying not only proficiency levels, but also readability. Readability measures typically classify texts into school grade levels or into a binary category of easy- vs. hard-to-read, but the term has also been used in the context of CEFR level classification, e.g. Xia, Kochmar and Briscoe (2016), François and Fairon (2012). In recent years a number of NLP-based readability models have been proposed not only for English (Collins-Thompson and Callan 2004; Schwarm and Ostendorf 2005; Graesser, McNamara and Kulikowich 2011; Vajjala and Meurers 2012; Collins-Thompson 2014), but also for other languages, e.g. Italian (Dell' Orletta, Montemagni and Venturi 2011) and Swedish (Heimann Mühlenbock 2013). The linguistic features explored so far for this task include information, among others, from part-of-speech (POS) taggers and dependency parsers. Cognitively motivated features have also been proposed, for example, in the Coh-Metrix (Graesser, McNamara and Kulikowich 2011). Although the majority of previous work focuses primarily on text-level analysis, the concept of sentence-level readability has also emerged and attracted an increasing interest in recent years (Pilán, Volodina and Johansson 2013; Vajjala and Meurers 2014; Dell'Orletta et al. 2014).

The prediction of proficiency levels for L2 teaching materials using supervised machine learning methods has been explored for English (Heilman et al. 2007; Huang et al. 2011; Zhang, Liu and Ni 2013; Salesky and Shen 2014; Xia, Kochmar and Briscoe 2016), French (François and Fairon 2012), Portuguese (Branco et al. 2014), Chinese (Sung et al. 2015) and, without the use of NLP, for Dutch (Velleman and van der Geest 2014).

Readability formulas for the Swedish language have a long tradition. One of the most popular, easy-to-compute formulas is LIX (*Läsbarhetsindex*, 'Readability index') proposed by Björnsson (1968). This measure combines the average number of words per sentence in the text with the percentage of long words, i.e. tokens consisting of more than six characters. Besides traditional formulas, supervised machine learning approaches have also been tested. A Swedish document-level readability model is described by Heimann Mühlenbock (2013) and Falkenjack, Heimann Mühlenbock and Jönsson (2013). In chapter 7, based on Pilán, Vajjala and Volodina (2016), on the other hand, we investigated L2 complexity for Swedish both at document and sentence level.

9.3 HitEx: a sentence selection framework and its implementation

In this section, we present our candidate sentence selection framework, HitEx (*Hitta Exempel* 'Find Exemples') and its implementation. After an overall description, we introduce and motivate each selection criteria in sections 9.3.2 to 9.3.7.

9.3.1 Overall description

In table 9.1, we show the selection criteria belonging to the proposed framework, grouped into broader categories. Each *criterion* is used to scan a sentence for the presence (or the absence) of linguistic elements associated to its "goodness", i.e. its suitability for the intended use. Most criteria target aspects that are negatively correlated to the goodness of a sentence. Certain selection criteria are associated with one (or more) numeric *parameter(s)* that users can set, e.g. the minimum and maximum number of tokens for the sentence length criterion. The categories concerning the search term, well-formedness and context independence can be considered *generic* criteria that are applicable for a number of different use cases, e.g. different exercise types, vocabulary examples, while the rest of the criteria are more *specific* for exercise item generation. In general, the sources that served as basis for these criteria include previous literature (section 9.2), L2 curricula and the qualitative results of previous user-based evaluations (Volodina, Johansson and Johansson Kokkinakis 2012; Pilán, Volodina and Johansson 2014).

Nr	Criterion	Nr	Criterion
	Search term		**Additional structural criteria**
1	*Absence of search term*	13	Negative formulations
2	Number of matches	14	*Interrogative sentence*
3	*Position of search term*	15	*Direct speech*
	Well-formedness	16	*Answer to closed questions*
4	*Dependency root*	17	Modal verbs
5	Ellipsis	18	Sentence length
6	*Incompleteness*		**Additional lexical criteria**
7	Non-lemmatized tokens	19	Difficult vocabulary
8	Non-alphabetical tokens	20	Word frequency
	Context independence	21	Out-of-vocabulary words
9	*Structural connective in isolation*	22	Sensitive vocabulary
10	Pronominal anaphora	23	Typicality
11	Adverbial anaphora	24	Proper names
12	**L2 complexity in CEFR level**	25	Abbreviations

Table 9.1: HitEx: a sentence selection framework.

We implemented a *hybrid system* which uses a combination of machine-learning methods for assessing L2 complexity and heuristic rules for all other criteria. The motivation behind using rules is, on the one hand, that certain linguistic elements are easily identifiable with such methods. On the other hand, a sufficient amount of training data encompassing the range of all possible exercise types would be extremely costly to create. Moreover, explicit rules make the sentence selection customizable to users' task-specific needs which increases the applicability of HitEx to a diverse set of situations. The criterion of L2 complexity has been implemented using machine learning methods since its assessment comprises multiple linguistic dimensions and data was available for approaching this sub-problem in a data-driven fashion. A few selection criteria in our framework are re-implementations of those described by Volodina, Johansson and Johansson Kokkinakis (2012) and Pilán, Volodina and Johansson (2014). Major additions to previous work include: (i) L2 complexity assessment on a 5-level scale, vs. a previously available binary classification model by Pilán, Volodina and Johansson (2014), (ii) typicality and (iii) the assessment of context dependence. Sensitive vocabulary filtering and the use of word frequencies from *SVALex* (François et al. 2016), a word list based on coursebook texts, are also novel aspects that we incorporated with the aim of making the sentence selection algorithm more pedagogically aware.

Our implementation relies on a number of different NLP resources. Our system searches for sentence candidates via *Korp* (Borin, Forsberg and Roxendal 2012), an online infrastructure providing access to a variety of (mostly) Swedish corpora. The concordance web service of Korp provides a list of corpus examples containing a certain user-specified search term, e.g. an uninflected word, *lemma* or a grammatical structure. Through Korp, a large variety of text genres are available such as novels, blogs, news and easy-to-read texts. All corpora are annotated at different linguistic levels, which include lemmatization, part-of-speech (POS) tagging and dependency parsing. HitEx assesses sentences based on these annotations as well as information from a number of Swedish lexical-semantic resources. A major lexical resource used is SALDO (Borin, Forsberg and Lönngren 2013) which is based on lexical-semantic closeness between word senses organized in a tree structure.

As a first step in our sentence *scoring algorithm*, for each candidate sentence $s \in S$, we apply a linguistic criterion $c \in C$ to s either as a *filter* $f \in F$ or as a *ranker* $r \in R$, that is $C = F \cup R$. The application of each criterion c_k to all the sentences, $c_k(S) = V_{c_k}$ is a set of criterion *values* ($v_{c_k} \in V_{c_k}$). $V_{c_k} = \{0,1\}$ when $c_k \in F$ and $V_{c_k} \in \mathbb{R}$ when $c_k \in R$; for instance, when c_k is the proper names criterion used as a ranker, $v_{c_k s_i}$ corresponds to the number of times a proper name appears in $s_i \in S$. If s_i contains an undesired linguistic element associated to $c_k \in F$, then $v_{c_k s_i} = 1$, and s_i is excluded from the ranking of suitable candidates. Further details about how we obtain V_{c_k} are outlined in sections 9.3.2 to 9.3.7. Some criteria encode binary characteristics (e.g. interrogative sentence), therefore, only $c_k \in F$ holds for these. We present these in italics in table 9.1.

To rank non-filtered sentences, we compute a *goodness* score $G_{s_i} \in \mathbb{N}$, which reflects the degree to which s_i is a suitable candidate based on R. When $c_k \in R$, $R = R^+ \cup R^-$, where $r^+ \in R^+$ is a *positive ranker* for positively correlated properties with goodness, namely typicality and SVALex frequencies; and $r^- \in R^-$ is a *negative ranker* that includes all other criteria. Based on V_{c_k}, we compute an intermediate (per-criterion) goodness score ($subG_{c_k s_i}$) for each s_i, by sorting S based on V_{c_k} and assigning the ranking position of s_i according to V_{c_k} to $subG_{c_k s_i}$. Consequently, the number of subscores will be equal to the number of $c_k \in R$ selected. During this sorting, for $s_i \in S$ and $s_j \in S$, for r^+ $subG_{c_k s_i} \geq subG_{c_k s_j} \Leftrightarrow v_{c_k s_i} \geq v_{c_k s_j}$ holds, and for r^- $subG_{c_k s_i} \geq subG_{c_k s_j} \Leftrightarrow v_{c_k s_i} \leq v_{c_k s_j}$ applies. In other words, we rank S based on an ascending order of goodness if $c_k \in R^+$ and a descending order of badness if $c_k \in R^-$. Therefore, more suitable candidates receive a higher $subG_{c_k s_i}$. For example, suppose r_k^- is proper names and s_i contains 2 proper names, while s_j contains none; then

$subG_{c_k s_i} = 1$ and $subG_{c_k s_j} = 2$. The score G_{s_i} is then computed by summing all subscores, that is $G_{s_i} = \sum subG_{c_k s_i}$. Finally, candidate sentences are ordered in a decreasing order based on G_{s_i}. A weighting scheme similar to GDEX would be possible with the availability of data specific for the end use of the sentences from where to estimate these weights. At the current stage, all ranking criteria contribute equally to G_{s_i}. Suboptimal sentences containing elements to filter can also be retained and ranked separately, if so wished, based on the amount of F matched. The final results include, for each $s_i \in S$, its summed overall score (G_{s_i}), its final rank and *detailed information* per selection criteria, as the screenshot presenting the system's graphical user interface in figure 9.1 in section 9.3.8 shows. In the following subsections, we present each criterion in detail, grouped into categories.

9.3.2 Search term

A *search term* corresponds to one (or more) linguistic element(s) that users would like the selected sentences to contain. It can be either a lexical element such as an inflected word or a lemma; or a grammatical pattern, e.g. verbs in a certain tense followed by a noun. The presence of a search term is guaranteed through the mere use of the Korp concordance web service which only returns sentences containing the searched expression. In some application scenarios, repeated matches of the search term may be considered suboptimal (Kosem, Husák and McCarthy 2011: p. 157), therefore we include this aspect among our criteria. Similarly, there might be a preference for the *position* of the search term in the sentence in some use cases such as dictionary examples (Kilgarriff et al. 2008).

9.3.3 Well-formedness

Good candidate sentences from corpora should be structurally and lexically well-formed (Kilgarriff et al. 2008; Husák 2010). We incorporate two criteria targeting the former aspect: one can check sentences for the presence of a dependency *root*, and *ellipsis*, i.e. the lack of a subject or a finite verb (all verb forms except infinitive, supine and participle) inspired by Volodina, Johansson and Johansson Kokkinakis (2012). The completeness criterion checks the beginning and the end of a sentence for orthographic clues such as capital letters and punctuation, in a similar fashion to what we described in chapter 8 based on Pilán (2016). A large amount of *non-lemmatized tokens*, i.e. tokens for which no matching dictionary form could be identified (in the SALDO lexicon in our case), are also preferably avoided (Husák 2010: p. 15). These are mostly cases of spelling or optical character recognition errors, foreign words, infrequent compounds, etc. A large portion of *non-alphabetical tokens* could be e.g. a sign of mark-up traces in web material, which has a negative influence on the L2 complexity and the usability of a sentence (Husák 2010: p. 15). Users can specify a constant as a threshold for these criteria to determine the allowed amount of non-lemmatized and non-alphabetical tokens in a sentence.

9.3.4 Context independence

Since sentences originally form part of coherent texts, a crucial aspect to take into consideration during selection is whether sentences would be meaningful also as a stand-alone unit without their original, larger context. The presence of linguistic elements responsible for connecting sentences at a syntactic or semantic level is therefore suboptimal (Kilgarriff et al. 2008). We incorporate a number of criteria for capturing this aspect which we described also in chapter 8 based on Pilán (2016).

Syntactic aspects include *structural connectives*, i.e. conjunctions and subjunctions (Webber et al. 2003). Two concepts connected by structural connectives may appear in separate sentences which give rise to context dependence. Our system considers sentences with connectives in sentence-initial position context dependent unless the sentence consists of more than one clause. Connectives that are paired conjunctions are also allowed (e.g. *antingen ... eller* ‘either ... or’).

Anaphoric expressions referring to previously mentioned information are aspects related to the semantic dimension. Our pronominal anaphora criterion targets mentions of the third person singular pronouns *den* ‘it’ (common gender) and *det* ‘it’ (neuter gender) as well as the demonstrative pronouns (e.g. *denna* ‘this’, *sådan* ‘such’ etc.). The non-anaphoric use of *det* (e.g. in clefts: *It is carrots that they eat.*), however, is not counted here. Such cases can be distinguished based on the output of the dependency parser: these occurrences of *det* are tagged as expletive (pleonastic). Pronouns followed by a relative clause introduced by *som* ‘which’ are also considered non-anaphoric.

Under *adverbial anaphora*, we count time and location adverbs that behave anaphorically (e.g. *då* ‘then’) (Webber et al. 2003). Another group of adverbs relevant for anaphora are those expressing logical relations (e.g. *istället* ‘instead’), which are also referred to as *discourse connectives* (Webber et al. 2003). Based on Teleman, Hellberg and Andersson (1999), a list of anaphoric adverbs has been collected and sentences are checked for the occurrence of any of the listed items.

9.3.5 L2 complexity

The aspect of L2 complexity has been assessed with the help of a supervised machine learning algorithm based on a number of different linguistic dimensions. We used the CEFR level classifier for sentences that we previously described in chapter 7, based on Pilán, Vajjala and Volodina (2016). The source of the training data was single sentences from COCTAILL (Volodina et al. 2014b), a corpus of coursebook texts for L2 Swedish. Such single sentences occurred either in the form of lists or so-called *language examples*, sentences exemplifying a lexical or a grammatical pattern. The feature set used for assessing L2 complexity is presented in table 9.2. This set consists of five subgroups of features: count-based, lexical, morphological, syntactic, and semantic features.

Count features are based on the number of characters and tokens (*T*), extra-long words being tokens longer than 13 characters. LIX, a traditional Swedish readability formula (see section 9.2) combines the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters (Björnsson 1968). Bi-logarithmic and a square root type-token ratio (TTR) related to vocabulary richness (Heimann Mühlenbock 2013) are also computed.

Lexical features incorporate information from the KELLY list (Volodina and Kokkinakis 2012), a word list with frequencies calculated from a corpus of web texts (thus completely independent of the sentences in the dataset). KELLY provides a suggested CEFR level per each listed lemma based on frequency bands. For some feature values, *incidence scores* (IS) normalized values per 1,000 tokens are computed, which reduces the influence of sentence length. Word forms or lemmas themselves are not used as features, the IS of their corresponding CEFR level is considered instead. *Difficult* tokens are those that belong to levels above the overall CEFR level of the text. Moreover, we consider the IS of tokens not present in KELLY (*OOV IS*), the IS of tokens for which the lemmatizer could not identify a corresponding lemma (*No lemma IS*), as well as average KELLY log frequencies.

Morphological features include both IS and variational scores, i.e. the ratio of a category to the ratio of lexical tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). The IS of all lexical categories as well as the IS of punctuation, particles, sub- and conjunctions (SJ, CJ) are taken into consideration. In Swedish, a special group of verbs ending in -s are called *s-verbs* (*S-VB*). These can indicate either a reciprocal verb, a passive construction or a deponent verb (active in meaning but passive in form). Due to their morphological and semantic peculiarity, they are explicitly targeted in L2 grammar books (Fasth and Kannermark 1997). Nominal ratio (Hultman and Westman 1977) is another readability formula proposed for Swedish that corresponds to the

Name	Type	Name	Type
Sentence length	COUNT	Modal V to V	MORPH
Avg token length	COUNT	Particle IS	MORPH
Extra-long token	COUNT	3SG pronoun IS	MORPH
Nr characters	COUNT	Punctuation IS	MORPH
LIX	COUNT	Subjunction IS	MORPH
Bilog TTR	COUNT	PR to N	MORPH
Square root TTR	COUNT	PR to PP	MORPH
Avg KELLY log freq	LEXICAL	S-V IS	MORPH
A1 lemma IS	LEXICAL	S-V to V	MORPH
A2 lemma IS	LEXICAL	ADJ IS	MORPH
B1 lemma IS	LEXICAL	ADJ variation	MORPH
B2 lemma IS	LEXICAL	ADV IS	MORPH
C1 lemma IS	LEXICAL	ADV variation	MORPH
C2 lemma IS	LEXICAL	N IS	MORPH
Difficult W IS	LEXICAL	N variation	MORPH
Difficult N&V IS	LEXICAL	V IS	MORPH
OOV IS	LEXICAL	V variation	MORPH
No lemma IS	LEXICAL	Function W IS	MORPH
Avg. DepArc length	SYNTACTIC	Neuter N IS	MORPH
DepArc Len $>$ 5	SYNTACTIC	CJ + SJ IS	MORPH
Max length DepArc	SYNTACTIC	Past PC to V	MORPH
Right DepArc Ratio	SYNTACTIC	Present PC to V	MORPH
Left DepArc Ratio	SYNTACTIC	Past V to V	MORPH
Modifier variation	SYNTACTIC	Supine V to V	MORPH
Pre-modifier IS	SYNTACTIC	Present V to V	MORPH
Post-modifier IS	SYNTACTIC	Nominal ratio	MORPH
Subordinate IS	SYNTACTIC	N to V	MORPH
Relative clause IS	SYNTACTIC	Lex T to non-lex T	MORPH
PP complement IS	SYNTACTIC	Lex T to Nr T	MORPH
Avg senses per token	SEMANTIC	Relative structure IS	MORPH
N senses per N	SEMANTIC		

Table 9.2: The feature set for L2 complexity assessment.

ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR), adverbs, and verbs. Relative structures consist of relative adverbs, determiners, pronouns and possessives.

Syntactic features are based, among others, on the length (depth) and the direction of dependency arcs (*DepArc*). These aspects are related to readers working memory load when processing sentences (Gibson 1998). For similar reasons, we consider also relative clauses as well as pre- and post-modifiers, which include, for example, adjectives and prepositional phrases respectively.

Semantic features draw on information from the SALDO lexicon. We use the average number of senses per token and the average number of noun senses per noun. Polysemous words can be demanding for readers as they need to be disambiguated for a full understanding of the sentence (Graesser, McNamara and Kulikowich 2011).

In Pilán, Vajjala and Volodina (2016) and chapter 7, utilizing the feature set described above, we reported 63.4% accuracy using a logistic regression classifier for the identification of CEFR levels with an exact match, and 92% accuracy for classifications within a distance of one CEFR level. Besides the features outlined above, also the lack of culture-specific knowledge can be a factor influencing L2 complexity, as well as learners' knowledge of other languages. We, however, do not address these dimensions in the current stage due to a lack of relevant data.

9.3.6 Additional structural criteria

Besides the aspects mentioned above, a number of additional structural criteria are available which proved to be relevant either based on previous evaluations (Volodina, Johansson and Johansson Kokkinakis 2012; Pilán, Volodina and Johansson 2013) or evidence from coursebooks (Volodina et al. 2014b). One such aspect is *negative wording* which is preferable to avoid in exercise items (Frey et al. 2005). All tokens with the dependency tag of negation adverbials fall under this criterion. Under the *interrogative sentence* criterion, we handle direct questions ending with a question mark. To detect *direct speech*, we have compiled a list of verbs denoting the act of speaking based on the Swedish FrameNet (Heppin and Gronostaj 2012). The list contains 107 verbs belonging to frames relevant to speaking (e.g. *viska* 'whisper' from the *Communication manner* frame). This is combined with POS tag patterns composed of a minor delimiter (e.g. dash, comma) or a pairwise delimiter (e.g. quotation marks), followed by a speaking verb (optionally combined with auxiliary verbs), followed by a pronoun or a proper name. Both questions and sentences containing direct speech tend to be less common as exercise items, incorporating these among our criteria allows users to avoid such sentences if so wished.

Answers to polar (or close-ended) questions are rarely employed as exercise items and they were also negatively perceived in previous evaluations (Volodina, Johansson and Johansson Kokkinakis 2012; Pilán, Volodina and Johansson 2013). This aspect is also relevant to the dependence of a sentence on a wider context. The algorithm tries to capture sentences of this type based on POS patterns: sentence-initial adverbs and interjections (e.g. *ja* ‘yes’, *nej* ‘no’) preceded and followed by minor delimiters where the initial delimiter is optional for interjections. *Modal verbs* were identified based on a small set of verbs used typically (but not exclusively) as modal verbs (e.g. *kan* ‘can’, ‘know’) where the dependency relation tag indicating a verb group excludes the non-auxiliary use. *Sentence length*, a criterion which is also part of GDEX, is measured as the number of tokens including punctuation in our system.

9.3.7 Additional lexical criteria

HitEx includes also options for filtering and ranking sentences based on information from lexical resources for ensuring an explicit control of this crucial aspect (Segler 2007). Sentences containing *difficult words*, i.e. words whose CEFR level is above the target CEFR level according to the KELLY list, can be penalized or filtered. Besides KELLY, we also integrated into our system information from the SVALex list based on word frequencies from coursebook texts. Sentences with words absent from SVALex or with words below the average frequency threshold for the target CEFR level are thus additional scoring criteria. Another criterion involves the presence of *proper names* which, although undesirable for dictionary examples (Kilgarriff et al. 2008), may be familiar and easy to understand for L2 students (Segler 2007). Both proper names and *abbreviations* were counted based on the POS tagger output.

In a pedagogical setting, certain *sensitive vocabulary items* and topics tend to be avoided by coursebook writers. These are also referred to as PARSNIPs, which stands for Politics, Alcohol, Religion, Sex, Narcotics, Isms[35] and Pork (Gray 2010). Some topics are perceived as taboos cross-culturally, such as swear words, while others may be more culture-bound. We compiled a word list starting with an initial group of seed words from more generally undesirable domains (e.g. swear words) collected from different lexical and collaborative online resources (e.g. Wikipedia)[36] complemented with a few manually added entries. Furthermore, we expanded this automatically with all child node senses from SALDO for terms which represent sensitive topics (e.g. *narkotika* ‘narcotics’, *svordom* ‘profanities’, *mörda* ‘murder’, etc.) so that synonyms and hyperonyms would also be included. A few common English swear words that are frequently used in Swedish in an informal context (e.g. blog texts)

[35] An ideology or practice, typically ending with the suffix *-ism*, e.g. *anarchism*.

[36] `https://www.wikipedia.org`.

were also incorporated. The current list of 263 items is not exhaustive and can be expanded in the future. The implementation allows teaching professionals to make the pedagogical decision of tailoring the subset of topics to use to a specific group of learners during the sentence selection.

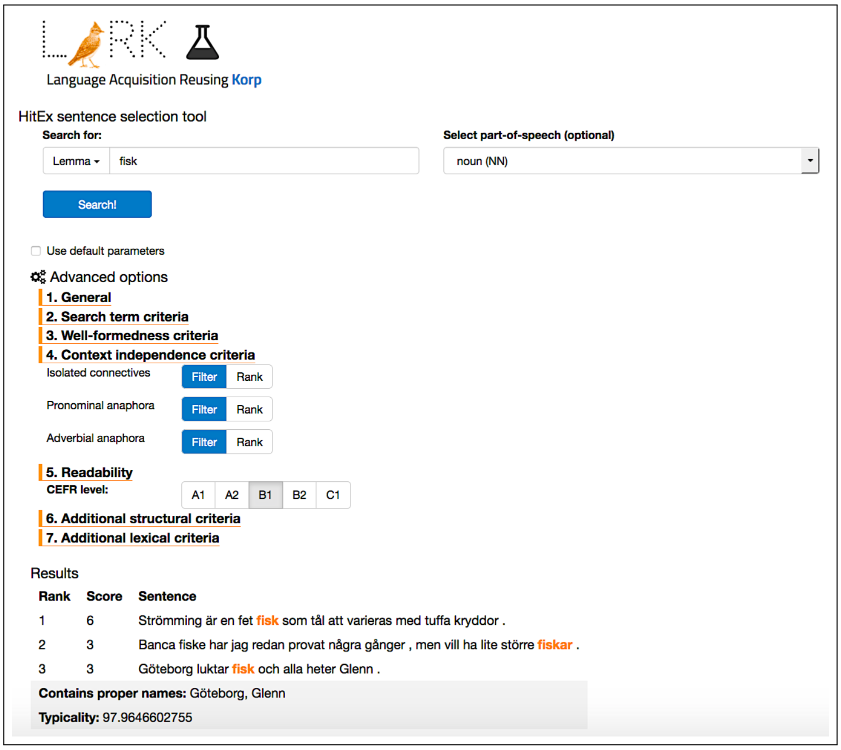
Typicality can be an indication of more or less easily recognizable meaning of concepts without a larger context (Barsalou 1982). We assessed the typicality of sentences with the help of a co-occurrence measure: *Lexicographers' Mutual Information* (LMI) score (Kilgarriff et al. 2004). LMI measures the probability of two words co-occurring together in a corpus and it offers the advantage of balancing out the preference of the Mutual Information score for low-frequency words (Bordag 2008). We used a web service offered by Korp for computing LMI scores based on Swedish corpora. As a first step in computing the LMI scores, we collected nouns and verbs in the KELLY and SVALex lists (removing duplicates), which resulted in a list of 12,484 items. Then using these, we estimated LMI scores for all noun-verb combinations (nouns being subjects or objects) as well as LMI for nouns and their attributes using Korp. Counts were based on 8 corpora of different genres amounting to a total of 209,110,000 tokens. The resulting list of commonly co-occurring word pairs consisted of 99,112 entries. Only pairs with a threshold of LMI $\geq$ 50 were included. The typicality value of a candidate sentence corresponded to the sum of all LMI scores available in the compiled list for each noun and verb in the sentence.

9.3.8 Integration into an online platform

To provide access to our sentence selection algorithm to others, we have integrated it into a freely available learning platform, Lärka.[37] With the help of a graphical interface, shown in figure 9.1, users can perform a sentence selection customized to their needs. Under the advanced options menu, users can choose which selection criteria presented in table 9.1 to activate as filters or rankers.

Moreover, the selection algorithm will serve as a seed sentence provider for automatically generated multiple-choice exercises for language learners within the same platform. The sentence selection algorithm is also available as a web service that others can easily integrate in their own systems.

[37] `https://spraakbanken.gu.se/larkalabb/hitex`.



HitEx sentence selection tool

Search for: Lemma ▾ fisk

Select part-of-speech (optional): noun (NN)

□ Use default parameters

1. General

2. Search term criteria

3. Well-formedness criteria

4. Context independence criteria

Isolated connectives: Filter | Rank

Pronominal anaphora: Filter | Rank

Adverbial anaphora: Filter | Rank

5. Readability

CEFR level: A1 A2 [✓ B1] B2 C1

6. Additional structural criteria

7. Additional lexical criteria

Results

Rank	Score	Sentence
1	6	Strömming är en fet **fisk** som tål att varieras med tuffa kryddor .
2	3	Banca fiske har jag redan provat några gånger , men vill ha lite större **fiskar** .
3	3	Göteborg luktar **fisk** och alla heter Glenn .

Contains proper names: Göteborg, Glenn

Typicality: 97.9646602755

Figure 9.1: The HitEx user interface with *fisk* ‘fish’ as search term.

9.4 A user-based evaluation

The main objective when developing our candidate selection algorithm was to identify seed sentences for L2 learning exercises. In absence of an annotated dataset for this task in Swedish, we tested the performance of HitEx with the help of a user-based evaluation. We assessed the goodness of the candidate sentences in two ways: (i) through L2 teachers confirming their suitability, (ii) by inspecting whether L2 learners’ degree of success in solving exercise items constructed based on these candidates matched what is typically expected in L2 teaching. This provided us with information about the extent to which the set of criteria proposed in section 9.3 was useful for identifying suitable seed sentences. The evaluation sentences and the associated results will be available as a dataset on `https://spraakbanken.gu.se/eng/resources`.

9.4.1 Participants

The participants consisted of 5 teachers of L2 Swedish from different institutions and 19 students from a language school targeting young adults newly arrived to Sweden. Participating students were between ages 16 and 19 with a variety of native languages including several Somali and Dari speakers. The proportion of female and male students was approximately equal. The CEFR level of students is assessed on a regular basis with a two-month interval in their school. In our evaluation, as a point of reference for students' CEFR level, we referred to the levels achieved on their latest assessment test. According to this, 3 students were at A1 level, and the remaining 16 were a 50–50% split between A2 and B1 levels.

9.4.2 Material and task

To create the evaluation material, we retrieved a set of sentences from Korp for CEFR levels A1–C1 using HitEx. To perform the Korp concordance search, we used lemmas from SVALex whose level corresponded to the level of the sentences we aimed at identifying. We used a lemma-based search and the parts of speech included nouns, verbs and adjectives. The sentences have been selected from 10 different corpora including novels and news texts. For each search lemma, a maximum of 300 matching Korp sentences were fed to the sentence selection algorithm, out of which only the top ranked candidate for each lemma was included in the evaluation material. Most selection criteria were used as filters, but typicality, proper names, KELLY and SVALex frequencies were used as rankers. Modal verbs were allowed in the sentences and the position of the search term was not restricted. Sentence length was set to a minimum of 6 and a maximum of 20 tokens. The threshold used for the percentage of non-alphabetic and non-lemmatized tokens was 30%.

Teachers received 330 sentences to evaluate, evenly distributed across the 5 CEFR levels A1–C1. The sentences were divided into two subgroups based on their level, at least two teachers rating each sentence. One set consisted of A1–B1 level sentences and the other of sentences within levels B1–C1. (B1 level sentences have been evenly split between the two subsets.) There was a *common subset* of 30 sentences from all 5 CEFR levels which was rated by all 5 teachers. Besides an overall score per sentence reflecting the performance of the combination of all criteria from table 9.1, we elicited teacher judgements targeting two criteria in particular, which were focal points during the implementation of HitEx, namely context independence and L2 complexity (see sections 9.3.4 and 9.3.5 respectively). No specific exercise type needed to be considered for evaluating these aspects, but rather a more application-neutral scenario of a learner reading the sentence. Teachers rated the three dimensions on a 4-point scale as defined in table 9.3. Besides these aspects, teachers were also required to suggest an alternative CEFR level if they did not agree with the one predicted by the system.

The sentence...	
1	*... doesn't satisfy the criterion.*
2	*... satisfies the criterion to a smaller extent.*
3	*... satisfies the criterion to a larger extent.*
4	*... satisfies the criterion entirely.*

Table 9.3: Evaluation scale.

To investigate further whether our selection criteria with the chosen setting produced good seed sentence candidates at the CEFR levels predicted by our L2 complexity criteria, we observed L2 learners' performance on exercise items created out of these sentences. Exercise generation requires a number of additional steps after the selection of seed sentences, many of which are open research problems. Therefore, we opted for a semi-automatic approach to the generation of these exercises. We manually controlled the combination of sentences into exercises and the selection of a *distractor*, an incorrect answer option which did not fit into any sentence, in order to reduce potential ambiguity in answer options. A subset of the sentences given to teachers were used as exercise items so that teachers' ratings and students' answers could be correlated.

The exercise type chosen was *word bank*, a type of matching exercise, since this posed less challenges when selecting distractors compared to multiple-choice items. Word bank exercises consist of a list of words followed by a list of sentences, each containing a gap. Learners' task is to identify which word is missing from which sentence. We created worksheets consisting of word bank exercises in Google Forms.[38] To lower the probability of answering correctly by chance, we added a distractor. Students had to provide their answers in a grid following the list of candidate words and the gapped sentences. The missing word to identify (and its position) corresponded to the search term used to retrieve the sentence from Korp. Worksheets consisted of 9 exercises with 5 sentences each, amounting to a total of 45 sentences. (The only exception was A1 level, where students had 2 exercises less.) Students had 60 minutes to work with the exercises, including 5 minutes of instructions. Students worked

[38] `https://docs.google.com/forms/`.

individually in a computer lab, access to external resources was not allowed.

The difficulty of the exercises varied along two dimensions: in terms of their CEFR level and in terms of the similarity of the morpho-syntactic form of the candidate words included in the word bank. A worksheet for a certain level contained 5 exercises from the same level as well as 2 exercises from one level below and one level above. In 5 exercises, the word bank consisted of lexical items with the same morpho-syntactic form (e.g. only plural neuter gender nouns), while 4 exercises had a word bank with mixed POS. The latter group of exercises was somewhat easier, since, besides lexical-semantic knowledge, students could identify the correct solution also based on grammatical clues such as inflectional endings.

9.4.3 Results and discussion

Below, we present teachers' and students' results on the evaluation material.

9.4.3.1 Teachers

To understand to which extent our set of criteria was able to select suitable seed sentences overall as well as specifically in terms of L2 complexity and context independence, we computed average values and standard deviation (StDev) over L2 teachers' ratings. (8 sentences had to be excluded between A1-B1 levels due to missing values.) The results are presented in table 9.4.

Criterion	**# of raters**	**Average**	**StDev**
L2 complexity	5	3.18	0.53
Context independence	5	3.05	0.56
Overall suitable (all criteria)	4	3.23	0.73

Table 9.4: Average teacher-assigned rating per criteria.

As for the criterion of context independence, 80% of the sentences were found suitable (received an average score higher than 2.5), and 61% of the sentences received score 3 or 4 from at least half of the evaluators. Besides rating the three dimensions in table 9.4, teachers also provided an alternative CEFR level in case they did not agree with the CEFR level suggested by the system. HitEx correctly assessed L2 complexity for 64% of sentences based on teachers' averaged CEFR label, and in 80% of the cases the system's CEFR level coincided with at least one teacher's decision. Besides comparing system-assigned and teacher-assigned levels, we measured also the inter-rater agreement (IAA) among the teachers. We used *Krippendorff's* α measuring observed and expected disagreement, since it is suitable for multiple raters. An $\alpha = 1$ corresponds to complete agreement, while $\alpha = 0$ is equivalent to chance agreement. The inter-rater agreement results among teachers are presented in table 9.5. The extent of agreement among teachers was considerably higher than chance agreement, but it still remained below what is commonly considered as reliability threshold in annotation tasks, namely $\alpha = 0.8$. CEFR level assignment for sentences thus seems to be a hard task even for teaching professionals.

SentID	# sents	# raters	CEFR	IAA
1-38	38	5	A1-C1	0.65
39-188	142	2	A1-B1	0.68
189-338	150	3	B1-C1	0.53
Tot/Avg	330	5	A1-C1	0.62

Table 9.5: Inter-rater agreement for CEFR level assignment.

Besides inter-rater agreement in terms of α, we considered also the distance between the CEFR levels assigned by all teachers compared both to each other and to HitEx (table 9.6). This would provide us information about the degree to which teachers' accepted our system's assessment of L2 complexity. CEFR levels were mapped to integer values for this purpose, and *averaged pairwise distances* between the levels have been computed in all cases. Surprisingly, teachers agreed with each other exactly on the CEFR level of sentences in only half of the cases, which shows that the exact CEFR level assignment on a 5 point scale is rather difficult even for humans. The percentage of sentences that teachers agreed on with HitEx completely (distance of 0) was slightly (4%) higher than the extent to which teachers agreed with each other. This may be due to the fact that teachers were confirming the system-assigned CEFR levels, but did not have information about each others' answers. Teacher-assigned CEFR levels remained within 1 level of difference when compared to each other in almost all cases and compared to the system for 92% of the sentences. All in all, the automatic CEFR levels predicted by HitEx were accepted by teachers in the majority of cases within 1 level distance.

Finally, we computed the *Spearman correlation coefficient* for teachers' scores of overall suitability and the two target criteria, L2 complexity and context independence, to gain insight into how strongly associated these two aspects were to seed sentence quality according to our evaluation data. The correlation over all sentences was $\rho = 0.34$ for L2 complexity and $\rho = 0.53$ for context

Level Distance	Teacher - Teacher	Teacher - System I
0	50.0	**53.9**
1	**49.4**	37.9
2	0.6	6.7
≥ 3	0.0	1.5

Table 9.6: Percentage of sentences per assigned CEFR label distance.

dependence. The maximum possible value is $\rho = 1$ for a positive correlation and $\rho = -1$ for a negative one. Both criteria were thus positively associated with overall suitability: the more understandable and context-independent a sentence was, the more suitable our evaluators found it overall. Out of the two criteria, context dependence showed a somewhat stronger correlation.

9.4.3.2 Students

First, based on students' responses, we computed *item difficulty* for each exercise item, which corresponds to the percentage of students correctly answering an item, a higher value thus indicating an easier item (Crocker and Algina 1986). The average item difficulty over all exercises was 0.62, corresponding to 62% of students correctly answering items on average. Table 9.7 shows additional average item difficulty scores divided per CEFR level, exercise type (distractors with same or different morpho-syntactic form) and POS. Values were averaged only over the exercise items that were of the same CEFR level as the answering students' level according to the system.

Ex. type	**SAME POS+INFL**			**Avg**	**MIXED POS+INFL**			**Avg**
CEFR	**A1**	**A2**	**B1**		**A1**	**A2**	**B1**	
Noun	0.67	0.83	0.73	**0.74**	0.67	0.52	0.65	**0.61**
Verb	0.50	0.69	0.69	**0.63**	0.0	0.62	0.77	**0.46**
Adjective	-	-	-	-	0.58	0.56	0.62	**0.59**
Avg	0.59	0.76	0.71	**0.69**	0.42	0.57	0.68	**0.55**
Overall				**0.62**				

Table 9.7: Average item difficulty per exercise item category, POS and CEFR level.

To be able to measure whether the item difficulty observed in our students' results matched the values one would typically expect in L2 teaching, we calculated the *ideal item difficulty* (IID) score for our exercises, which takes into consideration correct answers based on chance. We used the formula proposed by Thompson and Levitov (1985) presented in (13), where P_C is the probability of correct answers by chance.

$$IID = P_C + \frac{1 - P_C}{2} \tag{13}$$

Our exercises consisted of 5 gapped items and 6 answer options in the word bank. Students had, thus, a chance of 1/6 for filling in the first item, 1/5 for the second item etc., which corresponds to an average P_C of $(0.167 + 0.2 + 0.25 + 0.333 + 0.5)/5 = 0.29$ for the whole exercise and, consequently, an IID score of 0.645, that is 64.5% of students correctly answering. The observed item difficulty averaged over all students and exercise items of our evaluation was 62%, which is only slightly lower than the ideal item difficulty. If we break down this average to students' CEFR levels, we can notice that for A1 students the exercises were considerably more challenging than they should have been according to the ideal threshold. Only 51% of them responded correctly A1-level exercise items. Our sample size was particularly small, however, at this level, thus further evaluations with additional students would be required to confirm this tendency. A2 and B1 level students produced considerably better results: averaging over exercise types and POS, 66.5% and 69.5% of them respectively answered correctly the items of their levels. This indicates that the set of criteria proposed in section 9.3 can successfully select seed sentences for L2 exercises for students of A2 and B1 levels.

Contrary to what one might expect, exercise items with distractors bearing different morpho-syntactic forms proved to be actually harder for students compared to items with the same POS and inflection based on our evaluation data. The latter would be inherently harder since only lexico-semantic information can contribute to solving the exercises without the help of grammatical clues. As the item difficulty values show in table 9.7, approximately 14% more students answered correctly exercise items with distractors with the same morpho-syntactic tags, an outcome, which, however, may also depend on the inherent difficulty of the sentences presented. As mentioned in section 9.4.2, the work sheets also included exercises constructed with sentences belonging to one CEFR level higher and lower than students' level. This allowed us to further assess whether the CEFR levels suggested based on the L2 complexity criterion were appropriate. We display in figure 9.2 students' performance based on the CEFR level of exercise items comparing the system-assigned and the teacher-suggested CEFR levels for the items.

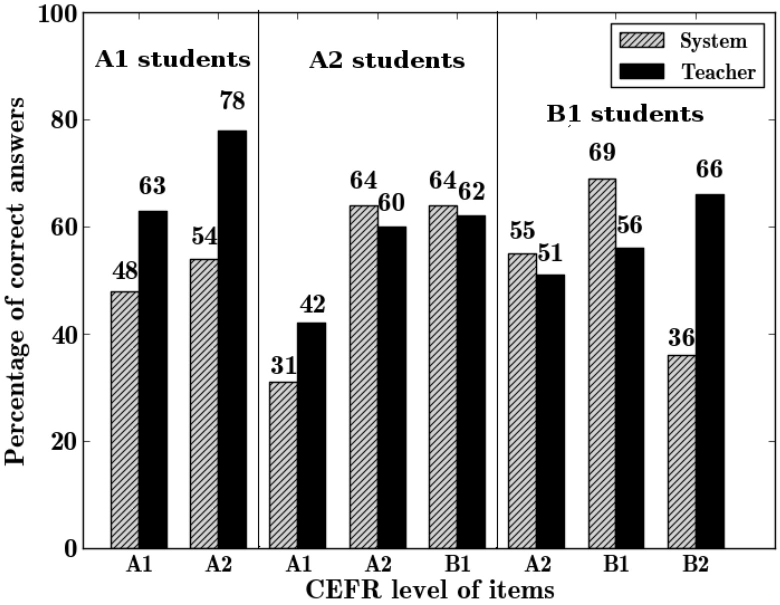


Figure 9.2: Correct answers per averaged teacher and system CEFR level.

As figure 9.2 shows, at A1 level students answered a larger amount of items according to the CEFR level determined by teachers (63%, vs. 48% with the system-assigned CEFR). The percentage of correct answers at A2 and B1 levels, however, showed more consistency with the levels assigned by our L2 complexity criterion: 64% (A2) and 69% (B1) correct answers based on our system's CEFR levels, vs. 60% (A2) and 56% (B1) with teacher-assigned levels. When considering these scores, however, it is worth noting that both teachers and the system were assessing only seed sentence difficulty, not the overall difficulty of the exercises. A few additional aspects play a role in determining the difficulty of exercise items, e.g. the selected distractors (Beinborn, Zesch and Gurevych 2014a), nevertheless the observed tendencies in error rates provide useful insights into the suitability of seed sentences in terms of L2 complexity.

9.5 Conclusion

We presented a comprehensive framework and its implementation for selecting sentences useful in the L2 learning context. The framework, among others, includes the assessment of L2 complexity in sentences and their independence of the surrounding context, both of which are relevant for a wide range of application scenarios. To our knowledge, this is the first comprehensive study addressing automatic seed sentence selection for L2 learning exercises. We invested considerable effort into creating a system that would yield pedagogically more relevant results. We conducted an empirical evaluation with L2 teachers and learners to gain insights into how successfully the proposed framework can be used for the identification of seed sentences for L2 exercises. Although the sample size was somewhat limited, the evaluation yielded very promising results. On average, the selected sentences lived up to teachers' expectations on L2 complexity, context independence and overall suitability. The exercises constructed with the use of the selected sentences were overall somewhat hard for beginners, but they were of an appropriate difficulty level for learners at subsequent stages. Moreover, learners' error rates at some levels correlated even slightly better with the CEFR levels predicted by our system than the averaged levels assigned by teachers. All in all, the evaluation indicated that the described system has good potentials to enhance language instruction by either assisting teaching professional when creating practice material or by providing self-learning opportunities for learners in the form of automatically generated exercises. Although our main focus was on seed sentences selection, the proposed system can be useful also for the identification of dictionary example sentences.

Future work could include a version of the system aware of word senses, both as search terms and as entries in the word lists applied. This would also enable searching for sentences belonging to specific topics or domains. Moreover, additional information about learners' lexical knowledge could be incorporated, for example, based on learner-written essays. Another valuable direction of further research would be the extension of the algorithm to multiple languages, for example through the use of universal POS and dependency tags. Finally, collecting additional data on how learners perform on the exercises constructed out of the selected sentences could also provide further indication on the quality and usefulness of the proposed algorithm.

Part III

Studies on learner text evaluation

10 Predicting proficiency levels in learner writings through domain transfer

This chapter is a postprint version of the following publication:

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 the* 26^{th} *International Conference on Computational Linguistics (COLING)*, 2101–2110.

Abstract. The lack of a sufficient amount of data tailored for a task is a well-recognized problem for many statistical NLP methods. In this paper, we explore whether data sparsity can be successfully tackled when classifying language proficiency levels in the domain of learner-written output texts. We aim at overcoming data sparsity by incorporating knowledge in the trained model from another domain consisting of input texts written by teaching professionals for learners. We compare different domain adaptation techniques and find that a weighted combination of the two types of data performs best, which can even rival systems based on considerably larger amounts of in-domain data. Moreover, we show that normalizing errors in learners' texts can substantially improve classification when in-domain data with annotated proficiency levels is not available.

10.1 Introduction

Data sparsity is a recognized problem in many machine learning based NLP approaches since the creation of data specifically collected and annotated for a certain task or language is time-consuming and costly. Previous attempts to overcome data sparsity include transferring knowledge between different types of data through the application of models from languages and tasks where sufficient data exists to the ones where data is unavailable or sparse (Daumé III and Marcu 2006). A common case of such a transfer learning scenario is *domain adaptation*, where training and test data belong to different domains (e.g. text genres) referred to as *source domain* and *target domain* respectively.

In our experiments, we aim at exploring the plausibility of domain adaptation as a strategy for overcoming data sparsity in the context of foreign and second language (L2) learning. More specifically, we operationalize *domain* as the type of text involved in the language learning process: on the one hand, texts from coursebooks intended for L2 learners (referred to as *L2 input texts* in this paper), and on the other hand, essays created by learners (*L2 output texts*). Our goal is to predict L2 language development stages in terms of linguistic complexity in the latter category, i.e. learner-produced texts. These stages are commonly referred to as *proficiency levels* in second language acquisition and language testing. Levels range from "absolute beginner" to "advanced language user" with increasing linguistic complexity as learners progress with the levels. A scale of such levels, very influential both in Europe and outside, is the CEFR – Common European Framework of Reference for Languages (Council of Europe 2001).

In previous work, NLP methods have been successfully applied to both assessing proficiency levels in L2 input texts collected from coursebooks and output texts written by learners (see section 10.2). However, the two text types have always been considered separately, while we argue that there is a shared linguistic content between the two that can be used for knowledge transfer. Specifically, the output of learners is a subset of the linguistic input that they are able to understand (Barrot 2015). Thus, incorporating knowledge from coursebook texts representing L2 input may improve the classification of proficiency levels in L2 output text. Decreasing the need for a large amount of L2 output data is particularly appealing since acquiring this type of text poses a number of challenges including copyright issues, anonymization of sensitive information, and often even digitizing hand-written material (Megyesi, Näsman and Palmér 2016; Mendes et al. 2016; Volodina et al. 2016c). Since an increasing amount of people learn foreign languages worldwide either out of necessity or as a personal interest, systems targeting the needs of this user group are especially valuable. Within this context, the automatic assessment of proficiency levels in learner-produced texts would be a powerful tool for increasing both learners' autonomy and teaching professionals' efficiency.

10.1.1 Research questions

In particular, this paper aims at answering the following research questions: (i) Can we overcome the lack of a sufficient amount of learner output data by incorporating knowledge from L2 input texts when performing proficiency level classification? (ii) What kind of domain adaptation technique performs best in this context? (iii) Does normalizing errors in L2 learner output benefit proficiency level classification in a domain adaptation setting?

The motivation behind error normalization is that learner output typically contains errors which may influence the performance of automatic taggers and parsers and thus, classification performance. Therefore, error normalization may allow for a more precise calculation of feature values and a more successful transfer from and to a non error-prone domain. The amount and type of errors, i.e. degree of incorrectness, however, is not explicitly considered as an indicator of proficiency for L2 learner output in our experiments in order to keep comparability with coursebook texts. Unlike linguistic complexity, incorrectness is not a relevant aspect for L2 input texts as these are authored by teaching professionals and are supposed to be relatively error-free examples of language use.

Our target language of choice is Swedish, a language considerably less resource-rich than English and for which a CEFR-level classification model of L2 learners' writing is not available yet, despite the clear need for breaking down CEFR descriptors into linguistic constituents that characterize proficiency levels for each individual language (Little 2011; North 2007).

10.1.2 Main findings

We find that, in the absence of annotated learner-written data, using a classification model trained only on coursebook texts is a viable alternative if learner errors are normalized. Furthermore, if a small amount of learner output data is available and it is combined with L2 input texts, it can even outperform a model trained only on the few in-domain instances, resulting in a prediction quality matching that of in-domain state-of-the-art systems for other languages. In a domain adaptation setting, normalizing learner errors proved to yield a substantial improvement for features based on token, character and sentence counts as well as for features based on the CEFR-level distribution of tokens.

10.2 Text categorization in the language learning context

The automated evaluation of learner output is primarily a text classification task which aims at determining the quality of writing and assigning an appropriate label from a given set, for example a score or grade on the continuum between pass-fail (*essay scoring*) or a level indicating learning progress (*proficiency level classification*). In a L2 learning scenario, a longer piece of learner-written text is a popular means to assess learners' proficiency level. The human assessment of learner output, however, is both time-consuming and prone to subjectivity. Different linguistic dimensions need to be taken into consideration usually requiring several iterations of re-reading and different factors may influence the decision, such as negative attitude to a learner, hunger, bad mood, and boredom. Therefore, the number of initiatives to complement (or even replace) human assessment with a more objective and more efficient supervised machine learning system has been increasing the past years, with essay grading (Burstein and Chodorow 2010) as an important application field.

10.2.1 Automatic essay scoring

Automatic essay scoring (AES) has been an active research area since 1990s, targeting mostly English (Burstein and Chodorow 2010; Miltsakaki and Kukich 2004; Page 2003). Recently, with the availability of annotated learner corpora for other languages, automatic essay grading has expanded to cover also other languages, e.g. German (Zesch, Wojatzki and Scholten-Akoun 2015) and Swedish (Östling et al. 2013), to name just a few.

In its nature, AES has mostly relied on machine learning approaches, exploring both supervised (Yannakoudakis, Briscoe and Medlock 2011) and unsupervised methods (Chen et al. 2010) with different degrees of success. Östling et al. (2013) have looked at Swedish upper secondary school essays, i.e. first language learner essays, and automatically assessed them in terms of a four-point scale of performance grades with an accuracy of 62%. The authors found that this result exceeded the agreement rate between two human assessors which was as low as 45.8% which might indicate that human-like performance is a rather uncertain goal. Linguistic parameters that have over time been presumed to be strong predictors of writing quality have varied from shallow ones like text and word length (Page 2003; Östling et al. 2013) to more sophisticated features using Latent Semantic Analysis (Landauer, Laham and Foltz 2003), cosine similarity (Attali and Burstein 2006), discourse structure and stylistic features (Attali and Burstein 2006).

10.2.2 Proficiency level classification

A closely related task to AES is classifying texts into L2 proficiency levels which consists of predicting at which language learning stage a text can be produced or understood by a L2 learner, rather than assigning a grade within a pass-fail range. The CEFR, the scale of proficiency levels adopted in our experiments, contains guidelines for the standardization of language teaching and assessment across languages and countries (Council of Europe 2001). It provides a common metalanguage to talk about objectives, assessment, (Little 2011), and it defines language competences at six proficiency levels (A1, A2, B1, B2, C1, C2) where A1 is the beginner level. Since the publication of the CEFR guidelines in 2001, several countries have adopted the system, but its practical application has proven to be rather non-straightforward since the descriptions of the competences at each level remain vague (Little 2011; North 2007).

The past few years have seen an increasing interest in the CEFR-level classification of both L2 input and output texts. In the case of coursebook texts such a classification has also been referred to as *L2 readability* and it has been investigated for, among others, French (François and Fairon 2012), Portuguese (Branco et al. 2014), Chinese (Sung et al. 2015), Swedish (chapter 7 based on Pilán, Vajjala and Volodina 2016), and English (Xia, Kochmar and Briscoe 2016).

Apart from L2 input texts, CEFR-level annotated L2 learner corpora are also available for a number of languages including but not limited to English (Nicholls 2003), Estonian (Vajjala and Lõo 2014) and German (Hancke and Meurers 2013). Moreover, MERLIN (Wisniewski et al. 2013) is a trilingual learner corpus comprised of written productions of L2 learners of Czech, German, and Italian also linked to CEFR levels. Despite the availability of annotated corpora for several languages, the number of projects targeting the automatic CEFR-level classification of learner essays has remained rather limited. Previously reported results for this task in terms of accuracy include 61% for German (Hancke and Meurers 2013) and 79% for Estonian (Vajjala and Lõo 2014).

10.2.3 Domain adaptation for tasks related to L2 learning

While there is a lot of previous work on domain adaptation in general, relatively few approaches exist in the field of assessing learner output texts. Previous applications of domain adaptation to learner essays focused on exploring the transfer of models between different writing tasks that prompted students to produce the essays, e.g. expressing an opinion on a topic vs. summarizing a news article (Zesch, Wojatzki and Scholten-Akoun 2015; Phandi, Chai and Ng 2015). Zesch, Wojatzki and Scholten-Akoun (2015) explore which features are transferable from one essay grading task to another task based on a different prompt. They find that by excluding some highly domain-specific features, the transfer loss can be reduced significantly without noticeable differences in overall performance.

A popular domain adaptation approach is EASYADAPT (Daumé III 2007) that augments the original feature space with source- and target-specific versions. Phandi, Chai and Ng (2015) successfully applied EASYADAPT for automatic essay scoring and Xia, Kochmar and Briscoe (2016) for the CEFR-level classification of L2 input texts with native language texts as source domain.

10.3 Datasets

For our experiments, we use L2 Swedish data including learners' output, i.e. error-prone essays written by learners, as well as L2 input data for learners, i.e. relatively error-free texts written by experts for L2 learners primarily intended as reading material. Both types of data are manually labeled for CEFR levels and automatically annotated across different linguistic dimensions including lemmatization, part-of-speech (POS) tagging, and dependency parsing using the Sparv (previously known as Korp) pipeline (Borin, Forsberg and Roxendal 2012).

10.3.1 L2 output texts

Our source of output texts is SweLL (Volodina et al. 2016c), a corpus consisting of L2 Swedish learner essays on a variety of topics, manually linked to CEFR levels. The essays also contain meta-information on learners' mother tongue(s), age, gender, education level, the exam setting, and, in certain cases, topic and genre. The distribution of essays per level is given in table 10.1.

The corpus includes some essays at A1 and C2 levels, but these classes were too under-represented to be included in our experiments. As for A1 level, this may depend on learners' limited ability to write due to the lack of familiarity with many linguistic constructs. In fact, the CEFR contains no descriptor for writing essays and reports at A1 level (Council of Europe 2001: 62). C2 is lacking since courses at this level are not provided, and it is in general characterized as a near-native language competence.

Since SweLL consists of learner-produced texts, it is likely that it contains some errors which, however, have not been annotated or normalized yet in the

		CEFR Levels				
		A2	**B1**	**B2**	**C1**	**Total**
Learner	**Texts**	83	75	74	88	**320**
	Tokens	18,349	29,814	32,691	60,095	**140,949**
Expert	**Texts**	157	258	288	115	**818**
	Tokens	37,168	79,124	101,297	71,723	**289,312**

Table 10.1: Overview of CEFR-level annotated Swedish datasets.

resource. The number of non-lemmatized tokens in the resource (i.e. tokens that could not be assigned baseforms during automatic annotation), which could indicate spelling errors or creative compounding at more advanced levels is higher at lower proficiency levels, but their amount always remains within a range of 5% and 8%.

10.3.2 L2 input texts

Our L2 input texts were collected from COCTAILL, a corpus of coursebooks used for teaching CEFR-based courses of L2 Swedish (Volodina et al. 2014b). The coursebooks are divided into lessons (book chapters), each of which is labeled for the CEFR level it is aimed at. Each lesson contains a variety of elements including reading texts, exercises, lists, etc. Out of these only the texts intended for reading have been included in our dataset, whose CEFR level was derived from the level of the lesson they occurred in. Table 10.1 gives an overview of the distribution of these texts per level. For the same reasons as in section 10.3.1, C2 was not included in this dataset and A1 level has been omitted to keep the classes consistent between the two datasets.

10.4 Feature set

We use the feature set presented in Pilán, Vajjala and Volodina (2016) (see also chapter 7) designed for modeling linguistic complexity in input texts for L2 Swedish learners. These features rely on morpho-syntactic tags, information about the CEFR level of tokens, and aspects inspired by L2 Swedish curricula. Five sub-group of features can be distinguished in this set: length-based, (weakly) lexical, morphological, syntactic, and semantic features. The detailed list of features is presented in table 10.2.

Count-based features rely on the number of characters and tokens (*tkn*), extra-long words being tokens longer than 13 characters. LIX (Läsbarhetsindex) is a traditional Swedish readability formula corresponding to the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters (Björnsson 1968). Rather than a simple type-token ratio (TTR), we use a bi-logarithmic and a square root equivalent following Vajjala and Meurers (2012).

Lexical features incorporate information from the KELLY list (Volodina and Kokkinakis 2012), a frequency-based word list compiled using a corpus of web texts (thus completely independent of our datasets), which also provides a suggested CEFR level per each lemma based on frequency bands. For some feature values, *incidence scores* (IS) are computed, in other words, instead of absolute counts, normalized values per 1000 tokens are considered to reduce the influence of sentence length. Lexical complexity is modeled with a set of weakly lexicalized features, i.e. we do not use word forms or lemmas themselves as features, but the IS of their corresponding CEFR levels instead. This aspect is especially important considering the limited size of our learner essay data. *Difficult* tokens are those that belong to levels above the overall CEFR level of the text. Moreover, we consider the IS of tokens not present in KELLY (OOV IS), the IS of tokens for which the lemmatizer could not identify a corresponding lemma (No lemma IS), as well as average KELLY log frequencies.

Morphological features include not only IS but also variational scores, i.e. the ratio of a category to the ratio of lexical tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). The IS of all lexical categories as well as the IS of punctuation, particles, sub- and conjunctions (SJ, CJ) are taken into consideration. Nominal ratio (Hultman and Westman 1977) is another readability formula proposed for Swedish that corresponds to the ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR) adverbs, and verbs. Relative structures consist of relative adverbs, determiners, pronouns and possessives. Some features are inspired by L2 teaching material (Fasth and Kannermark 1997) and they are based on fine-grained inflectional information such as the IS of neuter gender nouns and the ratio of different verb forms to all verbs.

Syntactic features are based, among others, on the length (depth) and the direction of dependency arcs (DepArc). Within this feature group, we consider also relative clauses as well as pre- and post-modifiers, which include, for example, adjectives and prepositional phrases respectively.

Semantic features build on information from the SALDO lexicon (Borin, Forsberg and Lönngren 2013). We use the average number of senses per token and the average number of noun senses per nouns.

Count	**Lexical**	**Syntactic**	**Morphological**	
Sentence length	A1 lemma IS	Avg DepArc length	Modal V to V	Verb IS
Avg token length	A2 lemma IS	DepArc Len > 5	Particle IS	V variation
Extra-long token	B1 lemma IS	Max length DepArc	3SG pronoun IS	Function W IS
Nr characters	B2 lemma IS	Right DepArc Ratio	Punctuation IS	Lex tkn to non-lex tkn
LIX	C1 lemma IS	Left DepArc Ratio	Subjunction IS	Lex tkn to Nr tkn
Bilog TTR	C2 lemma IS	Modifier variation	PR to N	Neuter N IS
Square root TTR	Difficult W IS	Pre-modifier IS	PR to PP	CJ + SJ IS
Semantic	Difficult N&V IS	Post-modifier IS	S-VB IS	Past PC to V
Avg senses per token	OOV IS	Subordinate IS	S-V to V	Present PC to V
N senses per N	No lemma IS	Relative clause IS	ADJ IS	Past V to V
	Avg. KELLY log freq	PP complement IS	ADJ variation	Present V to V
			ADV IS	Supine V to V
			ADV variation	Relative structure IS
			N IS	Nominal ratio
			N variation	N to V

Table 10.2: Feature set.

10.5 Experimental setup

For all experiments, we use SVMs as implemented in WEKA (Hall et al. 2009) and the feature set presented in detail in section 10.4. Results are obtained using 10-fold cross-validation. We report the F_1 score, i.e. the harmonic mean of precision and recall, as well as quadratic weighted kappa (κ^2), a distance-based scoring function taking into consideration also the degree of misclassifications.

10.5.1 Domain adaptation

In a domain adaptation scenario, data from a source domain (D_S) is used to predict labels in a different, target domain (D_T). To overcome data sparsity, especially relevant for our learner essay data, we experiment with improving CEFR level classification by transferring information from our D_S consisting of L2 coursebook texts to D_T consisting of Swedish L2 learners' essays.

As baselines, we employ both assigning the most frequent label in the dataset (MAJORITY) and an IN-DOMAIN setup using only the learner essays in a cross-validation setup. We compare these to different domain adaptation scenarios inspired mostly by Daumé III and Marcu (2006) and Pan and Yang (2010) which differ in the type and the amount of data used as detailed in table 10.3. We report the number of instances employed at the moment of training as well as the amount of instances from which information has been incorporated in some form in the final models.

In the **SOURCE-ONLY** setup, a model trained on all available source domain instances, i.e. coursebook texts, was applied directly to the target domain instances consisting of learner essays. **EASYADAPT** (Daumé III 2007) is a feature augmentation approach which consists of triplicating the feature space by including three versions of each feature in the augmented equivalent: a general, a source-specific and a target-specific version. In more formal terms, to each feature vector x, the mapping function $\phi^S(x) = \langle x, x, 0 \rangle$ is applied in the source domain and $\phi^T(x) = \langle x, 0, x \rangle$ in the target domain, 0 being a zero vector of length $|x|$. In **+FEATURE** we first train a model trained on the L2 input texts. Then, the CEFR label predicted by this system is incorporated as an additional feature for each essay instance and a new model is trained on the essays with this extra dimension. For **COMBINED** and **WEIGHTED** the training data includes not only D_S instances, but also 60% of D_T. In the WEIGHTED setup, an increased importance is given to D_T instances during training through the assignment of a higher weight (w). Finally, to obtain **WEIGHTED-INSTSEL**, we first train a model on the available D_T data and use that to classify D_S instances. Then those D_S instances that the essay-only

Experimental setup	Data used	# Training inst.	# Informing inst.
MAJORITY	D_T	288	320
IN-DOMAIN	D_T	288	320
SOURCE-ONLY	D_S	818	818
EASYADAPT	D_S and 60% of D_T with augmented features	1010	1010
+FEATURE	D_T with D_S prediction as feature	288	1138
COMBINED	D_S + 60% of D_T	1010	1010
WEIGHTED	D_S ($w = 1$) + 60% of D_T ($w = 10$)	1010	1010
WEIGHTED-INSTSEL	Correctly classified D_S ($w = 1$) + 60% of D_T ($w = 10$)	505	1138

Table 10.3: Domain adaptation experimental setups.

model correctly classified are combined with 60% D_T, the latter ones receiving a weight of 10. Compared to WEIGHTED, in this setup we discard D_S instances that might be misleading when making predictions on D_T, due to differences in the underlying distributions in the two domains. A similar approach is presented in Jiang and Zhai (2007).

10.5.2 Error normalization

Besides using learners' output texts in their original form, we investigate also the effects of error normalization on the domain-adapted strategies. By correcting errors we aim at bringing learners' texts closer to the standard language present in the coursebooks. Making the texts belonging to these two different domains more similar to each other may improve the domain-adapted classification performance. Moreover, since the annotation tools used were originally designed for dealing with standard Swedish, error normalization leads to a more reliable tagging and parsing, and hence to more precise feature values in the corrected learner output texts.

Previous error-normalization approaches include, among others, finite state transducers (Antonsen 2012) and a number of, mostly hybrid, systems created within the CoNLL Shared Task on grammatical error correction for L2 English (Ng et al. 2014).

We use LanguageTool[39] (Naber 2003), an open-source rule-based proofreading program available for multiple languages which detects not only spelling, but also some grammatical errors (e.g. inconsistent gender use in inflected forms). We propose a two-step algorithm consisting of first obtaining correction candidates from LanguageTool and then ranking these candidates based on a word co-occurrence measure. As a first step, we identify errors in the learner essays and a list of one or more LanguageTool correction suggestions, as well as the *context*, i.e. the surrounding tokens for the error within the same sentence. When more than one correction candidate is available, as an additional step, we make a selection based on *Lexicographers' Mutual Information* (LMI) scores (Kilgarriff et al. 2004). Here we assume a positive correlation between a correction candidate co-occurring with a context word and being the correct version of the word intended by the learner. We check LMI scores for each LanguageTool correction candidate paired with the lemma of each available noun, verb, and adjective in the context based on a pre-compiled list of LMI scores. We create this list using a Korp API (Borin, Forsberg and Roxendal 2012) providing LMI scores computed based on a customizable set of corpora. We use a variety of modern Swedish corpora totaling to more than

[39] `www.languagetool.org`

209 million tokens for our list of LMI scores. Only scores for noun-verb and noun-adjective combinations have been included with a threshold of LMI ≥ 50. When available, we select the correction candidate maximizing the sum of all LMI scores for the context words. In the absence of LMI scores for the pairs of correction candidates and context words, the most frequent word form in Swedish Wikipedia texts is chosen as a fallback.

Once correction candidates are ranked, each erroneous token identified by LanguageTool is replaced in the essays by the top ranked correction candidate. The normalized texts are then annotated again and feature values are re-computed.

10.6 Results and discussion

	ORIGINAL		ERROR-NORMALIZED	
	F_1	κ^2	F_1	κ^2
MAJORITY	.120	.000	.120	.000
IN-DOMAIN	.721	.886	.720	.872
SOURCE-ONLY	.438	.713	.620	.807
EASYADAPT	.503	.681	.533	.741
+FEATURE	.709	.879	**.802**	.864
COMBINED	.733	.863	.726	.885
WEIGHTED	**.747**	**.890**	.779	**.915**
WEIGHTED-INSTSEL	.733	.873	.795	.914

Table 10.4: Domain adaptation results with and without error normalization.

Table 10.4 presents the results of our domain adaptation experiments first without error normalization (*original*) and then with corrected errors (*error-normalized*). In the case of the non-normalized essays, the in-domain baseline obtained using only the small amount of learner output texts in a cross-validation setup is .721 F_1 and .886 κ^2. Compared to this, transferring a model based on coursebook texts directly (SOURCE-ONLY) results in a considerable performance drop (-.283 F_1 and -.173 κ^2). When using the essays in their original, noisy form, the best performing domain adaptation setup is the weighted combination of L2 input and output texts, which outperforms even the in-domain baseline both in terms of F_1 and κ^2.

The obtained domain-adaptation results are comparable to state-of-the-art in-domain systems for other languages, like the system for Estonian described in Vajjala and Lõo (2014) with an F_1 of .78, or the one for German (Hancke 2013) with .71 F_1 for a feature selected model distinguishing 5 classes. It is worth noting, however, that both of these systems required a considerably (about three times) larger annotated in-domain corpus. This shows that additional coursebook data can benefit the classification of language proficiency levels in learner output texts, especially if only a small amount of annotated in-domain data is available.

10.6.1 Error normalization

Our error-normalization method corrects in total 5,080 errors in the essays which amounts to 3.6% of all tokens in the data. In absence of error-annotated Swedish resources, we manually evaluate the method by inspecting 120 normalized items out of which we find 83 correct, corresponding to 69% accuracy. Out of the normalized tokens, about 87% are categorized as spelling errors by LanguageTool. Moreover, the choice of correction candidate is based on LMI scores in 24% of all cases.

Since our feature set does not target learner errors specifically (to be able to maintain comparability when applied to coursebook text), we do not expect error normalization to influence classification results with IN-DOMAIN. Our experiment results in table 10.4 show, in fact, that correcting learner errors does not have any statistically significant effect in the IN-DOMAIN setup, but it does improve performance to a great extent for most domain-adapted cases. This latter would support the hypothesis that correcting spelling and grammatical errors increases the similarity between the target and the source domain. The gain is especially large (+.182 F_1) in the case of the SOURCEONLY setup, which does not rely on annotated essays. EASYADAPT, which has been successfully used in an AES task previously (Phandi, Chai and Ng 2015), is outperformed by most other domain adaptation methods in our case, independently from error normalization.

In terms of F_1, +FEATURE using the predictions of a classifier trained on the L2 input texts performs best (.802 F_1), however, the degree of misclassifications indicated by κ^2 is smallest with WEIGHTED (.915), as in the case of the essays without error normalization. After error correction, WEIGHTED-INSTSEL achieves approximately the same quality of performance for all measures as the aforementioned two best performing models WEIGHTED and +FEATURE. These all improve over the IN-DOMAIN baseline by about .07 F_1 and .03 κ^2.

These results show that the knowledge transfer from L2 input texts can be substantially boosted by normalizing errors in the learner-produced texts.

10.6.2 Contribution of feature groups

In the next step, we investigate the contribution of individual feature groups to the classification performance both in- and cross-domain with the Source-Only setup which does not presuppose the availability of annotated in-domain data. Results for our ablation test are shown in table 10.5.

	In-domain		**Source-Only (Original)**		**Source-Only (Error-norm.)**	
Feature Group	F_1	κ^2	F_1	κ^2	F_1	κ^2
All	.721	.886	.438	.713	.620	.807
Count	.499	.740	.106	-.003	*.335*	***.708***
Lexical	**.625**	**.826**	**.318**	**.507**	**.378**	.626
Syntactic	.511	.665	.118	.066	.106	.030
Morphological	.538	.743	.297	.403	.291	.419
Semantic	.299	.198	.087	.000	.087	.000

Table 10.5: Performance of individual feature groups.

The most predictive features in- and cross-domain on both the original and on the normalized essays are lexical features measuring the proportion of tokens per CEFR level in the texts. Morphological features also preserve their strong predictive power when transferred between L2 input and output texts. The informativeness of syntactic and count features is very low in the cross-domain setting with the original essays, but the latter category transfers much better after error-normalizing L2 output texts. A potential explanation could be that error normalization includes also corrections of capitalization and whitespaces which might contribute to an improved detection of sentence boundaries, a central element in most of these features. Lexical features also benefit from error correction, presumably due to a more precise estimation of the CEFR-level distribution of tokens.

10.6.3 Direction of misclassifications

Finally, to investigate whether the transferred coursebook model predicts learner-written texts to be of higher or lower proficiency levels compared to the available annotations, we perform regression using SMO and the SourceOnly setup, transforming CEFR levels into numeric values. We use the normalized essays for this purpose since the automatic annotation is presumably more precise in these texts compared to their original version. Predictions within a distance of 0.5 from the numeric value representing the actual CEFR level are considered sufficiently close for being considered correct, thus the amount of errors is computed based only on cases exceeding this margin. The regression model produces .800 correlation and 1.120 RMSE (root mean squared error). We find that 64% of the erroneous predictions consider essays to be of a lower level than they actually are. This could be a data-driven confirmation of the pedagogical observation that learners' output texts are typically of a lower linguistic complexity compared to the L2 input texts written for them within the same CEFR level.

10.7 Conclusions

In this work we investigated the benefits of using texts from language learning coursebooks to classify proficiency levels in learner-written texts, since the latter type of data is especially costly to collect. Moreover, our experiments provide useful insights into how some simple domain adaptation techniques compare to each other for this task. Training only on source domain data did not yield a successfully transferable model between the L2 input and output texts if errors were not normalized in the learner-produced essays. With such a normalization, however, using only coursebook texts as training data produced a result rather close to what learning only from a small amount of essays did. Joining domains was useful, especially when weighted target domain instances were added to all, or a subset of the coursebook data, and learner errors were normalized. We showed that, with these two steps, it is possible to outperform a model based only on a limited amount of in-domain data. Furthermore, our results are competitive even compared to systems for other languages that make use of a considerably larger amount of in-domain data.

In the future, it would be informative to repeat the experiments for other languages, where we expect similar results. Additional domain adaptation techniques could also be explored for this task, for example, the identification of shared priors and kernel transformations. Alternatives to the current error normalization could be investigated in order to identify a broader range of incorrect tokens more precisely. More reliable error correction methods may yield further improvement to transferring classification models between these domains.

11 COURSEBOOK-BASED LEXICAL FEATURES FOR LEARNER WRITING EVALUATION

This chapter is a postprint version of the following publication:

Ildikó Pilán, Elena Volodina and David Alfter 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. *Proceedings of the COLING 2016 workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 120–126.

Abstract. We bring together knowledge from two different types of language learning data, texts learners read and texts they write, to improve linguistic complexity classification in the latter. Linguistic complexity in the foreign and second language learning context can be expressed in terms of proficiency levels. We show that incorporating features capturing lexical complexity information from reading passages can boost significantly the machine learning based classification of learner-written texts into proficiency levels. With an F_1 score of .8 our system rivals state-of-the-art results reported for other languages for this task. Finally, we present a freely available web-based tool for proficiency level classification and lexical complexity visualization for both learner writings and reading texts.

11.1 Introduction

Second or foreign (L2) language learners pass through different development stages commonly referred to as *proficiency levels*. A popular scale of such levels is the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). As learners advance to higher levels, the complexity of the linguistic input that they are able to comprehend (*receptive* skills) and the output that they produce (*productive* skills) increases in terms of both lexical and grammatical patterns. Although learners' receptive and productive knowledge overlap, they only do so partially, the latter being typically a subset of the former corresponding to a somewhat lower linguistic complexity overall (Barrot 2015).

In previous work, NLP methods have been successfully applied for assessing separately receptive and productive L2 levels (see section 11.2). We, on the other hand, hypothesize that, since a shared linguistic content exists between what L2 learners are exposed to (*L2 input texts*, e.g. reading passages from coursebooks) and what they produce (*L2 output texts*, e.g. essays), transferring knowledge from one text type may improve the classification of linguistic complexity levels in the other. We focus on the automatic prediction of CEFR levels for L2 learner essays for a number of reasons. Essay writing is a popular means to assess learners' proficiency level and it is a rather subjective and time-consuming task. Moreover, such data is rather scarce and cumbersome to collect (Volodina et al. 2016c). Our target language is Swedish since corpora for both L2 text types are available for this language.

We compare two different strategies aiming at improving L2 essay classification results without additional data of this type: (i) employing a word list based on a coursebook corpus for lexical features, (ii) *domain adaptation* experiments, i.e. training a machine learning model on L2 input texts and using it to classify the essays. We first compare the distribution of words per CEFR levels in the essays using two different word lists and find that a list based on L2 input texts correlates well with the manually assigned CEFR labels of the essays. Using this list in machine learning experiments produces a significant performance boost which exceeds our domain adaptation attempts and compares well also to previously reported results for this task. Finally, we present an online tool for assessing linguistic complexity in L2 Swedish input and output texts that performs a machine learning based CEFR level classification and a lexical complexity analysis supported by a color-enhanced visualization of words per level.

11.2 Background

Recently a number of attempts emerged at the classification of CEFR levels in input texts (also known as *L2 readability*) which include, among others, systems for French (François and Fairon 2012), Portuguese (Branco et al. 2014), Chinese (Sung et al. 2015), Swedish (see chapter 7 based on Pilán, Vajjala and Volodina 2016), and English (Xia, Kochmar and Briscoe 2016). The same type of classification for learner-written texts remains somewhat less explored. Investigations include Vajjala and Lõo (2014) for Estonian and Hancke (2013) for German reporting an F_1 score of .78 and .71 respectively. The systems above are based on supervised learning methods based on rich feature sets.

Relatively few studies exist in the field of assessing the complexity and quality of L2 texts with the use of domain adaptation. Experiments relying on such methods have been explored so far for transferring essay grading models between writing tasks based on different prompts (Zesch, Wojatzki and Scholten-Akoun 2015; Phandi, Chai and Ng 2015), and for L2 readability classification by transferring models trained on texts written for native language users to reading passages aimed at L2 learners (Xia, Kochmar and Briscoe 2016).

11.3 Receptive and productive L2 Swedish corpora

Two corpora with L2 focus are currently available for Swedish: SweLL (Volodina et al. 2016c), comprised of L2 output texts in the form of learner essays, and COCTAILL (Volodina et al. 2014b) containing L2 coursebooks written by experts for L2 learners. The essays in the **SweLL** corpus were written by adult L2 Swedish learners (with available metadata) and they address a variety of topics. In the case of the coursebook corpus, **COCTAILL**, instead of using it in its entirety, we only include reading passages in our dataset. Other coursebook components whose linguistic annotation may be less reliable (e.g. gapped exercises) are excluded. We derive the CEFR level of the reading texts from the level of the lesson (chapter) they occurr in. It is worth mentioning that these two corpora are *independent* from each other, i.e. the essays written by the learners are not based on, or inspired by, the reading passages.

Both corpora span beginner (A1) to advanced (C1) proficiency with texts manually labeled for CEFR levels, and automatically annotated across different linguistic dimensions. These include lemmatization, part-of-speech (POS) tagging and dependency parsing using the Sparv pipeline[40] (Borin, Forsberg and Roxendal 2012). Since A1 level is rather under-represented in both corpora, we exclude them from our experiments. The distribution of texts per type and CEFR level in our datasets is shown in table 11.1, where A2 corresponds to elementary level, B1 to intermediate, B2 to upper intermediate and C1 to advanced level.

[40] `https://spraakbanken.gu.se/sparv/`

Writer	Unit	A2	B1	B2	C1	Total
Learner	**Texts**	83	75	74	88	**320**
	Tokens	18,349	29,814	32,691	60,095	**140,949**
Expert	**Texts**	157	258	288	115	**818**
	Tokens	37,168	79,124	101,297	71,723	**289,312**

Table 11.1: Overview of CEFR-level annotated Swedish datasets.

11.4 L2 lexical complexity: a comparison of word lists

KELLY (Volodina and Kokkinakis 2012) is a popular L2 Swedish word lists, compiled based on web corpora. It contains 8,425 headwords with not only frequency information, but also suggested CEFR levels based on normalized frequencies. The list has been successfully applied previously in machine learning experiment for classifying CEFR levels in L2 input texts (chapter 7 based on Pilán, Vajjala and Volodina 2016).

SVALex (François et al. 2016) is another Swedish word list with an L2 focus, created recently. The list contains word frequencies based on reading passages from COCTAILL (see section 11.3), which, however, are not connected to suggested CEFR levels. Therefore, we propose an enhanced version of this list, **SVALex+**, that includes mappings from frequency distributions to a single CEFR label following the methodology described in Alfter et al. (2016). To create the mappings, as a first step, frequency counts are normalized. Part of this consists of taking the raw frequency counts from SVALex and calculating *per-million-word* (PMW) frequency distributions for all words. These distributions are complemented with *word diversity* distributions, i.e. information about how often a word is used in different coursebooks at each level in the COCTAILL corpus. The intuition is that, if a word is used often at a certain level, but only in one book, it is less representative of a level than if it appears in several coursebooks. We then combine these two distributions into one single *normalized frequency* ($Freq^n$) value for each word by taking the average of the PMW frequency distribution and the word diversity distribution.

The second step consists of mapping these normalized frequencies to CEFR levels. Rather than mapping to the CEFR level at which a word first appears, we establish a *significant onset of use*, a threshold indicating a difference between normalized frequency distributions that is sufficiently large for a level to qualify as mapping for a word. We set this threshold to 0.4 based on initial empirical investigations with L2 teachers during which the overlap between teacher- and system-assigned levels for a small subset of words have been compared. Thus, we map each word to the lowest CEFR level L for which $Freq_L^n - Freq_{L-1}^n > 0.4$ holds, with $L-1$ being the previous CEFR level and $Freq_{L-1}^n = 0$ if $L = \text{A1}$.

Table 11.2 compares the percentage of tokens belonging to different CEFR levels based on KELLY and SVALex+ (rows) per essay CEFR level (columns).

	Essay CEFR levels									
	KELLY					**SVALex+**				
	A1	**A2**	**B1**	**B2**	**C1**	**A1**	**A2**	**B1**	**B2**	**C1**
A1	69.0	72.91	72.56	73.3	70.91	74.28	77.86	**65.09**	**61.96**	**56.92**
A2	4.08	3.96	4.18	4.31	5.22	2.54	3.6	**8.01**	**9.31**	**10.67**
B1	1.2	1.79	1.52	2.4	3.16	1.73	2.91	**9.82**	**12.59**	**14.28**
B2	.67	.68	1.17	.83	1.11	.43	.66	.86	1.07	1.86
C1	.43	.31	.31	.4	.5	.14	.17	.33	.48	.81

Table 11.2: Distribution of token CEFR levels (in %) per essay CEFR levels.

We can observe that the distribution of tokens per CEFR level based on KELLY remains rather unchanged: A1 and C2 level essays contain, for instance, approximately the same amount of A1-C1 tokens. SVALex+, on the other hand, correlates better with the overall CEFR level of the essays. The highlighted cells show a decrease of lower level tokens in higher level essays and an increase of higher level tokens in more advanced essays. This would suggest that using SVALex+ may improve CEFR level classification performance for learner essays. The amount of B2 and C1 tokens seems still rather limited even at higher levels which can be explained to some extent by SVALex+ containing receptive vocabulary that learners might not be able to use productively.

11.5 Essay classification experiments

11.5.1 Feature set

For modeling linguistic complexity in L2 Swedish texts, we use the feature set that we described in chapters 7 and 10 based on Pilán, Vajjala and Volodina (2016) and Pilán, Volodina and Zesch (2016). The 61 features of this set can be divided into five sub-groups: *length-based* (e.g. average sentence and token length), *lexical* (e.g. amount of tokens per CEFR level), *morphological* (e.g. past verbs to verbs ratio), *syntactic* (e.g. average dependency length) and *semantic* features (e.g. number of senses). For a more detailed description of the feature set see the cited works.

11.5.2 Experimental setup

We use the sequential minimal optimization algorithm from WEKA (Hall et al. 2009) and the feature set mentioned above for all experiments. Results are obtained using 10-fold cross-validation, unless otherwise specified. Reported measures include F_1 and quadratic weighted kappa (κ^2), a distance-based scoring function taking into consideration also the degree of misclassifications. Our baselines consist of assigning the most frequent label in the dataset to each instance (MAJORITY) and cross-validated results on the learner essays using KELLY (E-KELLY) for lexical features.

We compare these to two models using information from SVALex+ (E-SVALEX+ with SVALex+ instead of KELLY and E-KELLY&SVALEX+ including both lists), as well as to two simple domain adaptation setups inspired by Daumé III and Marcu (2006). In a domain adaptation scenario, data from a source domain is used to predict labels in a different, target domain. In our SOURCE-ONLY setup, a model trained on coursebook texts is applied to the essays, our target domain. In +FEATURE the CEFR levels predicted by a model trained on coursebook texts is used as an additional feature when training a classifier for the essays. For both the SOURCE-ONLY and the +FEATURE setup the KELLY list has been used.

11.5.3 Classification results

The results of our experiments are presented in table 11.3.

Substituting KELLY-based features with their SVALex+ based equivalents increases classification performance substantially, from .721 to .822 in terms of F_1. This is most likely connected to the fact that word frequencies based on the general (web) corpus, KELLY, reflect less precisely learners' progression in terms of lexical complexity compared to SVALex+, which is based on texts explicitly intended for L2 learners (see table 11.2). Combining both KELLY and SVALex+ achieves a slight gain, but the performance difference remains rather negligible compared to using SVALex+ alone. The high κ^2 values for the SVALex+ based models indicate that very few misclassifications occur with a distance of more than one CEFR level. By inspecting the confusion matrices we find that only two instances fall into this category for the E-SVALex+ model, and none for E-KELLY&SVALex+.

Applying a coursebook model to the essays (SOURCE-ONLY) results in a radical performance drop compared to the in-domain models, which indicates that the distribution of feature values in L2 input and output texts differ to a rather large extent. For the same reason, adding the output of a coursebook

Essays (baselines)			Essays (using SVALex+)			Coursebooks → Essays		
	F_1	κ^2		F_1	κ^2		F_1	κ^2
MAJORITY	.120	.000	E-SVALEX+	**.808**	**.922**	SOURCE-ONLY	.438	.713
E-KELLY	.721	.886	E-KELLY&SVALEX+	**.816**	**.930**	+FEATURE	.709	.879

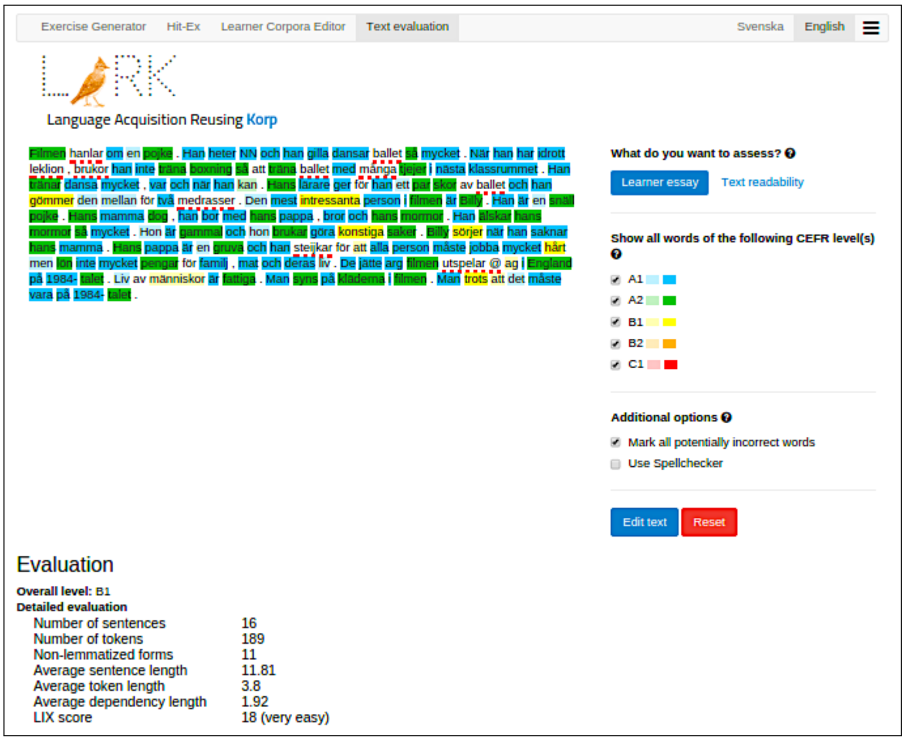
Table 11.3: Results for different classification improvement strategies.

based classifier (+FEATURE) performs less accurately than the E-KELLY baseline. These results, however, do not exclude the possibility of a successful transfer between these domains. Additional domain adaptation techniques may be able to bridge the gap between the source and target domain distributions.

Our SVALex+ based models achieve state-of-the-art performance compared to CEFR level classification systems for other languages such as the German system with .71 F_1 from Hancke (2013), and the Estonian one with .78 F_1 by Vajjala and Lõo (2014). Both of these systems, however, were built using an approximately three times larger annotated in-domain corpus.

11.6 An online tool for L2 linguistic complexity analysis

To put our L2 linguistic complexity analysis methods to practical use, we have made them available as a free online tool.[41] Figure 11.1 shows the web interface of the current version of our system.



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 . Den mest intressanta person i filmen är Billy . Han är en snäll
pojke . Hans mamma dog , han bor med hans pappa , bror och hans mormor . Han älskar hans
mormor så mycket . Hon är gammal och hon brukar göra konstiga saker . Billy sörjer när han saknar
hans mamma . Hans pappa är en gruva och han steijkar för att alla person måste jobba mycket hårt
men lön inte mycket pengar för familj , mat och deras liv . De jätte arg filmen utspelar @ ag i England
på 1984- talet . Liv av människor är fattiga . Man syns på kläderna i filmen . Man trots att det måste
vara på 1984- talet .

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

Evaluation

Overall level: B1

Detailed evaluation

Number of sentences	16
Number of tokens	189
Non-lemmatized forms	11
Average sentence length	11.81
Average token length	3.8
Average dependency length	1.92
LIX score	18 (very easy)

Figure 11.1: The interface for linguistic complexity analysis

[41] `https://spraakbanken.gu.se/larkalabb/texteval`

Users can type or paste a text in a text box and indicate whether the text was written by experts as reading material ("Text readability") or by learners ("Learner essay"). The text is then automatically analyzed in several steps. First, it undergoes an automatic linguistic annotation with Sparv (e.g. POS tags, dependency relations). Then the annotated text is fed to a machine learning algorithm based on the feature set described in section 11.5 that assesses the overall linguistic complexity of the text provided in terms of CEFR levels. Some simple statistics and values for traditional readability measures (e.g. average token length, LIX (Björnsson 1968)) are also included in the final results at the bottom of the page.

In addition to an overall assessment, a detailed visual L2 lexical complexity analysis can be performed. Users can highlight words of different CEFR levels in their text by ticking one (or more) of the check boxes in the right-side menu. The visualization highlights receptive and productive vocabulary items within the same CEFR level with the darker and lighter shade of the same color respectively. The highlighting is based on information from two vocabulary list: SVALex+ for receptive vocabulary and a word list based on SweLL (Alfter et al. 2016), for productive vocabulary, created using the mapping approach described in section 11.4 for SVALex+.

11.7 Conclusions

We described an exploration of different methods to improve the classification of texts produced by L2 Swedish learners into proficiency levels reflecting L2 linguistic complexity. By incorporating information from coursebooks in the form of lexical features indicating the distribution of CEFR levels per token in the texts, we created a system that reaches state-of-the-art performance reported for other languages for this task. Finally, we presented an online tool for linguistic complexity analysis of L2 texts. In the future, additional domain adaptation techniques could be tested for these text types and the effects of incorporating a learner essay based word list on the classification of L2 input texts could also be investigated.

Part IV

Cross-dataset experiments for feature selection

12 The importance of individual linguistic complexity features

This chapter is based on the following article in submission:

Pilán, Ildikó and Elena Volodina. Investigating the importance of linguistic complexity features across different datasets related to language learning. Submitted.

Abstract. We present the results of our investigations aiming at identifying the most informative linguistic complexity features for classifying language learning levels in three different datasets. The datasets vary across two dimensions: the size of the instances (texts vs. sentences) and the language learning skill they involve: reading comprehension texts vs. texts written by learners themselves. We present a subset of the most predictive features for each datasets, taking into consideration significant differences in their per-class mean values and show that these subsets lead not only to simpler models, but also to an improved classification performance. Furthermore, we pinpoint fourteen central features that are good predictors of learning levels regardless of the size of the linguistic unit analyzed or the type of skill involved which include both morpho-syntactic and lexical dimensions.

12.1 Introduction

The increase in international mobility for work, leisure or necessity has resulted in a large number of language learners world-wide (Castles, De Haas and Miller 2013). Exposure to a sufficient amount of comprehensible input has an important role in the learning process (Council of Europe 2001) and mastering the language of a host country is indispensable for a successful societal integration and accessing crucial information about healthcare, rights, rules and regulations.

We believe that the automatic analysis of linguistic complexity can be a useful means for determining whether non-native speakers can understand or produce certain linguistic input in a second or foreign language (L2) at different learning levels.

Linguistic complexity, especially in cross-linguistic studies, is often approached in absolute terms, describing complexity as a property of a linguistic system in terms of e.g. number of contrastive sounds (Moran and Blasi 2014). In this paper, however, we investigate a *relative* type of linguistic complexity from a cognitive perspective, our focus being the ability of L2 learners to process or produce certain linguistic elements in writing at different stages of proficiency. We operationalize the term *linguistic complexity* as the set of lexico-semantic, morphological and syntactic characteristics reflected in texts (or sentences) that determine the magnitude of the language skills and competences required to process or produce them. The scale of learning (*proficiency*) levels adopted in this work is the CEFR, the Common European Framework of Reference for Languages (Council of Europe 2001). The CEFR offers a common ground for language learning and assessment and it proposes a six-point scale of proficiency levels: from A1 (beginner) to C2 (advanced) level.

Large corpora in the language learning domain are rather scarce due to either copy-right issues, privacy reasons or the need for digitizing them. For the Swedish language, a number of resources have become available recently (Volodina et al. 2014b, 2016c), which, although somewhat small in size, encompass texts involving different skills and CEFR levels. This allows for investigations about the similarities and differences between linguistic complexity observable at different proficiency levels for different skill types, namely *receptive* skills, required when learners process passages produced by others and *productive* skills, when learners produce the texts themselves. We perform linguistic complexity analyses across two different dimensions: the type of learner skills involved when dealing with the texts and the size of the linguistic context investigated. In the latter case, we carry out experiments both at the text and at the sentence level. Throughout the years, a large number of linguistic features related to complexity has been proposed. Typically, out of the features suggested for a specific task some are more useful than others. Eliminating redundant features can result in simpler and improved models that are not only faster, but might also generalize better on unseen data (Witten et al. 2011: 308). In this paper, we investigate therefore the importance of individual linguistic complexity features for predicting proficiency levels across different L2 datasets. The two main research questions we investigate are: (i) Which linguistic complexity features are most useful for determining proficiency levels for each L2 dataset? (ii) Are there features that are relevant regardless of the context size and the type of skill considered? Our contributions include, on the one hand, a subset of the most informative features for each dataset whose use leads to improved classification results. On the other hand, we identify some lexical, morphological and syntactic features that are good indicators of complexity across all three datasets, namely, reading comprehension texts, essays and sentences.

In Section 12.2, we provide an overview of previous work related to linguistic complexity analysis, followed by the description of our datasets in Section 12.3. In Section 12.4, we present the set of features used and highlight their relevance for modeling linguistic complexity in the L2 context. We then describe our experiments and their results in Section 12.5, presenting the most informative features and their effect on classification performance. Finally, we conclude our results and outline future work in Section 12.6.

12.2 Previous literature on linguistic complexity for predicting L2 levels

Linguistic complexity analysis can be used for predicting both readability levels and proficiency (CEFR) levels. Readability analysis and proficiency level classification focus on different type of language users and skills. The former typically targets reading skills of native (L1) speakers with low reading levels or cognitive impairment, while proficiency level analysis is employed to assess a variety of skills for L2 speakers. Nevertheless, part of the linguistic features and the proposed approaches such as machine learning techniques (Collins-Thompson and Callan 2004; Heimann Mühlenbock 2013) for these two tasks are shared. Although both readability and scales of proficiency levels include also a number of additional aspects, some criteria connected to linguistic complexity heavily underlies both (Dale and Chall 1949; Council of Europe 2001) and it is the one aspect that most NLP systems providing such analyses explicitly or implicitly capture. In this section, we provide an overview of previous work related to applying linguistic complexity analysis in the L2 context.

12.2.1 Expert-written texts targeting receptive skills

In the L2 context, specific scales reflecting progress in language proficiency have been proposed. One such scale is the CEFR, introduced in section 1. An alternative to the CEFR is the 7-point scale of the Interagency Language Roundtable (ILR), common in the United States.

In table 12.1, we provide an overview of studies targeting L2 receptive complexity and compare the target language, the type and amount of training data and the methods used. The studies are ordered alphabetically based on the target language of the linguistic complexity analysis. We only include previous work here that shares the following characteristics: (i) texts rather than single sentences are the unit of analysis; (ii) receptive linguistic complexity is measured; and (iii) NLP tools are combined with machine learning algorithms. Under dataset size, we report the number of texts used (except for Heilman et al. (2007)), where whole books were employed), followed by the number of tokens in parenthesis when available. In some cases, the corpus used for the experiments was collected from L2 coursebooks and exams (François and Fairon 2012; Karpov, Baranova and Vitugin 2014; Xia, Kochmar and Briscoe 2016; Pilán, Vajjala and Volodina 2016). Other studies used authentic texts written primarily for L1 readers, which then were graded either by teaching professionals (Salesky and Shen 2014; Sung et al. 2015) or by L2 learners (Zhang, Liu and Ni 2013).

CEFR-based studies have been more commonly treated as a classification problem, a popular choice of classifier being support vector machines (SVM). A particular aspect distinguishing Xia, Kochmar and Briscoe (2016) from the rest of the studies mentioned in table 12.1 is the idea of using L1 data to improve the classification of L2 texts. For the sake of comparability, the information in table 12.1 describes only the experiments using the L2 data reported in this study. The state-of-the-art performance reported for the CEFR-based classification described in the studies included in table 12.1 ranges between 75% and 80% accuracy (Curto, Mamede and Baptista 2015; Sung et al. 2015; Xia, Kochmar and Briscoe 2016; Pilán, Vajjala and Volodina 2016).

A large number of features have been proposed and tested in this context. Count-based measures (e.g. sentence and token length, type-token ratio) and syntactic features (e.g. dependency length) have been confirmed to be influencing factors in L2 complexity (Curto, Mamede and Baptista 2015; Reynolds 2016). Lexical information based on either n-gram models (Heilman et al. 2007) or frequency information from word lists (François and Fairon 2012; Reynolds 2016) and Google search results (Huang et al. 2011) has proven to be, however, one of the most predictive dimensions. Beinborn, Zesch and Gurevych (2014b) offer an in-depth investigation of the role of lexical features in L2 complexity and propose taking into consideration cognates. Heilman et al. (2007) found that lexical features outperform grammatical ones, which, although more important for L2 than L1 complexity, still remain less predictive for L2 English complexity. Nevertheless, the authors mention that this may depend on the morphological richness of a language. Reynolds (2016), in fact, finds that morphological features are among the most influential ones for L2 Russian texts.

Study	**Target language**	**CEFR**	**Dataset size in # texts**	**Text type**	**# levels**	**Method**
Salesky and Shen (2014)	Arabic, Dari English, Pashto	No	4 × 1400	Non-L2	7	Regression
Sung et al. (2015)	Chinese	Yes	1578	L2	6	Classification
Heilman et al. (2007)	English	No	4 books (200,000)	L2	4	Regression
Huang et al. (2011)	English	No	187	Both	6	Regression
Xia et al. (2016)	English	Yes	331	L2	5 (A2-C2)	Both
Zhang et al. (2013)	English	No	15	Non-L2	1-10	Regression
François and Fairon (2012)	French	Yes	1852 (510,543)	L2	6	Classification
Branco et al. (2014)	Portuguese	Yes	110 (12,673)	L2	5 (A1-C1)	Regression
Curto et al. (2015)	Portuguese	Yes	237 (25,888)	L2	5 (A1-C1)	Classification
Karpov et al. (2014)	Russian	Yes	219	Both	4 (A1-B1, C2)	Classification
Reynolds (2016)	Russian	Yes	4689	Both	6	Classification
Pilán et al. (2016)	Swedish	Yes	867	L2	5 (A1-C1)	Both

Table 12.1: An overview of studies on L2 receptive complexity.

12.2.2 Learner-written texts

Similarly to L2 texts targeting reading skills, also texts produced by L2 learners manifest varying degrees of complexity at different stages of proficiency. Typically however, receptive linguistic complexity is somewhat higher than its productive counterpart for a learner at a given CEFR level (Barrot 2015). Previous studies aiming at classifying CEFR levels in learner-written texts include Hancke and Meurers (2013) for L2 German and Vajjala and Lõo (2014) for L2 Estonian. The most predictive features for L2 German include lexical and morphological features. Morphological features (e.g. amount of distinct cases used) are also among the most informative ones for L2 Estonian at all L2 development stages.

A fundamental difference between assessing receptive and productive texts is that, while receptive texts are expected to be relatively error free, the latter ones typically contain a varying amount of L2 errors, which have also been used to inform features. Errors are usually counted based on the output of a spell checker (Hancke and Meurers 2013; Tack et al. 2017) or with using hand-crafted rules (Tack et al. 2017). The reported state-of-the-art performance for CEFR-level classification in L2 learner texts in terms of accuracy is between 61% (Hancke and Meurers 2013) and 79% (Vajjala and Lõo 2014) accuracy.

12.2.3 Smaller linguistic units

Besides the text-level analyses in table 12.1, studies targeting smaller units also appear in the literature. Linguistic complexity in single sentences from an L2 perspective has been explored in Karpov, Baranova and Vitugin (2014) and in Pilán, Volodina and Johansson (2014). Both studies are CEFR-related, but rather than classifying sentences into individual CEFR levels, a binary distinction is made (below or at B1 level vs. above B1). Pilán, Vajjala and Volodina (2016) report 63% accuracy for a 5-way CEFR level classification of Swedish coursebook sentences. Another approach to linguistic complexity analysis focusing on smaller units is proposed in Ströbel et al. (2016). The authors present Cocogen (Complexity Contour Generator), a system analyzing complexity locally, within specific sliding window sizes, which build up a distribution of linguistic complexity (complexity contour) in the text. As for productive complexity, research on the automatic assessment of short answers to open-ended questions in terms of using CEFR has been investigated in Tack et al. (2017) for L2 English. The authors proposed an ensemble method consisting of integrating the votes of a number of traditional classification methods into a single prediction. Sentence and word length, lexical features and information about the age of acquisition of words were found especially predictive.

12.3 Datasets

12.3.1 Text-level datasets

We used two L2 Swedish corpora consisting of texts in our experiments: SweLL (Volodina et al. 2016c) comprised of essays written by L2 learners and COCTAILL (Volodina et al. 2014b) containing L2 coursebooks authored or adapted by experts for L2 learners. The SweLL corpus consists of essays produced by adult learners of L2 Swedish on a variety of topics (TEXT-E). From the coursebook corpus, we only include whole texts meant for reading comprehension practice (TEXT-R) since the linguistic annotation of other coursebook elements (e.g. gap-filling exercises) may be prone to automatic linguistic annotation errors. These two corpora cover five CEFR levels (A1 to C1). Each SweLL essay has been assigned a CEFR level by teachers. For reading texts, CEFR levels were derived from the level of the lesson (chapter) they occur in. It is worth mentioning that these two corpora are *independent* from each other, i.e. the essays written by the learners are not based on, or inspired by, the reading passages. The distribution of texts per type and CEFR level in the datasets is shown in table 12.2. The total number of tokens in the coursebook-based dataset was 289,312, while in the learner essay data it was 43,033.

12.3.2 A teacher-evaluated dataset of sentences

At the sentence level, we use a small dataset (SENT) based on a user evaluation of a corpus example selection system, HitEx, described in Pilán, Volodina and Borin (2017) and section 3.1.5 in this thesis. HitEx aims at identifying sentences from corpora suitable as exercise items. The sentences in this dataset have been automatically assessed for their CEFR level and have been filtered for their well-formedness, independence from the rest of their textual context and some additional lexical and structural criteria (e.g. abbreviations, interrogative form) using HitEx. Out of the original 330 sentences from the evaluation material, we only included in this dataset the subset of sentences: (i) that were found overall suitable (with an evaluation score $>=$ 2.5 out of 4); and (ii) where a majority of teachers agreed with the CEFR level assigned automatically by HitEx. This subset was complemented with 90 sentences for the otherwise insufficiently represented A1 level from the COCTAILL corpus. Only individually occurring sentences in lists and non-gapped exercises were considered, thus these are not a subset of the text-level dataset described above. The distribution of sentences per CEFR level in the dataset is presented in table 12.2. The total number of tokens in the dataset is 4,060.

Writer	**Unit**	**A1**	**A2**	**B1**	**B2**	**C1**	**Total**
Learner	**Texts**	16	83	75	74	88	**336**
Expert	**Texts**	49	157	258	288	115	**867**
Expert	**Sentences**	98	82	58	92	45	**375**

Table 12.2: CEFR-level annotated Swedish datasets.

All three corpora are equipped also with automatic linguistic annotation which includes lemmatization, part-of-speech (POS) tagging and dependency parsing based on the Sparv pipeline[42] (Borin, Forsberg and Roxendal 2012).

12.4 A flexible feature set for linguistic complexity analysis

In this section, we provide a detailed description of the set of features used and relate them to cognitive aspects of linguistic complexity. The feature set is "flexible" in the sense that it can be applied to different types of L2 data and units of analysis (e.g. texts or sentences) since it does not incorporate text-level features (e.g. discourse-related aspects) or learner language specific ones (e.g. L2 error features). The feature set is comprised of 61 features in total and it has previously been proposed for CEFR classification experiments in Pilán, Volodina and Zesch (2016), see chapter 10. Table 12.3 shows the complete feature set divided into five sub-categories based on the type of NLP tools and resources used: *count-based*, *lexical*, *morphological*, *syntactic* and *semantic*.

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

COUNT	SYNTACTIC	MORPHOLOGICAL
Sentence length	Avg. DepArc length	Function W IncSc
Avg token length	DepArc Len > 5	Particle IncSc
Extra-long token	Max length DepArc	3SG pronoun IncSc
Nr characters	Right DepArc Ratio	Punctuation IncSc
LIX	Left DepArc Ratio	Subjunction IncSc
Bilog TTR	Modifier variation	PR to N
Square root TTR	Pre-modifier IncSc	PR to PP
LEXICAL	Post-modifier IncSc	Relative structure IncSc
Avg KELLY log freq	Subordinate IncSc	S-V IncSc
A1 lemma IncSc	Relative clause IncSc	S-V to V
A2 lemma IncSc	PP complement IncSc	ADJ IncSc
B1 lemma IncSc	MORPHOLOGICAL	ADJ variation
B2 lemma IncSc	Neuter N IncSc	ADV IncSc
C1 lemma IncSc	CJ + SJ IncSc	ADV variation
C2 lemma IncSc	Past PC to V	N IncSc
Difficult W IncSc	Present PC to V	N variation
Difficult N&V IncSc	Past V to V	V IncSc
OOV IncSc	Supine V to V	V variation
No lemma IncSc	Present V to V	Lex T to Nr T
SEMANTIC	Nominal ratio	Lex T to non-lex T
Avg senses per token	N to V	
N senses per N	Modal V to V	

Table 12.3: Feature set for linguistic complexity assessment in L2 data.

12.4.1 Count-based features

The feature set includes seven indicators that are based on simple counts or traditional readability measures. One such measure for Swedish is *LIX* (*Läsbarhetsindex* 'Readability index') proposed in Björnsson (1968). LIX combines the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters. Sentence length is measured both as the number of tokens and that of characters. Sentence length can indicate syntactic difficulty and it can be a sign of e.g. multiple clauses or larger noun phrases. Average token (*T*) length is computed based on the number of characters. Extra-long words, i.e. tokens longer than 13 characters, are also counted since compounding, frequent in Swedish, can result in particularly long words (Heimann Mühlenbock 2013). Compounds can be more challenging to process since the boundaries between the parts might not be obvious, and their meaning is not necessarily *compositional*, a sum of the meaning of the parts. Type-token ratio (TTR), the ratio of unique tokens to all tokens, is an indicator of lexical richness (Graesser et al. 2004). Using a more varied set of vocabulary items in a text increases its lexical complexity as the links between words referring to similar concepts need to be recognized. A bi-logarithmic and a square root TTR are used which decrease the effect of text and sentence length (Vajjala and Meurers 2012).

12.4.2 Word-list based lexical features

Besides richness, the frequency of words also influences lexical complexity as repeated exposure facilitates their processing (Graesser et al. 2004). Frequency information is collected from the KELLY list (Volodina and Kokkinakis 2012), based on web texts. Using log frequencies better reflects reading times since the effect of high frequency function words and other common, but less frequent words, is balanced out (Graesser et al. 2004).

Instead of n-grams, weakly lexicalized features are employed to increase the generalizability of the models on unseen data. Each token is represented by its corresponding CEFR level. Unlike Pilán, Volodina and Zesch (2016) in chapter 10 employing KELLY, the per-token CEFR level information is retrieved from two word lists compiled based on the L2 corpora described in Section 12.3. To guarantee the independence of the word lists from the datasets, we use SweLLex (Volodina et al. 2016b), a frequency list based on the learner essays when classifying CEFR levels in coursebook texts and SVALex (François et al. 2016), containing frequencies from coursebooks for making predictions on the essays. For sentences, SVALex has been used since it is independent from the dataset, but both reflect receptive linguistic complexity. Frequency distributions in these lists have been mapped to single CEFR levels based on the difference in per-level normalized frequency between adjacent levels as described in Alfter et al. (2016).

Instead of absolute counts, a normalized value, an *incidence score* (IncSc) is used to reduce the influence of sentence length as shown in (14), where N_t is the total number of tokens and N_c is the count of a certain category of tokens in the text or sentence (Graesser et al. 2004).

$$\textsc{IncSc} = \frac{1000}{N_t} \times N_c \tag{14}$$

The IncSc of *difficult* tokens is also computed, that is, tokens above a certain reference CEFR level, which can be the level of an L2 learner writing a text or whom the text would be presented to as reading material. This value is also computed separately for nouns and verbs, since these are crucial for conveying meaning. Moreover, the IncSc of tokens not present in the L2 word lists, i.e. out-of-vocabulary words (*OOV* INCSC) is also considered as well as the INCSC of non-lemmatized tokens (*No lemma* INCSC).

12.4.3 Morphological features

Morphological features include not only INCSC of different morpho-syntactic categories, but also variational scores, i.e. the ratio of a category to the ratio of *lexical* tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). Some specific features for L2 Swedish are the ratio of different verb forms to verbs which are typically introduced at varying stages of L2 learning. *S-verbs* (*S-VB*) are a group of Swedish verbs ending in *-s* that are peculiar in terms of morphology and semantics. They indicate either reciprocity, a passive construction or are *deponent* verbs, i.e. verbs active in meaning, but passive in form. Neuter gender nouns are also considered since they can indicate the abstractness of a concept (Graesser et al. 2004). Among relative structures relative adverbs, determiners, pronouns and possessives are counted. *Nominal ratio* (Hultman and Westman 1977) corresponds to the ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR), adverbs, and verbs. Its simplified version is the ratio of nouns to verbs, and it is meant to measure the information load of a text or reveal its genre (e.g. spoken vs. news text). A higher value corresponds to higher degrees of complexity and a more elaborate genre.

INCSC for punctuation marks as well as sub- and conjunctions (SJ, CJ) are also computed since their presence in larger quantities can indicate a more complex syntactic structure. Particles can change the meaning of verbs considerably, similarly to English phrasal verbs (Heimann Mühlenbock 2013). The INCSC of the third person singular (3SG) pronoun inspired by Zhang, Liu and Ni (2013) is also included since this is often used referentially, which can further increase the difficulty of processing.

12.4.4 Syntactic features

Syntactic aspects are related to readers' working memory load when processing sentences which can be increased by ambiguity or embedded constituents (Gibson 1998; Graesser et al. 2004). Here, the average length (depth) of dependency arcs (*DepArc*) and their direction is considered. Right dependencies indicating a syntactic head appearing after its dependent tend to be harder to process (Sjöholm 2012). Relative clauses, pre- and post-modifiers (e.g. adjectives and prepositional phrases), prepositional complements as well as subordinates, commonly used in previous research on linguistic complexity (Heimann Mühlenbock 2013; Schwarm and Ostendorf 2005), are also counted.

12.4.5 Semantic features

The two features in this category quantify available word senses per lemma based on the SALDO lexicon (Borin, Forsberg and Lönngren 2013). Both the average number of senses per token and the average number of noun senses per noun are considered. Polysemous words can be demanding for readers as they need to be disambiguated for a full understanding of the sentence (Graesser, McNamara and Kulikowich 2011).

12.5 Cross-dataset feature selection experiments

In this section, we describe the results of our feature selection experiments on the three datasets presented in Section 12.3. These experiments differ from the ones described previously by Pilán, Vajjala and Volodina (2016) and by Pilán, Volodina and Zesch (2016) in chapter 10 in a number of respects. In this work, the worth of individual features is evaluated rather than that of the complete set of features or groups of features. Moreover, as mentioned in section 12.4, most lexical features are based on L2 word lists rather than KELLY.

12.5.1 Experimental setup

We use 85% of each dataset as development set (DEV) for identifying the most informative features. The reported classification results using this part of the data are based on a stratified 5-fold cross-validation setup, that is, the original distribution of instances per CEFR level in the dataset has been preserved in all folds. We evaluated the generalizability of the selected subset of features on the remaining 15% of the data (TEST). As learning algorithm for these models, we used *LinearSVC* as implemented in scikit-learn (Pedregosa et al. 2011), which has been successfully applied in recent years in a number of NLP areas, such as native language identification (Chan et al. 2017).

12.5.2 Feature selection method

As a pre-processing step before training our classifiers, we used a *univariate feature selection* method, also available in scikit-learn, to identify the most informative features scored with *analysis of variance* (ANOVA).[43] This feature selection method is suitable for multi-class problems, it is independent of the learning method used and it has been previously adopted for NLP tasks, e.g. by Carbon, Fujii and Veerina (2014) and by Ljubešic and Kranjcic (2014). ANOVA is a statistical test that can be used to measure how strong the relationship between each feature and the output class is (CEFR levels in our case). It relies on *F-tests*, which can be employed to score features based on significant differences in their per-class mean values. To detect these differences indicating dependencies, first, the *variance*, i.e. the dispersion of the data in terms of its distance from the mean, is measured both *within* and *between* classes for each feature. Then, the F-statistic can be computed as the ratio of the variance between class means and the variance within a class.

12.5.3 Results

The results of the models with and without feature selection in terms of accuracy and F_1 are presented in table 12.4.

Data	**Features**	**SENT**		**TEXT-R**		**TEXT-E**	
		Acc	F_1	Acc	F_1	Acc	F_1
DEV	ALL	0.62	0.61	0.68	0.68	0.73	0.71
DEV	*k*-BEST	0.73	0.71	0.70	0.70	0.81	0.81
TEST	*k*-BEST	0.81	0.79	0.73	0.73	0.84	0.82
Number of *k*-BEST		21		54		24	

Table 12.4: Accuracy with feature selection across datasets.

Although using the complete set of features yielded a performance that considerably exceeded a baseline of always predicting the most frequent label in the dataset (MAJORITY), reducing it to the subset of most informative features improved the results further. The most substantial boost (+0.11 accuracy) was obtained for sentences. The models with selected features generalized well also on the held-out test sets. Moreover, while for SENT and TEXT-E only about one third of the features have been selected, almost all features were included in the *k* number of best ones for TEXT-R. The selected features ranked based on ANOVA are presented in table 12.5. For TEXT-R, features with low importance

[43] We experimented also with using χ^2 for scoring features, which, however, produced inferior results that we opted for not presenting here.

(a merit of < 3) are not listed separately. These are only indicated when they overlap with a feature selected by the other models (with a rank > 24).

Feature name	**Rank**		
	Sent	Text-R	Text-E
Nr characters	1	4	-
Square root TTR	2	7	9
A1 lemma IncSc	3	3	2
Punctuation IncSc	4	11	12
Sentence length	5	5	-
Relative clause	6	> 24	8
Difficult N&V IncSc	7	1	1
Avg. DepArc length	8	10	14
Max length DepArc	9	6	13
Bilog TTR	10	24	-
DepArc Len > 5	11	8	-
S-V IncSc	12	> 24	-
Present PC to V	13	18	17
Past PC to V	14	> 24	18
Particle IncSc	15	> 24	16
V variation	16	15	10
Difficult W IncSc	17	2	4
V IncSc	18	22	-
C1 lemma IncSc	19	> 24	5
3SG pronoun IncSc	20	> 24	-
N to V	21	> 24	20
OOV IncSc	-	9	-
LIX	-	12	-
Extra-long token	-	13	6
Lex T to Nr T	-	14	15
PR to PP	-	16	-
Past V to V	-	17	19
B1 lemma IncSc	-	19	3
Function W IncSc	-	20	-
Right DepArc Ratio	-	21	-
Avg token length	-	23	7
B2 lemma IncSc	-	> 24	11
N senses per N	-	> 24	21
PR to N	-	> 24	22
Nominal ratio	-	> 24	23

Feature name	Rank		
	SENT	TEXT-R	TEXT-E
N INCSC	-	> 24	24

Table 12.5: The *k*-best features and their rank across different datasets.

Fourteen features were among the most informative ones across all three datasets, which are highlighted in bold in table 12.5. One such feature was the count-based measure of square root TTR, thus it seems that a varied way of expression, through e.g. the use of synonyms, is a good indicator of linguistic complexity in the L2 context. Among the word-list based lexical features, besides the proportion of difficult lexica, the amount of tokens at the extremes of the CEFR scale, namely the lowest, A1 level and the advanced, C1 level (the highest available in our L2 lists) were also useful predictors. Interestingly, two out of the three strong indicators of L2 English essays quality identified in Crossley and McNamara (2011) were lexical diversity, closely related to our Square root TTR feature, and lexical frequency, based on the same type of information as our word-list features. Lexical variation in terms of TTR as well as verb variation were also found highly predictive for L2 Estonian learner texts Vajjala and Lõo (2014). These findings indicate the predictive strength of these features across languages.

Furthermore, syntactic features relative to the length of dependency arcs and verb-related morphological features (e.g. INCSC of participles and s-verbs) were among the *k*-best for all datasets. Such verb forms are, in fact, typically introduced explicitly to L2 learners at higher CEFR levels (Fasth and Kannermark 1997). The amount of punctuation and particles was also indicative of complexity. The former can, for example, indicate clause boundaries and hence more complex sentences. Particles, on the other hand, can be challenging for language learners, since they alter the meaning of verbs.

For the two datasets related to receptive skills, SENT and TEXT-R, a number of count features were strongly predictive. Unlike for TEXT-E, sentence length in terms of both the number of tokens and the number of characters were highly informative for determining receptive complexity. Although the proportion of lexical tokens to all tokens was not informative at the sentence level, it proved to be a good indicator of linguistic complexity at the text level. The traditional readability measure, LIX was informative only for TEXT-R, which could be explained by the fact that this dataset was the most similar to the intended use of LIX, namely determining readability at the text level. On the other hand, the other traditional formula, nominal ratio, was more useful across datasets, especially in its simplified version (*N to V*). It would be useful to investigate further whether this also depends on a difference in the genre of the text.

A limitation of our study is the relatively small size of our datasets, which is especially true in the case of the A1 level learner essays. Considering the difficulties in having access to similar types of L2 data, and the extension of our experiments to cross-dataset observations, the results could still provide valuable insights for teaching experts and members of the NLP community targeting similar tasks. Moreover, additional aspects relevant for other languages known by L2 learners, especially their mother tongue, can also influence relative linguistic complexity. If the language being learned is genealogically related or geographically close to a language already known by learners, part of the grammatical and lexical peculiarities are likely to be already familiar and, consequently, less complex for them (Moran and Blasi 2014). Modeling this aspect, however, would require additional data.

12.6 Conclusion and future work

In this work, we described the results of a feature selection method applied to different language learning related datasets. We found a small number of features that proved useful across all datasets regardless of the length of the linguistic input or the type of relevant language learning skill. We showed that besides lexical frequency and variation, the length of dependencies and the amount and type of verbs carry valuable information for predicting proficiency levels. To our knowledge, the usefulness of single features across receptive and productive L2 data of different sizes has not been previously explored. We aimed at finding the optimal number and types of features to use in order to boost performance and decrease computation time for these types of predictions. An improved CEFR level classification is especially important for its integration into NLP applications aiming at on-the-fly assessment of texts or exercise generation. In the future, extending this investigation of feature importances to datasets in other languages could contribute to a deeper understanding about which indicators are more universally useful. Furthermore, the selected subset of features could be evaluated also with the help of teaching experts to confirm their usefulness.

References

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*, 1–7. Linköping University Electronic Press.

Amaral, Luiz A and Detmar Meurers 2011. On using Intelligent Computer-Assisted Language Learning in real-life foreign language teaching and learning. *ReCALL* 23 (1): 4–24.

Antonsen, Lene 2012. Improving feedback on L2 misspellings-an FST approach. *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, 1–10. Linköping University Electronic Press.

Arregik, Itziar Aldabe 2011. Automatic exercise generation based on corpora and natural language processing techniques. Ph.D. diss., Universidad del País Vasco.

Artstein, Ron and Massimo Poesio 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 (4): 555–596.

Attali, Yigal and Jill Burstein 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4, no. 3.

Barrot, Jessie Saraza 2015. Comparing the linguistic complexity in receptive and productive modes. *GEMA Online® Journal of Language Studies* 15, no. 2.

Barsalou, Lawrence W. 1982. Context-independent and context-dependent information in concepts. *Memory & Cognition* 10 (1): 82–93.

Beinborn, Lisa, Torsten Zesch and Iryna Gurevych 2012. Towards fine-grained readability measures for self-directed language learning. *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, Volume 80, 11–19.

Beinborn, Lisa, Torsten Zesch and Iryna Gurevych 2014a. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics* 2: 517–529.

Beinborn, Lisa, Torsten Zesch and Iryna Gurevych 2014b. Readability for foreign language learning: The importance of cognates. *ITL-International Journal of Applied Linguistics* 165 (2): 136–162.

Björnsson, Carl Hugo 1968. *Läsbarhet*. Liber.

Bordag, Stefan 2008. A comparison of co-occurrence and similarity measures as simulations of context. *International Conference on Intelligent Text Processing and Computational Linguistics*, 52–63. Springer.

Borin, Lars 2002a. Where will the standards for Intelligent Computer-Assisted language learning come from? *Proceedings of the Workshop on International Standards of Terminology and Language Resources Management*. Uppsala universitet.

Borin, Lars 2002b. What have you done for me lately? The fickle alignment of NLP and CALL. *Reports from Uppsala Learning Lab*.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj and Dimitrios Kokkinakis 2010. The past meets the present in Swedish FrameNet++. 14^{th} *European Association for Lexicography International Congress*, 269–281.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Anne Schumacher and Roland Schäfer 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. *Swedish Language Technology Conference*.

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 and Johan Roxendal 2012. Korp – the corpus infrastructure of Språkbanken. *LREC*, 474–478.

Borin, Lars and Anju Saxena 2004. Grammar, incorporated. Peter Juel Henrichsen (ed.), *CALL for the Nordic languages*, 125–145. Copenhagen: Samfundslitteratur.

Boullosa, Beto, Richard Eckart de Castilho, Alexander Geyken, Lothar Lemnitzer and Iryna Gurevych 2017. A tool for extracting sense-disambiguated example sentences through user feedback. *Proceedings of the European Chapter of the Association for Computational Linguistics*, pp. 69–72.

Boyatzis, Richard Eleftherios 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.

Branco, António, João Rodrigues, Francisco Costa, João Silva and Rui Vaz 2014. Rolling out text categorization for language learning assessment supported by language technology. *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, 256–261. Springer.

Braun, Virginia and Victoria Clarke 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3 (2): 77–101.

vor der Brück, Tim, Sven Hartrumpf and Hermann Helbig 2008. A readability checker with supervised learning using deep indicators. *Informatica* 32, no. 4.

Brysbaert, Marc, Evelyne Lagrou and Michael Stevens 2017. Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition* 20 (3): 530–548.

Burrows, Steven, Iryna Gurevych and Benno Stein 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education* 25 (1): 60–117.

Burstein, Jill 2003. The e-rater scoring engine: Automated essay scoring with natural language processing. *Lawrence Erlbaum Associates, Inc.*

Burstein, Jill and Martin Chodorow 2010. Progress and new directions in technology for automated essay evaluation. *Oxford Handbook of Applied Linguistics*.

Burstein, Jill, Jane Shore, John Sabatini, Brad Moulder, Steven Holtzman and Ted Pedersen 2012. The Language Muse^{sm} system: linguistically focused instructional authoring. *ETS Research Report Series* 2012, no. 2.

Carbon, Kyle, Kacyn Fujii and Prasanth Veerina 2014. Applications of machine learning to predict Yelp ratings. Stanford University.

Carroll, J.B., P. Davies and B. Richman 1971. *The American Heritage word frequency book.* Houghton Mifflin Boston.

Castles, Stephen, Hein De Haas and Mark J Miller 2013. *The age of migration: International population movements in the modern world.* Palgrave Macmillan.

Chall, Jeanne Sternlicht 1958. *Readability: An appraisal of research and application*. Ohio State University.

Chan, Sophia, Maryam Honari Jahromi, Benjamin Benetti, Aazim Lakhani and Alona Fyshe 2017. Ensemble methods for native language identification. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 217–223.

Chen, Yen-Yu, Chien-Liang Liu, Chia-Hoang Lee, Tao-Hsing Chang et al. 2010. An unsupervised automated essay scoring system. *IEEE Intelligent systems* 25 (5): 61–67.

Chinkina, Maria and Detmar Meurers 2016. Linguistically aware information retrieval: Providing input enrichment for second language learners. *Pro-*

ceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 188–198.

Cobb, Tom 1997. Is there any measurable learning from hands-on concordancing? *System* 25 (3): 301–315.

Collins-Thompson, Kevyn 2014. Computational assessment of text readability: a survey of current and future research. *International Journal of Applied Linguistics* 165 (2): 97–135.

Collins-Thompson, Kevyn and James P Callan 2004. A language modeling approach to predicting reading difficulty. *Proceedings of the HLT/NAACL Annual Conference*, 193–200.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Press Syndicate of the University of Cambridge.

Cresswell, Andy 2007. Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. *Language and Computers* 61 (1): 267–287.

Crocker, Linda and James Algina 1986. *Introduction to classical and modern test theory.* ERIC.

Crossley, Scott A and Danielle S McNamara 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning* 21 (2-3): 170–191.

Curto, Pedro, Nuno J Mamede and Jorge Baptista 2015. Automatic text difficulty classifier – Assisting the selection of adequate reading materials for European Portuguese teaching. *Proceedings of the International Conference on Computer Supported Education*, 36–44.

Dale, Edgar and Jeanne S Chall 1949. The concept of readability. *Elementary English* 26 (1): 19–26.

Daumé III, Hal 2007. Frustratingly easy domain adaptation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 256–263. Prague, Czech Republic: Association for Computational Linguistics.

Daumé III, Hal and Daniel Marcu 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pp. 101–126.

DeKeyser, Robert 2007. *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*. Cambridge University Press.

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.

Dell'Orletta, Felice, Martijn Wieling, Andrea Cimino, Giulia Venturi and Simonetta Montemagni 2014. Assessing the readability of sentences: which corpora and features? *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pp. 163–173.

Didakowski, Jörg, Lothar Lemnitzer and Alexander Geyken 2012. Automatic example sentence extraction for a contemporary German dictionary. *Proceedings of the conference of the European Association for Lexicography*, 343–349.

Diependaele, Kevin, Kristin Lemhöfer and Marc Brysbaert 2013. The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology* 66 (5): 843–863.

Falkenjack, Johan, Katarina Heimann Mühlenbock and Arne Jönsson 2013. Features indicating readability in Swedish text. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 27–40.

Fasth, Cecilia and Anita Kannermark 1997. *Form i focus: övningsbok i svensk grammatik. del B.* Lund: Folkuniv. förlag.

Fellbaum, Christiane 1998. *WordNet.* Wiley Online Library.

Feng, Lijun, Martin Jansche, Matt Huenerfauth and Noémie Elhadad 2010. A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 276–284. Association for Computational Linguistics.

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.

François, Thomas, Nuria Gala, Patrick Watrin and Cédrick Fairon 2014. FLELex: a graded lexical resource for French foreign learners. *Proceedings of the International Conference on Language Resources and Evaluation*, 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. *Proceedings of the 10th international conference on language resources and evaluation.*

Frey, Bruce B, Stephanie Petersen, Lisa M Edwards, Jennifer Teramoto Pedrotti and Vicki Peyton 2005. Item-writing rules: collective wisdom. *Teaching and Teacher Education* 21 (4): 357–364.

Garcia, Ignacio 2013. Learning a language for free while translating the web. Does duolingo work? *International Journal of English Linguistics* 3 (1): 19.

Geyken, Alexander, Christian Pölitz and Thomas Bartz 2015. Using a Maximum Entropy Classifier to link "good" corpus examples to dictionary senses. *Proceedings of the Electronic Lexicography in the 21st Century Conference*, 304–314.

Gibson, Edward 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68 (1): 1–76.

Graesser, Arthur C, Danielle S McNamara and Jonna M Kulikowich 2011. Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher* 40 (5): 223–234.

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* 36 (2): 193–202.

Gray, John 2010. *The construction of English: Culture, consumerism and promotion in the ELT global coursebook*. Palgrave Macmillan.

Gundel, Jeanette K, Nancy Hedberg and Ron Zacharski 2005. Pronouns without explicit antecedents: how do we know when a pronoun is referential. *Anaphora processing: linguistic, cognitive and computational modelling*, pp. 351–364.

Halácsy, Péter, András Kornai and Csaba Oravecz 2007. HunPos: an open source trigram tagger. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics on interactive poster and demonstration sessions*, 209–212. Association for Computational Linguistics.

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.

Hämäläinen, Wilhelmiina and Mikko Vinni 2006. Mitsuru Ikeda, Kevin D. Ashley and Tak-Wai Chan (eds), *Comparison of Machine Learning Methods for Intelligent Tutoring Systems*. Springer Berlin Heidelberg.

Hancke, Julia 2013. Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. Master's thesis, University of Tübingen.

Hancke, Julia and Detmar Meurers 2013. Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research Conference*, 54–56.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman 2009. Unsupervised learning. *The elements of statistical learning*, 485–585. Springer.

Heilman, Michael, Le Zhao, Juan Pino and Maxine Eskenazi 2008. Retrieval of reading materials for vocabulary and reading practice. *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, 80–88. Association for Computational Linguistics.

Heilman, Michal J., Kevyn Collins-Thompson, Jamie Callan and Maxine Eskenazi 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 460–467.

Heimann Mühlenbock, Katarina 2013. I see what you mean—assessing readability for specific target groups. *Data linguistica*, no. 24.

Heppin, Karin Friberg and Maria Toporowska Gronostaj 2012. The rocky road towards a Swedish FrameNet – creating SweFN. *Proceedings of the International Conference on Language Resources and Evaluation*, 256–261.

Holmes, Philip and Ian Hinchliffe 2003. *Swedish: A comprehensive grammar*. Psychology Press.

Horbach, Andrea, Alexis Palmer and Manfred Pinkal 2013. Using the text to evaluate short answers for reading comprehension exercises. *Second Joint Conference on Lexical and Computational Semantics, North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 286–295.

Huang, Yi-Ting, Hsiao-Pei Chang, Yeali Sun and Meng Chang Chen 2011. A robust estimation scheme of reading difficulty for second language learners. *11th IEEE International Conference on Advanced Learning Technologies (ICALT)*, 58–62. IEEE.

Hultman, Tor G and Margareta Westman 1977. *Gymnasistsvenska*. Liber.

Husák, Miloš 2010. *Automatic retrieval of good dictionary examples*. Bachelor Thesis.

Jiang, Jing and ChengXiang Zhai 2007. Instance weighting for domain adaptation in nlp. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 264–271.

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.

Kilgarriff, Adam 2009. Corpora in the classroom without scaring the students. *Proceedings from the 18th international symposium on english teaching*.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel 2014. The Sketch Engine: ten years on. *Lexicography*, pp. 7–36.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlỳ 2008. GDEX: automatically finding good dictionary examples in a corpus. *Proceedings of the European Association for Lexicography International Congress*.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell 2004. The Sketch Engine. *Proceedings of the European Association for Lexicography International Congress*, 105–116.

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.

Kokkinakis, Dimitrios, Maria Toporowska-Gronostaj and Karin Warmenius 2000. Annotating, disambiguating & automatically extending the coverage of the Swedish SIMPLE lexicon. *Proceedings of the International Conference on Language Resources and Evaluation*.

Kosem, Iztok, Miloš Husák and Diana McCarthy 2011. GDEX for Slovene. *Proceedings of the Electronic Lexicography in the 21st Century Conference*, 151–159.

Krashen, Stephen D 1987. *Principles and practice in second language acquisition*. New York.

Landauer, Thomas K, Darrell Laham and Peter W Foltz 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated essay scoring: A cross-disciplinary perspective*, pp. 87–112.

Larsson, Patrik 2006. Classification into readability levels: implementation and evaluation. Master's thesis, Uppsala University.

Lee, John and Mengqi Luo 2016. Personalized exercises for preposition learning. *Proceedings of ACL-2016 System Demonstrations*, 115–120.

Lemnitzer, Lothar, Christian Pölitz, Jörg Didakowski and Alexander Geyken 2015. Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. *Proceedings of Electronic lexicography in the 21st century 2015*, 21–31.

Li, Yifan, Petr Musilek, Marek Reformat and Loren Wyard-Scott 2009. Identification of pleonastic *it* using the web. *Journal of Artificial Intelligence Research*, pp. 339–389.

Little, David 2011. The Common European Framework of Reference for Languages: A research agenda. *Language Teaching* 44 (3): 381–393.

Ljubešic, Nikola and Denis Kranjcic 2014. Discriminating between very similar languages among Twitter users. *Proceedings of the Ninth Language Technologies Conference*, 90–94.

Ljubešić, Nikola and Mario Peronja 2015. Predicting corpus example quality via supervised machine learning. *Proceedings of the Electronic Lexicography in the 21st Century Conference*, 477–485.

Lu, Xiaofei 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45 (1): 36–62.

Megyesi, Beáta, Jesper Näsman and Anne Palmér 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Mendes, Amália, Sandra Antunes, Maarten Janssen and Anabela Gonçalves 2016. The COPLE2 corpus: a learner corpus for Portuguese. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Menn, Lise and Cecily Jill Duffield 2014. Looking for a 'Gold Standard' to measure language complexity: what psycholinguistics and neurolinguistics can (and cannot) offer to formal linguistics. *Measuring grammatical complexity*, pp. 281–302.

Meurers, Detmar 2012, *Natural Language Processing and Language Learning*. Blackwell Publishing Ltd.

Meurers, Detmar, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf and Niels Ott 2010. Enhancing authentic web pages for language learners. *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, 10–18. Association for Computational Linguistics.

Miltsakaki, Eleni and Karen Kukich 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10 (01): 25–55.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi and Bonnie Webber 2004. Annotating discourse connectives and their arguments. *Proceedings of the Workshop on Frontiers in Corpus Annotation at the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 9–16.

Miltsakaki, Eleni and Audrey Troutt 2008. Real-time web text classification and analysis of reading difficulty. *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, 89–97. Association for Computational Linguistics.

Mitkov, Ruslan 1998. Robust pronoun resolution with limited knowledge. *Proceedings of the 17th International Conference on Computational Linguistics*, 869–875.

Mitkov, Ruslan 2014. *Anaphora resolution.* Routledge.

Mitkov, Ruslan, Ha Le An and Nikiforos Karamanis 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering* 12 (2): 177–194.

Moran, Steven and Damián Blasi 2014. Cross-linguistic comparison of complexity measures in phonological systems. pp. 217–240.

Naber, Daniel 2003. A rule-based style and grammar checker. Master's thesis, Bielefeld University, Bielefeld, Germany.

Ng, Hwee Tou, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto and Christopher Bryant (eds) 2014. *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared task.* Baltimore, Maryland: Association for Computational Linguistics.

Ng, Vincent 2010. Supervised noun phrase coreference research: The first fifteen years. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1396–1411. Association for Computational Linguistics.

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.

Nilsson, Kristina 2010. Hybrid methods for coreference resolution in Swedish. Ph.D. diss., Department of Linguistics, Stockholm University.

Nilsson, Kristina and Lars Borin 2002. Living off the land: The Web as a source of practice texts for learners of less prevalent languages. *Proceedings of the International Conference on Language Resources and Evaluation.*

Nilsson Björkenstam, Kristina 2013. SUC-CORE: A balanced corpus annotated with noun phrase coreference. *Northern European Journal of Language Technology NEJLT* 3: 19–39.

Nivre, Joakim, Johan Hall and Jens Nilsson 2006. MaltParser: A data-driven parser-generator for dependency parsing. *Proceedings of LREC*, Volume 6, 2216–2219.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13 (02): 95–135.

North, Brian 2007. The CEFR illustrative descriptor scales. *The Modern Language Journal* 91 (4): 656–659.

Nyborg, Roger and Nils-Owe Pettersson 1991. *Svenska utifrån*. Svenska institutet.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter 2007. *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Östling, Robert, André Smolentzov, Björn Tyrefors and Erik Höglin 2013. Automated essay scoring for Swedish. *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Padó, Ulrike 2016. Get semantic with me! The usefulness of different feature types for short-answer grading. *Proceedings of the International Conference on Computational Linguistics*, 2186–2195.

Page, Ellis Batten 2003. Project essay grade: PEG. *Automated essay scoring: A cross-disciplinary perspective*, pp. 43–54.

Pan, Sinno Jialin and Qiang Yang 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–1359.

Paroubek, Patrick, Stéphane Chaudiron and Lynette Hirschman 2007. Principles of evaluation in Natural Language Processing. *Traitement Automatique des Langues* 48 (1): 7–31.

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.

Phandi, Peter, Kian Ming A. Chai and Hwee Tou Ng 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 431–439. Association for Computational Linguistics.

Pilán, Ildikó 2016. Detecting context dependence in exercise item candidates selected from corpora. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 151–161.

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)*, 120–126.

Pilán, Ildikó and Elena Volodina (Submitted). Investigating the importance of linguistic complexity features across different datasets related to language learning.

Pilán, Ildikó, Elena Volodina and Lars Borin 2017. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues (TAL) Journal, Special issue on NLP for Learning and Teaching* 57 (3): 67–91.

Pilán, Ildikó, Elena Volodina and Richard Johansson 2014. Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, 174–184.

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 the 26th International Conference on Computational Linguistics*, 2101–2111.

Pilán, Ildikó, Sowmya Vajjala and Elena Volodina 2016. A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications (IJCLA)* 7 (1): 143–159.

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, Proceedings of EUROCALL*, 218–225.

Pino, Juan and Maxine Eskenazi 2009. Semi-automatic generation of cloze question distractors effect of students' L1. *Proceedings of the Workshop on Speech and Language Technology in Education*, 65–68.

Pitler, Emily and Ani Nenkova 2009. Using syntax to disambiguate explicit discourse connectives in text. *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 13–16. Association for Computational Linguistics.

Poesio, Massimo, Simone Ponzetto and Yannick Versley 2011. Computational models of anaphora resolution: A survey. <`http://wwwusers.di.uniroma1.it/~ponzetto/pubs/poesio10a.pdf`>.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi and Bonnie L Webber 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the International Conference on Language Resources and Evaluation.*

Recasens, Marta, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio and Yannick Versley 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. *Pro-*

ceedings of the 5th International Workshop on Semantic Evaluation, 1–8. Association for Computational Linguistics.

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.

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*, 36–46. Linköping University Electronic Press.

Salamoura, Angeliki and Nick Saville 2010. Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, no. 1:101–132.

Salesky, Elizabeth and Wade Shen 2014. Exploiting morphological, grammatical, and semantic correlates for improved text difficulty assessment. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, 155–162.

Scherrer, Paula Levy and K. Lindemalm 2007. *Rivstart: A1+ A2,Textbok*. Stockholm: Natur & Kultur.

Schwarm, Sarah E and Mari Ostendorf 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 523–530.

Segler, Thomas M 2007. Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German. Ph.D. diss., University of Edinburgh.

Settles, Burr and Brendan Meeder 2016. A trainable spaced repetition model for language learning. *Proceedings of the Annual Meeting on Association for Computational Linguistics*.

Singh, Abhinav Deep, Poojan Mehta, Samar Husain and Rajkumar Rajakrishnan 2016. Quantifying sentence complexity based on eye-tracking measures. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, 202–212.

Sjöholm, Johan 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköping University.

Smith, Simon, PVS Avinesh and Adam Kilgarriff 2010. Gap-fill tests for language learners: corpus-driven item generation. *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, 1–6.

Søgaard, Anders 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies* 6 (2): 1–103.

Srinivasan, Latha and Jem Treadwell 2005. An overview of service-oriented architecture, web services and grid computing. *HP Software Global Business Unit*, vol. 2.

Ströbel, Marcus, Elma Kerz, Daniel Wiechmann and Stella Neumann 2016. CoCoGen-complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pp. 23–31.

Sumita, Eiichiro, Fumiaki Sugaya and Seiichi Yamamoto 2005. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. *Proceedings of the 2nd Workshop on Innovative Use of NLP for Building Educational Applications*, 61–68.

Sung, Yao-Ting, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang and Yu-Chia Chen 2015. Leveling L2 texts through readability: combining multilevel linguistic features with the CEFR. *The Modern Language Journal* 99 (2): 371–391.

Statistics Sweden 2016. Finland och Irak de två vanligaste födelseländerna bland utrikes födda. <`http://www.scb.se/sv_/Hitta-statistik/Artiklar/Finland-och-Irak-de-tva-vanligaste-fodelselanderna-bland-utrikes-fodda`>.

Tack, Anaïs, Thomas François, Sophie Roekhaut and Cédrick Fairon 2017. Human and automated CEFR-based grading of short answers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 169–179.

Teleman, Ulf, Staffan Hellberg and Erik Andersson 1999. *Svenska Akademiens grammatik*. Svenska Akademien/Norstedts ordbok (distr.).

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*, 1821–1824.

Thompson, Bruce and Justin E Levitov 1985. Using microcomputers to score and evaluate items. *Collegiate Microcomputer* 3 (2): 163–168.

Tiberius, Carole and Dirk Kinable 2015. Using and configuring GDEX for Dutch. Slides presented at the ENeL COST Action meeting, <http://www.elexicography.eu/wp-content/uploads/2015/04/ENeLWG3_GDEX4Dutch.pdf>.

Tolmachev, Arseny and Sadao Kurohashi 2017. Automatic extraction of high-quality example sentences for word learning using a determinantal point process. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 133–142.

Tomanek, Katrin, Udo Hahn, Steffen Lohmann and Jürgen Ziegler 2010. A cognitive cost model of annotations based on eye-tracking data. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1158–1167.

Vajjala, Sowmya and Kaidi Lõo 2014. Automatic CEFR level prediction for Estonian learner text. *NEALT Proceedings Series Vol. 22*, pp. 113–127.

Vajjala, Sowmya and Detmar Meurers 2012. On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, 163–173.

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*.

Vapnik, Vladimir 1998. *Statistical learning theory. 1998.* Wiley, New York.

Velleman, Eric and Thea van der Geest 2014. Online test tool to determine the CEFR reading comprehension level of text. *Procedia Computer Science* 27: 350–358.

Volodina, Elena 2008. *From corpus to language classroom: reusing Stockholm Umeå Corpus in a vocabulary exercise generator SCORVEX.*

Volodina, Elena, Richard Johansson and Sofie Johansson Kokkinakis 2012. Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation. *Proceedings of the Workshop on NLP for Computer Assisted Language Learning*, Volume 80, 59–70.

Volodina, Elena and Sofie Johansson Kokkinakis 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. *Proceedings of the International Conference on Language Resources and Evaluation*, 1040–1046.

Volodina, Elena, Beáta Megyesi, Mats Wirén, Lena Granstedt, Julia Prentice, Monica Reichenberg and Gunlög Sundberg 2016a. A friend in need?: Research agenda for electronic second language infrastructure. *The Sixth Swedish Language Technology Conference (SLTC).* Umeå University.

Volodina, Elena, Ildikó Pilán, Lars Borin and Therese Lindström Tiedemann 2014a. A flexible language learning platform based on language resources and web services. *Proceedings of the International Conference on Language Resources and Evaluation*, 3973–3978.

Volodina, Elena, Ildikó Pilán, Stian Rødven Eide and Hannes Heidarsson 2014b. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language. *Proceedings of the 3rd Workshop on NLP for Computer Assisted Language Learning*, 128–144.

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*, 76–84. Linköping University Electronic Press.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg and Monica Sandell 2016c. SweLL on the rise: Swedish learner language corpus for European Reference Level studies. *Proceedings of the Tenth International Conference on Language Resources and Evaluation.*

Webber, Bonnie, Matthew Stone, Aravind Joshi and Alistair Knott 2003. Anaphora and discourse structure. *Computational Linguistics* 29 (4): 545–587.

Wisniewski, Katrin, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel and Jirka Hana 2013. MERLIN: An online trilingual learner corpus empirically grounding the European reference levels in authentic learner data. *ICT for Language Learning 2013, Conference Proceedings.*

Witten, Ian H, Eibe Frank, Mark A Hall and Christopher J Pal 2011. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wojatzki, Michael, Oren Melamud and Torsten Zesch 2016. Bundled gap filling: A new paradigm for unambiguous cloze exercises. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 172–181.

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.

Yannakoudakis, Helen, Ted Briscoe and Ben Medlock 2011. A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies – Volume 1, 180–189. Association for Computational Linguistics.

Zesch, Torsten, Michael Wojatzki and Dirk Scholten-Akoun 2015. Task-independent features for automated essay grading. *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*.

Zhang, Lixiao, Zaiying Liu and Jun Ni 2013. Feature-based assessment of text readability. *7th International Conference on Internet Computing for Engineering and Science (ICICSE)*, 51–54. IEEE.

A LIST OF ADDITIONAL PUBLICATIONS NOT INCLUDED IN THE THESIS

A.1 Publications as main author

Additional publications as main author include:

- Pilán, Ildikó, David Alfter and Elena Volodina 2017. Lärka: an online platform where language learning meets natural language processing. *Proceedings of the 7th Workshop on Speech and Language Technology in Education (SLaTE).*

- Pilán, Ildikó 2015. Helping Swedish words come to their senses: word-sense disambiguation based on sense associations from the SALDO lexicon. *Proceedings of the Nordic Conference of Computational Linguistics*, 275–279. Linköping University Electronic Press.

- Pilán, Ildikó, Elena Volodina 2014. Reusing Swedish FrameNet for training semantic roles. In *Proceedings of the International Conference on Language Resources and Evaluation*, 1359–1363.

- Pilán, Ildikó, Elena Volodina and Richard Johansson 2014. Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the 9th Workshop on Building Educational Applications Using NLP*, 174–184.

- 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*, 218–225.

A.2 Other publications

Other co-authored publications, where the main author was not the candidate, include:

- Alfter, David, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, 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*, 1–7. Linköping University Electronic Press.

- Alfter, David, Ildikó Pilán (To appear). SB@GU at the Complex Word Identification 2018 Shared Task. *Proceedings of the 13th Workshop on Building Educational Applications Using NLP*.

- François, Thomas, Elena Volodina, Ildikó Pilán, Anaïs Tack 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 3766-3773.

- Volodina, Elena, Lars Borin, Ildikó Pilán, Anaïs Tack, Thomas François 2017. SVALex. En andraspråksordlista med CEFR-nivåer. *Svenskans beskrivning 35.*, 369-382.

- Volodina, Elena, Ildikó Pilán, David Alfter 2016. Classification of Swedish learner essays by CEFR levels. *Proceedings of EuroCALL 2016*.

- Volodina, Elena, Ildikó Pilán, Lars Borin, Therese Lindström Tiedemann 2014. A flexible language learning platform based on language resources and web services. *Proceedings of the International Conference on Language Resources and Evaluation*, 3973–3978.

- Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, Monica Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 206–212.

- Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, Thomas François 2016. SweLLex: second language learners' productive vocabulary. *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 76–84. Linköping University Electronic Press.

- Volodina, Elena, Ildikó Pilán, Stian Rødven Eide and Hannes Heidarsson 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, 128–144.

B LINGUISTIC ANNOTATION

We present the tags sets used for different parts of speech and dependency relations by the annotation tools integrated in the Sparv pipeline.

Tag	POS category (Swedish)	POS category (English)
AB	Adverb	Adverb
DT	Determinerare, bestämningsord	Determiner
HA	Frågande/relativt adverb	Interrogative/Relative Adverb
HD	Frågande/relativ bestämning	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivuttryck	Interrogative/Relative Possessive
IE	Infinitivmärke	Infinitive Marker
IN	Interjektion	Interjection
JJ	Adjektiv	Adjective
KN	Konjunktion	Conjunction
NN	Substantiv	Noun
PC	Particip	Participle
PL	Partikel	Particle
PM	Egennamn	Proper Noun
PN	Pronomen	Pronoun
PP	Preposition	Preposition
PS	Possessivuttryck	Possessive
RG	Räkneord: grundtal	Cardinal Number
RO	Räkneord: ordningstal	Ordinal Number
SN	Subjunktion	Subjunction
UO	Utländskt ord	Foreign Word
VB	Verb	Verb

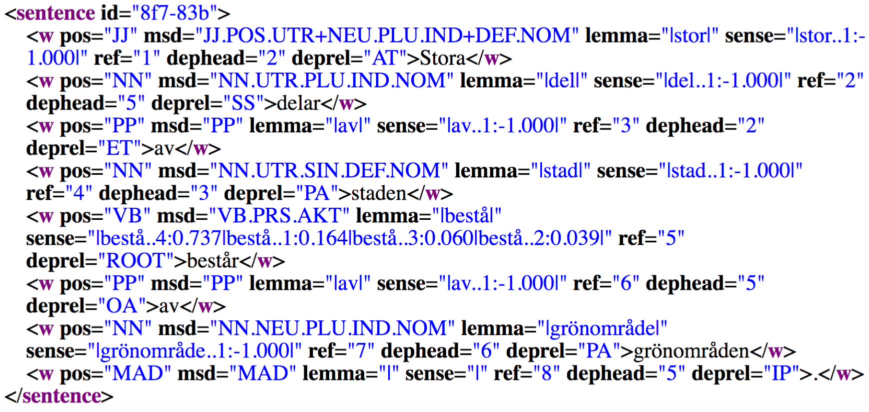
Table B.1: Part of speech tags.

Tag	**Dependency relation**
++	Coordinating conjunction
+A	Conjunctional adverbial
+F	Coordination at main clause level
AA	Other adverbial
AG	Agent
AN	Apposition
AT	Nominal (adjectival) pre-modifier
CA	Contrastive adverbial
DB	Doubled function
DT	Determiner
EF	Relative clause in cleft
EO	Logical object
ES	Logical subject
ET	Other nominal post-modifier
FO	Dummy object
FP	Free subjective predicative complement
FS	Dummy subject
FV	Finite predicate verb
I?	Question mark
IC	Quotation mark
IG	Other punctuation mark
IK	Comma
IM	Infinitive marker
IO	Indirect object
IP	Period
IQ	Colon
IR	Parenthesis
IS	Semicolon
IT	Dash
IU	Exclamation mark
IV	Nonfinite verb
JC	Second quotation mark
JG	Second (other) punctuation mark
JR	Second parenthesis
JT	Second dash
KA	Comparative adverbial
MA	Attitude adverbial
MS	Macrosyntagm
NA	Negation adverbial

Tag	Dependency relation
OA	Object adverbial
OO	Direct object
OP	Object predicative
PL	Verb particle
PR	Preposition
PT	Predicative attribute
RA	Place adverbial
SP	Subjective predicative complement
SS	Other subject
TA	Time adverbial
TT	Address phrase
UK	Subordinating conjunction
VA	Notifying adverbial
VO	Infinitive object complement
VS	Infinitive subject complement
XA	Expressions like "så att säga" (so to speak)
XF	Fundament phrase
XT	Expressions like "så kallad" (so called)
XX	Unclassifiable grammatical function
YY	Interjection phrase
	New Categories
CJ	Conjunct (in coordinate structure)
HD	Head
IF	Infinitive verb phrase minus infinitive marker
PA	Complement of preposition
UA	Subordinate clause minus subordinating conjunction
VG	Verb group

Table B.2: Dependency relation tags.

An example of the XML annotation provided by Sparv is shown in figure B.1. The sentence is the same as the one presented in Figure 5.1. The annotations include, for every word element (*w*): its part of speech (*pos* attribute); morpho-syntactic information (*msd*) such as gender, number, tense; lemmatization (*lemma*), the position of the token in the sentence (*ref*), its dependency head (*dephead*) and its dependency relation tag (*deprel*). Additional information is also available in the recently updated pipeline such as compound analysis (not included here) and word sense disambiguation (*sense* attribute).



```
<sentence id="8f7-83b">
  <w pos="JJ" msd="JJ.POS.UTR+NEU.PLU.IND+DEF.NOM" lemma="|stor|" sense="|stor..1:-
  1.000|" ref="1" dephead="2" deprel="AT">Stora</w>
  <w pos="NN" msd="NN.UTR.PLU.IND.NOM" lemma="|del|" sense="|del..1:-1.000|" ref="2"
  dephead="5" deprel="SS">delar</w>
  <w pos="PP" msd="PP" lemma="|av|" sense="|av..1:-1.000|" ref="3" dephead="2"
  deprel="ET">av</w>
  <w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|stad|" sense="|stad..1:-1.000|"
  ref="4" dephead="3" deprel="PA">staden</w>
  <w pos="VB" msd="VB.PRS.AKT" lemma="|bestå|"
  sense="|bestå..4:0.737|bestå..1:0.164|bestå..3:0.060|bestå..2:0.039|" ref="5"
  deprel="ROOT">består</w>
  <w pos="PP" msd="PP" lemma="|av|" sense="|av..1:-1.000|" ref="6" dephead="5"
  deprel="OA">av</w>
  <w pos="NN" msd="NN.NEU.PLU.IND.NOM" lemma="|grönområde|"
  sense="|grönområde..1:-1.000|" ref="7" dephead="6" deprel="PA">grönområden</w>
  <w pos="MAD" msd="MAD" lemma="|" sense="|" ref="8" dephead="5" deprel="IP">.</w>
</sentence>
```

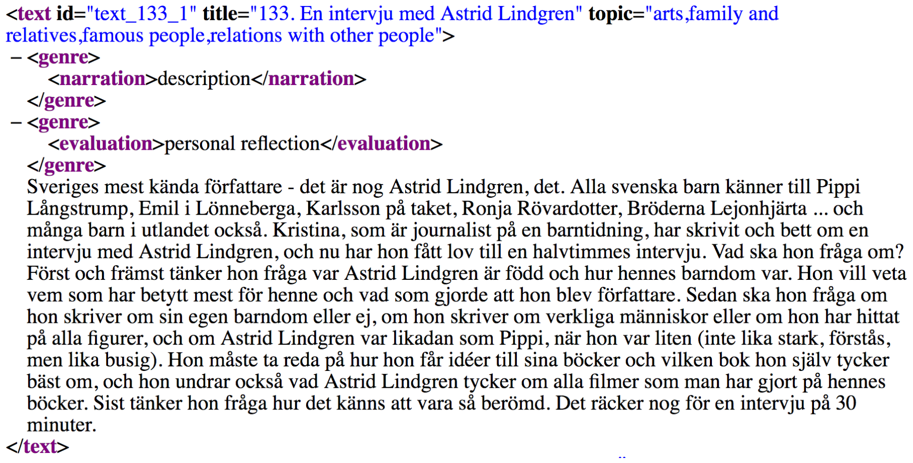
Figure B.1: An example of the linguistic annotation in Sparv.

C Dataset instance examples

In this chapter, we present examples for the different types of data used in our experiments: a text and some sentences from our coursebook corpus as well as a learner-written essay.

C.1 Texts and sentences from coursebooks

In this section, we illustrate the type of texts and sentences included in our datasets for receptive complexity analysis. All examples are from *Svenska utifrån* (Nyborg and Pettersson 1991). Figure C.1 shows a text intended for reading comprehension at B1 level. The XML format includes the metadata available in the corpus: the title ('An interview with Astrid Lingren'), the genre and the topic of the text.



```
<text id="text_133_1" title="133. En intervju med Astrid Lindgren" topic="arts,family and
relatives,famous people,relations with other people">
– <genre>
    <narration>description</narration>
  </genre>
– <genre>
    <evaluation>personal reflection</evaluation>
  </genre>
  Sveriges mest kända författare - det är nog Astrid Lindgren, det. Alla svenska barn känner till Pippi
  Långstrump, Emil i Lönneberga, Karlsson på taket, Ronja Rövardotter, Bröderna Lejonhjärta ... och
  många barn i utlandet också. Kristina, som är journalist på en barntidning, har skrivit och bett om en
  intervju med Astrid Lindgren, och nu har hon fått lov till en halvtimmes intervju. Vad ska hon fråga om?
  Först och främst tänker hon fråga var Astrid Lindgren är född och hur hennes barndom var. Hon vill veta
  vem som har betytt mest för henne och vad som gjorde att hon blev författare. Sedan ska hon fråga om
  hon skriver om sin egen barndom eller ej, om hon skriver om verkliga människor eller om hon har hittat
  på alla figurer, och om Astrid Lindgren var likadan som Pippi, när hon var liten (inte lika stark, förstås,
  men lika busig). Hon måste ta reda på hur hon får idéer till sina böcker och vilken bok hon själv tycker
  bäst om, och hon undrar också vad Astrid Lindgren tycker om alla filmer som man har gjort på hennes
  böcker. Sist tänker hon fråga hur det känns att vara så berömd. Det räcker nog för en intervju på 30
  minuter.
</text>
```

Figure C.1: Example of a reading comprehension (receptive) text from a coursebook.

For the sentence-level dataset, rather than using sentences occurring in coherent texts, sentences appearing in isolation were collected from lists and language examples. Language examples illustrate the usage of a lexical or a grammatical pattern. Figure C.2 shows sentences exemplifying the use of the construction *det finns* ‘there is’ followed by a noun in indefinite form from an A2 level lesson.

```
<language_example id="langex_29_4" title="§ DET FINNS + OBESTÄMT
SUBSTANTIV" skill="grammar" unit="full_sentences,single_words">
  Det finns en sportavdelning på varuhuset. Det finns ett varuhus i staden. Det finns
  många avdelningar på ett varuhus.
</language_example>
```

Figure C.2: Example of individual sentences in language examples.

In figure C.3, we present an example for lists of sentences at A2 level. The sentences illustrate the use of prepositions connected to the topic of furniture and interior decoration.

```
<list id="list_38_1" title="38. Prepositioner" skill="grammar,vocabulary"
unit="full_sentences">
  Blomman står i hörnet. Teven står mitt emot soffan. Tavlorna hänger på väggen.
  Hunden ligger under bordet. Mattan ligger på golvet.
</list>
```

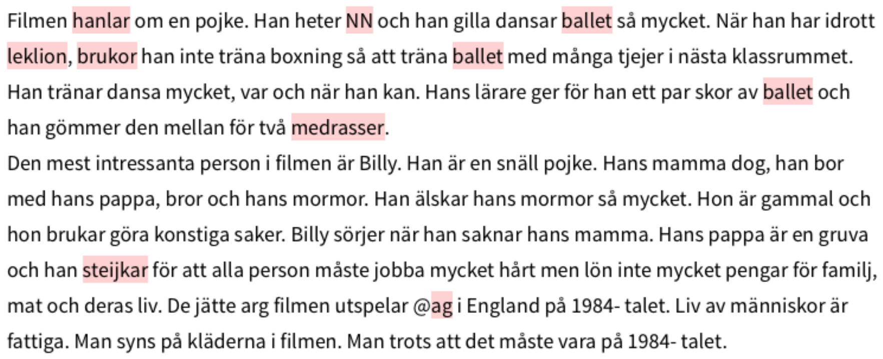
Figure C.3: Example of individual sentences occurring in lists.

C.2 Learner essays

Here we show part of a B1 level text written by a learner from the SweLL corpus. Besides text-related information, metadata about the learner is also included. (Part of this has been omitted here for privacy reasons.) This information includes, among others, age, gender, native language (*l1*) and residence time in number of month. The example text has been written in an exam setting without the use of additional resources. The token *NN* represents an anonymized token, while @ stands for an illegible character. This essay describing a film contains a number of learner errors, as it can be observed in its original version in figure C.4. Figure C.5, on the other hand, shows a number of highlighted tokens corresponding to the errors identified by LanguageTool, the software used to locate errors and retrieve correction candidates in the study presented in chapter 10. Finally, the error-normalized version of the essay is shown in figure C.6. Correction candidates were chosen based on word co-occurrence information as described in section 10.5.2.

```
<essay age="19" cefr="B1" education="upper-secondary-3-4years" essay_id="SpIn69_4"
gender="female" l1="Vietnamese" permit="public" residence="10" resource="none"
setting="exam" subcorpus="SpIn" topic="arts">
  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. Den mest intressanta person i filmen är
  Billy. Han är en snäll pojke. Hans mamma dog, han bor med hans pappa, bror och hans mormor.
  Han älskar hans mormor så mycket. Hon är gammal och hon brukar göra konstiga saker. Billy
  sörjer när han saknar hans mamma. Hans pappa är en gruva och han steijkar för att alla person
  måste jobba mycket hårt men lön inte mycket pengar för familj, mat och deras liv. De jätte arg
  filmen utspelar @ag i England på 1984- talet. Liv av människor är fattiga. Man syns på kläderna
  i filmen. Man trots att det måste vara på 1984- talet.
</essay>
```

Figure C.4: An example learner essay.



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.

Den mest intressanta person i filmen är Billy. Han är en snäll pojke. Hans mamma dog, han bor med hans pappa, bror och hans mormor. Han älskar hans mormor så mycket. Hon är gammal och hon brukar göra konstiga saker. Billy sörjer när han saknar hans mamma. Hans pappa är en gruva och han steijkar för att alla person måste jobba mycket hårt men lön inte mycket pengar för familj, mat och deras liv. De jätte arg filmen utspelar @ag i England på 1984- talet. Liv av människor är fattiga. Man syns på kläderna i filmen. Man trots att det måste vara på 1984- talet.

Figure C.5: LanguageTool output for the essay.

```
<essay age="19" cefr="B1" education="upper-secondary-3-4years" essay_id="SpIn69_4"
gender="female" l1="Vietnamese" permit="public" residence="10" resource="none"
setting="exam" subcorpus="SpIn" topic="arts">
  Filmen handlar om en pojke. Han heter NN och han gilla dansar balett så mycket. När han har
  idrott lektion, brukar han inte träna boxning så att träna balett med många tjejer i nästa
  klassrummet. Han tränar dansa mycket, var och när han kan. Hans lärare ger för han ett par skor
  av balett och han gömmer den mellan för två madrasser. Den mest intressanta person i filmen är
  Billy. Han är en snäll pojke. Hans mamma dog, han bor med hans pappa, bror och hans mormor.
  Han älskar hans mormor så mycket. Hon är gammal och hon brukar göra konstiga saker. Billy
  sörjer när han saknar hans mamma. Hans pappa är en gruva och han strejkar för att alla person
  måste jobba mycket hårt men lön inte mycket pengar för familj, mat och deras liv. De jätte arg
  filmen utspelar @ar i England på 1984- talet. Liv av människor är fattiga. Man syns på kläderna
  i filmen. Man trots att det måste vara på 1984- talet.
</essay>
```

Figure C.6: The error-normalized version of the essay.

D Example SALDO entries

Figure D.1 shows some example entries from SALDO including both core and peripheral senses.

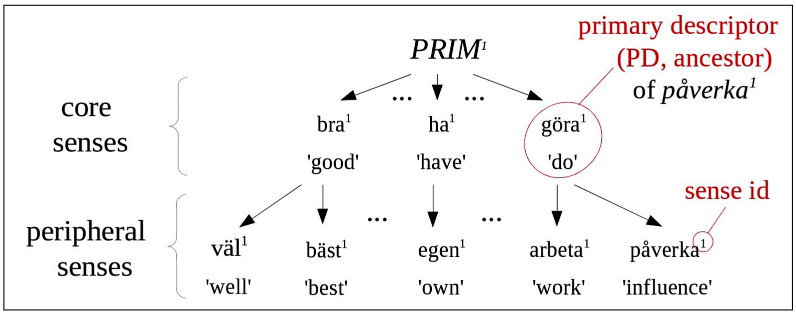
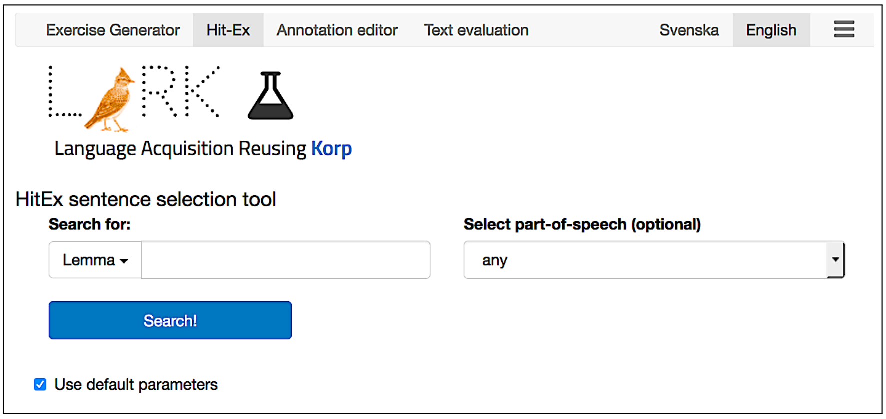


Figure D.1: Example SALDO senses.

E HitEx: a sentence selection tool

E.1 User interface

In this section, we present the HitEx user interface in detail. First, we show the initial page, set up for a simple search based on a default setting of the criteria and parameters (figure E.1). By clicking on the arrow on the right side of the button displaying the word *Lemma*, it is possible to switch from a lemma-based search to one based on wordforms, or CQP expressions for more complex target search patterns (e.g. any verb in simple past tense followed by a certain noun).



HitEx sentence selection tool

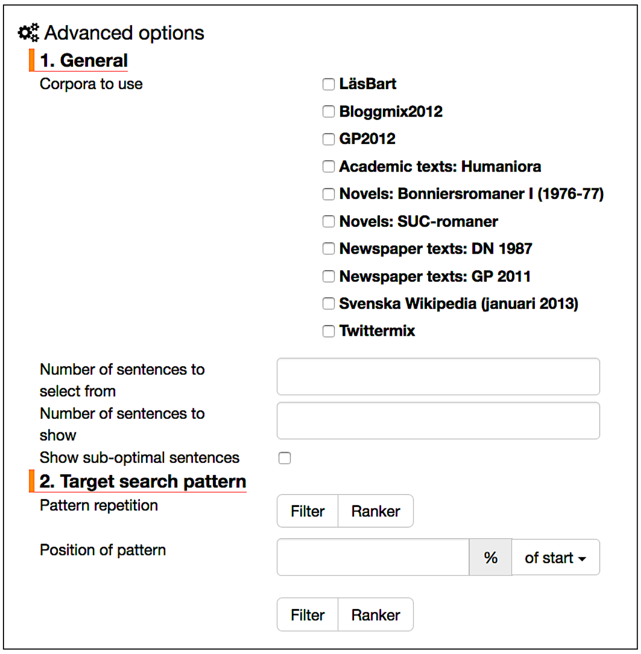
Search for: Lemma

Select part-of-speech (optional): any

☑ Use default parameters

Figure E.1: Opening page of HitEx: simple search with default options.

When the box next to *Use default parameters* is unticked, the advance search options become available. In figures E.2, E.3 and E.4, we show all advanced search options available in HitEx at the time of writing. The selection criteria are grouped into seven broader categories. Each filter or ranker option can be activated or left unselected if not wished to be used. All criteria, except for typicality and word frequency (in green font in figure E.4), are negatively correlated with the goodness of a sentence. The list of corpora in *General* is only a partial list of the available resources.



Advanced options

1. General

Corpora to use

□ LäsBart
□ Bloggmix2012
□ GP2012
□ Academic texts: Humaniora
□ Novels: Bonniersromaner I (1976-77)
□ Novels: SUC-romaner
□ Newspaper texts: DN 1987
□ Newspaper texts: GP 2011
□ Svenska Wikipedia (januari 2013)
□ Twittermix

Number of sentences to select from

Number of sentences to show

Show sub-optimal sentences □

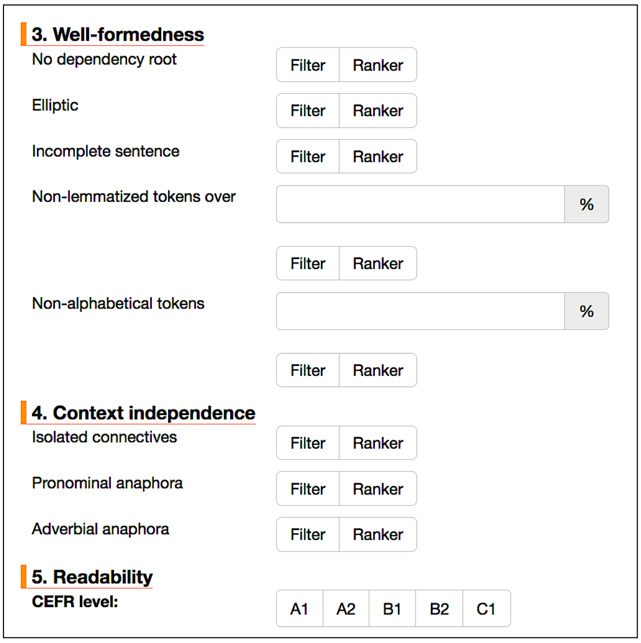
2. Target search pattern

Pattern repetition Filter | Ranker

Position of pattern % of start ▾

Filter | Ranker

Figure E.2: General and target pattern specific search options.



3. Well-formedness

No dependency root Filter | Ranker

Elliptic Filter | Ranker

Incomplete sentence Filter | Ranker

Non-lemmatized tokens over %

Filter | Ranker

Non-alphabetical tokens %

Filter | Ranker

4. Context independence

Isolated connectives Filter | Ranker

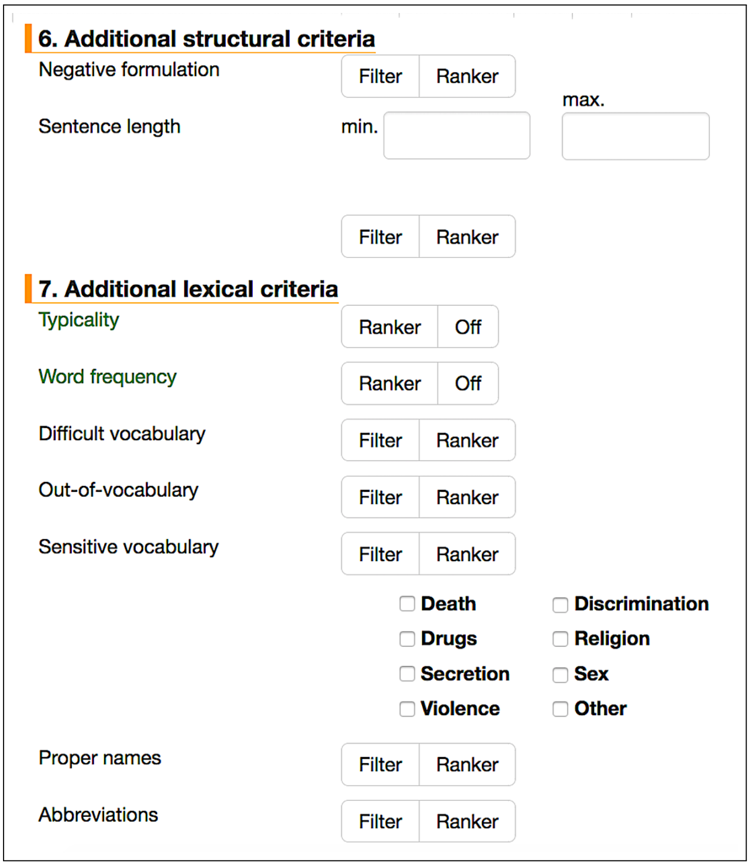
Pronominal anaphora Filter | Ranker

Adverbial anaphora Filter | Ranker

5. Readability

CEFR level: A1 | A2 | B1 | B2 | C1

Figure E.3: Options for ensuring well-formedness, context independence and readability.



6. Additional structural criteria

Negative formulation — Filter | Ranker

Sentence length — min. ___ max. ___

Filter | Ranker

7. Additional lexical criteria

Typicality — Ranker | Off

Word frequency — Ranker | Off

Difficult vocabulary — Filter | Ranker

Out-of-vocabulary — Filter | Ranker

Sensitive vocabulary — Filter | Ranker

□ **Death** □ **Discrimination**
□ **Drugs** □ **Religion**
□ **Secretion** □ **Sex**
□ **Violence** □ **Other**

Proper names — Filter | Ranker

Abbreviations — Filter | Ranker

Figure E.4: Additional structural and lexical criteria.

E.2 User evaluation

E.2.1 Settings used for the selection criteria and parameters

In tables E.1 and E.2, we present the parameters and criteria used during the user-based evaluation of HitEx. The list of corpora we searched sentence candidates in included both fiction and newspaper texts, namely: the novels published by Norstedts in 1999; the local newspaper, Göteborgs-Posten (2010–13) and its magazine-style attachment, Två dagar 'Two days'; the easy-to-read newspaper Åtta sidor 'Eigth pages'; LäSBarT, a corpus of easy-to-read and children's texts; the Stockholm-Umeå corpus and the Swedish treebank, Talbanken.

Parameter	**Value**
# Korp concordances to select from	300
# results to show	20
Search pattern near	sentence end
Appear within ... from sentence edge	50%
Target CEFR level	A2
% of words above the target CEFR	0
Minimum sentence length	6
Maximum sentence length	20
Threshold for non-alphabetical tokens	30
Threshold for non-lemmatized tokens	30
Sensitive vocabulary categories	all

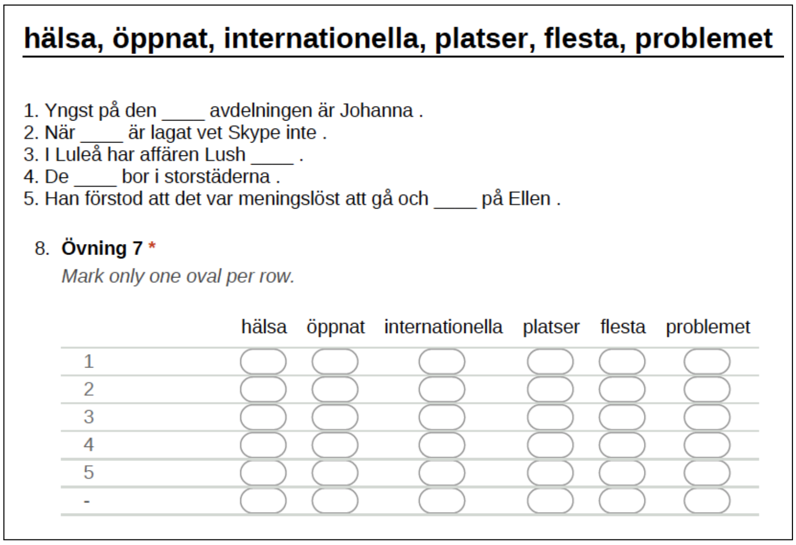
Table E.1: Parameter settings used during the user evaluation.

Criterion	**Value**
Well-formedness	
Dependency root	filter
Incompleteness	filter
Elliptic	filter
Non-alphabetical tokens	filter
Non-lemmatized tokens	filter
Context independence	
Isolated structural connective	filter
Pronominal anaphora	filter
Adverbial anaphora	filter
L2 complexity (CEFR)	filter
Additional structural criteria	
Negative formulations	inactive
Interrogative sentence	filter
Direct speech	filter
Answer to yes/no questions	filter
Modal verbs	inactive
Sentence length	filter
Additional lexical criteria	
Difficult vocabulary (KELLY)	ranker
Word frequency (SVALex)	ranker
Out-of-vocabulary (SVALex)	filter
Sensitive vocabulary	filter
Typicality	ranker
Proper names	ranker
Abbreviations	filter

Table E.2: Criteria used during the user evaluation..

E.2.2 Example learner exercises

In figures E.5 and E.6, we show two examples of the word bank exercise format presented to the learners, which had been semi-automatically constructed based on the sentences selected with HitEx. Figure E.5 illustrates an A2-level exercise with keywords of different parts of speech, while figure E.6 is a B1-level exercise with nouns only as keywords, all of which are of the same morpho-syntactic form.



hälsa, öppnat, internationella, platser, flesta, problemet

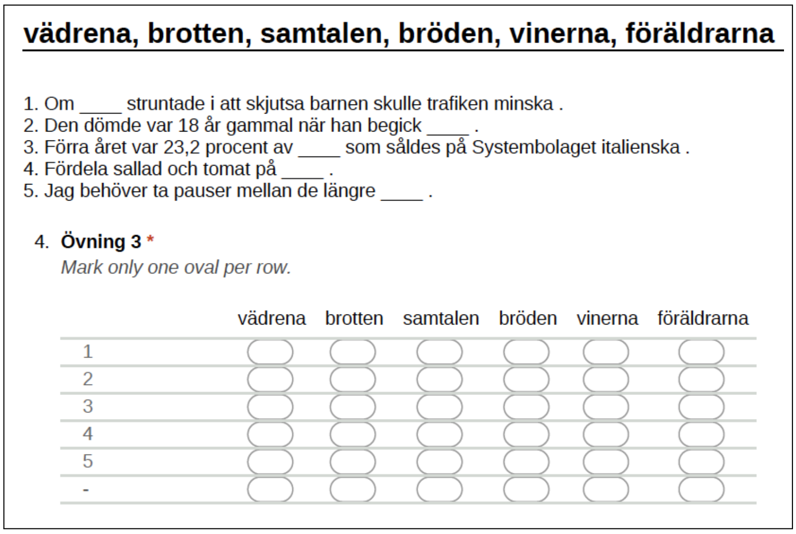
1. Yngst på den ____ avdelningen är Johanna .
2. När ____ är lagat vet Skype inte .
3. I Luleå har affären Lush ____ .
4. De ____ bor i storstäderna .
5. Han förstod att det var meningslöst att gå och ____ på Ellen .

8. **Övning 7** *

Mark only one oval per row.

	hälsa	öppnat	internationella	platser	flesta	problemet
1	○	○	○	○	○	○
2	○	○	○	○	○	○
3	○	○	○	○	○	○
4	○	○	○	○	○	○
5	○	○	○	○	○	○
-	○	○	○	○	○	○

Figure E.5: A2-level word bank exercise with mixed parts of speech.



vädrena, brotten, samtalen, bröden, vinerna, föräldrarna

1. Om ____ struntade i att skjutsa barnen skulle trafiken minska .
2. Den dömde var 18 år gammal när han begick ____ .
3. Förra året var 23,2 procent av ____ som såldes på Systembolaget italienska .
4. Fördela sallad och tomat på ____ .
5. Jag behöver ta pauser mellan de längre ____ .

4. **Övning 3** *

Mark only one oval per row.

	vädrena	brotten	samtalen	bröden	vinerna	föräldrarna
1	○	○	○	○	○	○
2	○	○	○	○	○	○
3	○	○	○	○	○	○
4	○	○	○	○	○	○
5	○	○	○	○	○	○
-	○	○	○	○	○	○

Figure E.6: B1-level word bank exercise with nouns of the same form.

After reading both the candidate words and the gapped sentences, learners' task consisted of matching them using the numbered rows of the grid. Five sentences were clustered together in one exercise in every case and a sixth word not fitting any of the sentences (a distractor) was also included to reduce the probability of correct answers by chance.

F Feature selection experiments

In this chapter, we illustrate the performance of our linguistic complexity features for both our sentence-level (SENT) and text-level datasets comprised of reading comprehension texts (TEXT-R) and learner essays (TEXT-E).

F.1 Informative features for receptive texts

In table F.1, we present the list of the remaining k-best features with a rank of >24 which were not included in table 12.4.

Rank	Feature	Rank	Feature
25	Post-modifier INCSC	40	A2 lemma INCSC
26	Relative clause INCSC	41	Relative structure INCSC
27	Particle INCSC	42	N variation
28	ADV INCSC	43	3SG pronoun INCSC
29	PR to N	44	Nominal ratio
30	Present PC to V	45	Avg KELLY log freq
31	Past PC to V	46	Modal V to V
32	S-V INCSC	47	C1 lemma INCSC
33	N INCSC	48	N senses per N
34	N to V	49	Left DepArc Ratio
35	ADJ INCSC	50	Neuter N INCSC
36	Avg senses per token	51	Supine V to V
37	PP complement INCSC	52	B2 lemma INCSC
38	Pre-modifier INCSC	53	ADJ variation
39	S-V to V	54	CJ + SJ INCSC

Table F.1: The k-best features for READINGTEXTS with a rank of >24.

F.2 ANOVA F-values of selected features

In tables F.2 – F.4, we present the merit of each feature based on ANOVA F-values. Each feature had $p < 0.01$ except for the ones marked with $*$ where $0.01 \leq p < 0.05$ applies.

Rank	**Feature**	**Merit**
1	Nr characters	148.0
2	Square root TTR	143.85
3	A1 lemma INCSC	128.43
4	Punctuation INCSC	121.46
5	Sentence length	119.03
6	Relative clause	108.66
7	Difficult N&V INCSC	85.53
8	Avg. DepArc length	73.4
9	Max length DepArc	73.3
10	Bilog TTR	44.12
11	DepArc Len > 5	42.41
12	S-V INCSC	41.06
13	Present PC to V	31.67
14	Past PC to V	31.67
15	Particle INCSC	29.26
16	V variation	27.96
17	Difficult W INCSC	23.35
18	V INCSC	18.82
19	C1 lemma INCSC	16.96
20	3SG pronoun INCSC	16.35
21	N to V	16.04

Table F.2: All *k*-best features for SENT.

Rank	Feature	Merit
1	Difficult N&V IncSc	251.19
2	Difficult W IncSc	247.86
3	A1 lemma IncSc	64.42
4	Nr characters	48.39
5	Sentence length	45.5
6	Max length DepArc	36.3
7	Square root TTR	35.14
8	DepArc Len $>$ 5	25.06
9	OOV IncSc	18.37
10	Avg. DepArc length	17.33
11	Punctuation IncSc	15.96
12	LIX	11.56
13	Extra-long token	9.48
14	Lex T to Nr T	9.19
15	V variation	8.16
16	PR to PP	7.06
17	Past V to V	6.19
18	Present PC to V	5.52
19	B1 lemma IncSc	4.37
20	Function W IncSc	3.8
21	Right DepArc Ratio	3.4
22	V IncSc	3.35*
23	Avg token length	3.27*
24	Bilog TTR	3.09*

Table F.3: The *k*-best features for TEXT-R with a merit of >3.

Rank	**Feature**	**Merit**
1	Difficult N&V IncSc	159.38
2	A1 lemma IncSc	158.07
3	B1 lemma IncSc	129.01
4	Difficult W IncSc	111.41
5	C1 lemma IncSc	72.88
6	Extra-long token	58.41
7	Avg token length	57.64
8	Relative clause IncSc	39.49
9	Square root TTR	39.06
10	V variation	36.49
11	B2 lemma IncSc	35.98
12	Punctuation IncSc	32.3
13	Max length DepArc	32.21
14	Avg. DepArc length	30.43
15	Lex T to Nr T	29.5
16	Particle IncSc	27.76
17	Present PC to V	25.08
18	Past PC to V	25.08
19	Past V to V	16.87
20	N to V	16.1
21	N senses per N	15.26
22	PR to N	15.18
23	Nominal ratio	13.3
24	N IncSc	13.18

Table F.4: All *k*-best features for Text-E.

F.3 Effects of incremental feature inclusion

Figures F.1 – F.3 show the development of classification accuracy while increasing the number (*k*) of features included in the model. The features were ordered with ANOVA analysis in decreasing order of their merit. For additional details including the list of ranked features and their overall performance, see section 5.4.

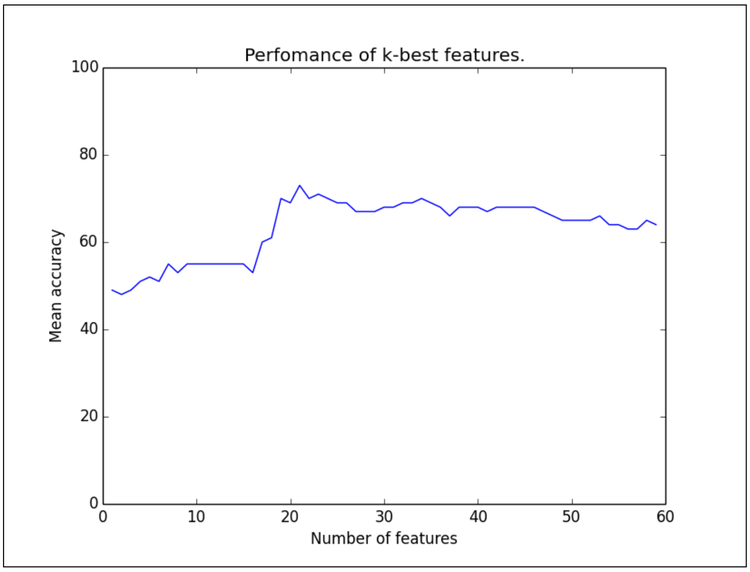


Figure F.1: Performance of *k*-best features for SENT.

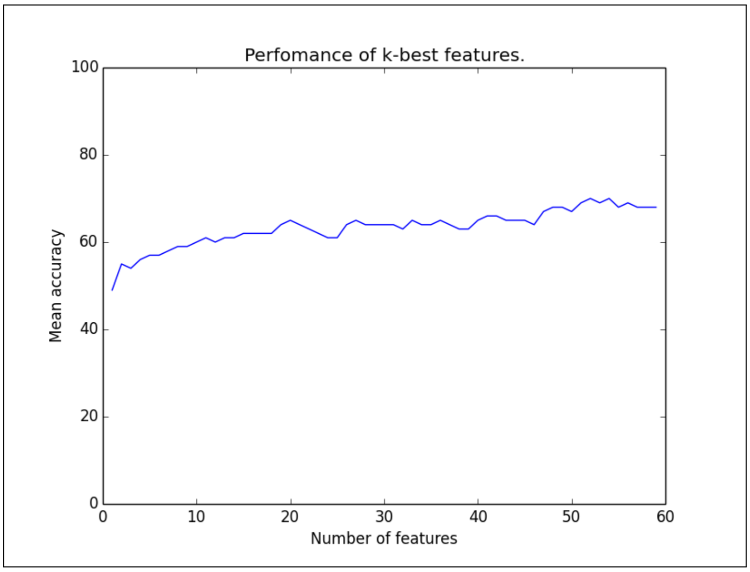


Figure F.2: Performance of *k*-best features for TEXT-R.

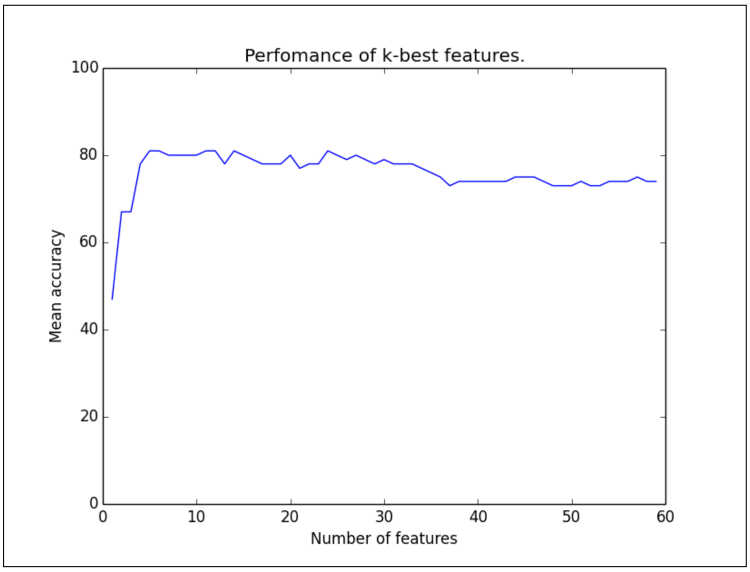


Figure F.3: Performance of *k*-best features for TEXT-E.

F.4 Confusion matrices with *k*-best features

Tables F.5 – F.7 present the confusion matrices on the test sets using the *k*-best features for all 3 datasets. Correct predictions are highlighted in bold font.

Predictions					
A1	A2	B1	B2	C1	
14	0	0	1	0	L
0	**12**	0	0	0	a
0	3	**2**	4	0	b
0	1	2	**11**	0	e
0	0	0	0	**7**	l

Table F.5: Confusion matrix for SENT with *k*-best features.

Predictions					
A1	A2	B1	B2	C1	
7	0	0	0	0	L
1	**15**	7	1	0	a
0	4	**22**	11	2	b
0	1	10	**33**	0	e
0	0	0	0	**17**	l

Table F.6: Confusion matrix for TEXT-R with *k*-best features.

Predictions					
A1	A2	B1	B2	C1	
3	0	0	0	0	L
0	**12**	1	0	0	a
0	3	**4**	4	0	b
0	0	0	**11**	0	e
0	0	0	0	**17**	l

Table F.7: Confusion matrix for TEXT-E with *k*-best features.