# Predicting and Manipulating the Difficulty of Text-Completion Exercises for Language Learning

Vom Fachbereich Informatik der Technischen Universität Darmstadt genehmigte

#### Dissertation

zur Erlangung des akademischen Grades Doktor der Naturwissenschaften

vorgelegt von

Lisa Marina Beinborn, M.Sc. geboren in Münster

Tag der Einreichung: 28. April 2016Tag der Disputation: 5. Juli 2016

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt

Prof. Dr. Torsten Zesch, Duisburg Prof. Dr. Detmar Meurers, Tübingen

Darmstadt 2016

D17

Please cite this document as URN: urn:nbn:de:tuda-tuprints-56479

URL: http://tuprints.ulb.tu-darmstadt.de/56479/

This document is provided by tuprints, E-Publishing-Service of the TU Darmstadt http://tuprints.ulb.tu-darmstadt.de mailto: tuprints@ulb.tu-darmstadt.de



This work is published under the following Creative Commons license: Attribution – Non Commercial – No Derivative Works 4.0 International https://creativecommons.org/licenses/by-nc-nd/4.0/

## **Abstract**

The increasing levels of international communication in all aspects of life lead to a growing demand of language skills. Traditional language courses compete nowadays with a wide range of online offerings that promise higher flexibility. However, most platforms provide rather static educational content and do not yet incorporate the recent progress in educational natural language processing. In the last years, many researchers developed new methods for automatic exercise generation, but the generated output is often either too easy or too difficult to be used with real learners. In this thesis, we address the task of predicting and manipulating the difficulty of text-completion exercises based on measurable linguistic properties to bridge the gap between technical ambition and educational needs. The main contribution consists of a theoretical model and a computational implementation for exercise difficulty prediction on the item level. This is the first automatic approach that reaches human performance levels and is applicable to various languages and exercise types.

The exercises in this thesis differ with respect to the exercise content and the exercise format. As theoretical basis for the thesis, we develop a new difficulty model that combines content and format factors and further distinguishes the dimensions of text difficulty, word difficulty, candidate ambiguity, and item dependency. It is targeted at text-completion exercises that are a common method for fast language proficiency tests. The empirical basis for the thesis consists of five difficulty datasets containing exercises annotated with learner performance data. The difficulty is expressed as the ratio of learners who fail to solve the exercise. In order to predict the difficulty for unseen exercises, we implement the four dimensions of the model as computational measures. For each dimension, the thesis contains the discussion and implementation of existing measures, the development of new approaches, and an experimental evaluation on sub-tasks. In particular, we developed new approaches for the tasks of cognate production, spelling difficulty prediction, and candidate ambiguity evaluation. For the main experiments, the individual measures are combined into an machine learning approach to predict the difficulty of C-tests, X-tests and cloze tests in English, German, and French. The performance of human experts on the same task is de-

termined by conducting an annotation study to provide a basis for comparison. The quality of the automatic prediction reaches the levels of human accuracy for the largest datasets.

If we can predict the difficulty of exercises, we are able to manipulate the difficulty. We develop a new approach for exercise generation and selection that is based on the prediction model. It reaches high acceptance ratings by human users and can be directly integrated into real-world scenarios. In addition, the measures for word difficulty and candidate ambiguity are used to improve the tasks of content and distractor manipulation. Previous work for exercise difficulty was commonly limited to manual correlation analyses using learner results. The computational approach of this thesis makes it possible to predict the difficulty of text-completion exercises in advance. This is an important contribution towards the goal of completely automated exercise generation for language learning.

# Zusammenfassung

Durch die zunehmende internationale Kommunikation in allen Lebensbereichen werden Sprachkenntnisse immer gefragter. Traditionelle Sprachkurse konkurrieren mittlerweile mit zahlreichen Online-Angeboten, die eine höhere Flexibilität versprechen. Die meisten E-learning-Kurse beruhen jedoch auf eher statischen Lerninhalten und berücksichtigen noch nicht die jüngsten Entwicklungen der Sprachtechnologie. In den letzen Jahren wurden viele neue Methoden zur automatischen Aufgabengenerierung entwickelt, aber die resultierenden Aufgaben sind häufig zu einfach oder zu schwierig um in echten Lernszenarien eingesetzt zu werden. In dieser Dissertation geht es darum, die Schwierigkeit von Sprachlernaufgaben anhand messbarer linguistischer Merkmale automatisch vorherzusagen und anzupassen. Der Hauptbeitrag besteht in der Entwicklung eines theoretischen Modells und seiner Implementierung für die automatische Schwierigkeitsvorhersage von Aufgaben zur Textvervollständigung. Die entwickelte Methode ist die erste, die das Qualitätsniveau menschlicher Experten für die Schwierigkeitsvorhersage erreicht und für verschiedene Aufgabentypen und Sprachen eingesetzt werden kann.

Die theoretische Grundlage für diese Arbeit wird durch die Entwicklung eines neuen Schwierigkeitsmodells gelegt, das sowohl den Aufgabeninhalt als auch das Aufgabenformat berücksichtigt und dabei vier Dimensionen unterscheidet: die Textschwierigkeit, die Wortschwierigkeit der Lösung, die Ambiguität der Lösungskandidaten und die gegenseitige Abhängigkeit von Teilaufgaben. Im Fokus stehen Aufgaben, die dem Prinzip der reduzierten Redundanz folgen und vom Lerner die Vervollständigung eines lückenhaften Textes erfordern. Die empirische Grundlage besteht aus fünf Datensätzen mit Aufgaben und den entsprechenden Lösungsergebnissen von Sprachlernern. Die Schwierigkeit einer Aufgabe ergibt sich aus dem Anteil der Lerner, die die Aufgabe nicht lösen können. Um die Schwierigkeit von neuen Aufgaben vorherzusagen, werden automatische Methoden entwickelt, die die vier Dimensionen des Modells abbilden. Für jede Dimension werden existierende Methoden diskutiert und implementiert und neue Ansätze entwickelt und evaluiert. Die neuen Methoden beinhalten insbesondere Lösungen für die Produktion von Kognaten, die

Vorhersage von Rechtschreibschwierigkeiten und die Abschätzung der Ambiguität von Lösungskandidaten.

Die vier Dimensionen werden dann in einem maschinellen Lernverfahren zusammen geführt, um die Schwierigkeit von sogenannten C-tests, X-tests und Cloze Tests für Englisch, Französisch und Deutsch vorherzusagen. Die menschliche Fähigkeit, die Schwierigkeit von Aufgaben einzuschätzen wird zudem in einer Annotationsstudie erhoben, um eine Vergleichsbasis heranziehen zu können. Es zeigt sich, dass die automatische Vorhersagequalität der einer erfahrenen Lehrkraft entspricht, wenn die Datensätze groß genug sind.

Die guten Ergebnisse der Schwierigkeitsvorhersage werden schließlich verwendet, um die Schwierigkeit von Aufgaben automatisch zu justieren. Der neu entwickelte Ansatz zur Generierung und Auswahl von Aufgaben erzielt hohe Akzeptanzwerte bei pädagogischen Fachkräften und kann direkt in ihre Arbeitsroutinen integriert werden. Darüber hinaus werden die Methoden für die Bestimmung der Wortschwierigkeit und der Ambiguität zwischen Lösungskandidaten eingesetzt, um die Schwierigkeit von Aufgabentexten und Lösungskandidaten zu manipulieren. Frühere Arbeiten zur Aufgabenschwierigkeit waren weitestgehend auf manuelle Korrelationsanalysen von Lernerdaten beschränkt. Der automatische Ansatz dieser Arbeit ermöglicht es, die Schwierigkeit von Aufgaben vorab zu bestimmen. Dies ist ein wichtiger Beitrag auf dem Weg zur komplett automatischen Aufgabengenerierung für Sprachlerner.

# Acknowledgements - Danksagung

Diese Dissertation wäre ohne die Hilfe vieler Förderer, Ratgeber und Freunde nicht möglich gewesen. Ich danke zunächst der Klaus-Tschira-Stiftung für die finanzielle Förderung. Auf der wissenschaftlichen Seite haben meine drei Betreuer eine große Rolle gespielt. Ich danke vor allem Iryna Gurevych, die mir die Chance gegeben hat, in die Forschung zurück zu kehren. Sie hat in den UKP-Gruppen an der *TU Darmstadt* und am *DIPF* hervorragende Forschungsbedingungen geschaffen und mir einen klaren Betreuungsrahmen gegeben, der diese Arbeit ermöglicht hat. Es hat mich tief beeindruckt, mit wie viel Energie sie sich stets neuen interdisziplinären Fragestellungen widmet. Ich danke Torsten Zesch, weil er ein großartiger Mentor war und sich immer Zeit für meine Fragen genommen hat. Ich habe unglaublich viel von ihm gelernt. Ich danke Detmar Meurers für seinen Enthusiasmus und sein leidenschaftliches Interesse an den Forschungsthemen dieser Arbeit. Die Diskussionen mit ihm waren sehr motivierend.

Die Erstellung und Verwendung der Datensätze wurde durch die Zusammenarbeit mit sprachdidaktischen Einrichtungen ermöglicht. Ich danke den Mitarbeitern des Sprachenzentrums der *TU Darmstadt* dafür, dass sie sich auf die Pilotstudie zur digitalen Durchführung der Einstufungstests eingelassen haben und für viele Rückfragen und für Annotationsstudien zur Verfügung standen. Mein Dank gilt insbesondere Britta Hufeisen und Martha Gibson. Die Hiwis Jens Haase und Fabian Rokohl haben mich bei der Erstellung des Interfaces für die Testphasen hervorragend unterstützt. Ich danke Thomas Eckes vom *TestDaF*-Institut und Dirk Scholten-Akoun und Anna Mashkovskaya von der *Universität Duisburg-Essen* für die Bereitstellung weiterer Datensätze.

Ich danke Dorothea Krampen, die ihr Wissen über Item-Abhängigkeiten bereitwillig mit mir geteilt hat und eine sehr hilfsbereite Ansprechpartnerin war. Mein Dank gilt außerdem allen Mitgliedern des DFG-Netzwerks *INDUS* für viele inspirierende Diskussionen und wertvolle Ratschläge.

Ich habe das große Glück, Teil einer starken und engagierten Forschungsgruppe im *UKP Lab* zu sein. Jeder in dieser Gruppe hat Einfluss auf meine Ansichten und ich danke allen dafür, den Büroalltag so lebenswert zu machen. Stellvertretend für die zahlreichen

Wissenschaftler mit denen ich Artikel diskutiert, Kaffeepausen verbracht und über sprachliche Spitzfindigkeiten gelacht habe, möchte ich diejenigen nennen, mit denen ich das Büro geteilt habe: Nico Erbs, Richard Eckart de Castilho, Pedro Santos, Hans-Peter Zorn, Mohamed Khemakhem, Beto Boullosa, Christian Meyer, Judith Eckle-Kohler, Sven Bräutigam und Margot Mieskes. Im Rahmen einer interdisziplinären Kooperation war ich außerdem mit dem *Deutschen Institut für internationale pädagogische Forschung* assoziiert. Ich danke den Kollegen am *DIPF* und insbesondere meinen Schreibtischnachbarn Lucie Flekova, Daniel Bengs, Jinseok Nam und Susanne Neumann. Ich danke außerdem der Gruppe *Human Language Technologies* der *Fondazione Bruno Kessler* in Trento, bei der ich ein sehr spannendes Forschungspraktikum verbringen durfte.

Konzentrierte Forschung funktioniert nur, wenn man auch Abstand gewinnen kann. Ich danke meiner Familie und meinen Freunden dafür, dass sie immer für mich da sind und mit mir die vielen anderen schönen Seiten des Lebens teilen. Und ich danke Stefan. Für alles.

# Contents

1	Intr	Introduction 1						
	1.1	Computer-Assisted Language Learning	2					
	1.2	Exercise Difficulty	3					
	1.3	Thesis Outline	5					
	1.4	Contributions	6					
	1.5	Publication Record	7					
2	Exe	Exercises for Language Learning						
	2.1	Educational Criteria for Language Learning Exercises	9					
	2.2	NLP for Language Learning Exercises	18					
	2.3	Text-Completion Exercises	31					
	2.4	Chapter Summary	41					
3	The Concept of Difficulty							
	3.1	Difficulty Model	43					
	3.2	Thesis Goals	45					
	3.3	Difficulty Datasets	46					
	3.4	Measuring Difficulty	50					
	3.5	Chapter Summary	59					
4	Tex	t Difficulty Prediction	51					
	4.1	Readability Measures	62					
	4.2	Readability and Reduced Redundancy Exercises	68					
	4.3	L2 Readability	72					
	4.4	Chapter Summary	78					
5	Word Difficulty Prediction 7							
	5.1		80					
	5.2	Cognates	85					

	5.3	Spelling Difficulty	4					
	5.4	Chapter Summary	6					
6	The	Influence of the Exercise Format	7					
	6.1	The Candidate Space	9					
	6.2	Candidate Ambiguity	4					
	6.3	Item Dependencies	9					
	6.4	Chapter Summary	3					
7	Exe	rcise Difficulty Prediction 13	5					
	7.1	Human Difficulty Prediction	6					
	7.2	Predicting C-Test Difficulty	8					
	7.3	Predicting X-test Difficulty	9					
	7.4	Predicting Cloze Test Difficulty	2					
	7.5	Learning Curve	6					
	7.6	The Influence of the Exercise Text	7					
	7.7	Chapter Summary	8					
8	Diff	iculty Manipulation 15	9					
	8.1	Manipulating Exercise Content	0					
	8.2	Manipulating Exercise Format	9					
	8.3	Chapter Summary	5					
9	Cor	onclusions 177						
	9.1	Summary	7					
	9.2	Outlook	9					
In	dex	18	3					
Li	st of	Figures 18	7					
Li	st of	Tables 18	9					
Bi	bliog	graphy 19	1					
Ą۱	ppen	dix 21	9					
-,	A	Readability Features						
	В	Word Difficulty Features						
	С	Candidate Ambiguity Features	:3					
	D	Item Dependency Features	4					

#### CHAPTER 1

## Introduction

"In the lives of individuals and societies, language is a factor of greater importance than any other. For the study of language to remain solely the business of a handful of specialists would be a quite unacceptable state of affairs."

- Ferdinand de Saussure

In a labor market that is increasingly globalized, knowledge of foreign languages is more relevant than ever before. Multilingual skills are also required for private communication as friendships stretch across geographical and linguistic borders. In addition, almost a million refugees have been crossing the Mediterranean Sea to seek asylum in Europe in 2015 and are looking for cheap and fast solutions to acquaint themselves with the new languages.<sup>1</sup>

The static time frame of conventional language courses is often not compatible with learners' unstable working conditions and lifestyles. As an alternative, many learners turn to online portals for language learning. These portals are becoming increasingly popular although the provided contents are often rather inflexible and limited.

According to Vygotsky's zone of proximal development (Vygotsky, 1978), the range of suitable exercises for a learner is very narrow. Exercises that do not challenge the learner easily lead to boredom and stagnation, whereas overly complex exercises might result in frustration. Creating exercises that fit this narrow target zone is a tedious and time-consuming task. For optimal support of language learning, it is thus important to adapt the exercise difficulty according to the specific needs of the learner. In order to properly manipulate difficulty, it is necessary to measure difficulty.

So far, difficulty has mainly been a subjective estimate based on the experience of teachers. However, the teachers already know the solutions and cannot always anticipate the

<sup>&</sup>lt;sup>1</sup>997,125 non-European first-time asylum applicants have been reported from October 2014 to September 2015 by Eurostat: http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum\_quarterly\_report, accessed: December 11, 2015

confusion an exercise might cause for learners. This results in a subjective difficulty estimation that often lacks consistency and can only be evaluated based on posterior performance measures. As a consequence, the exercises that are used in high-stakes language testing (the *TOEFL* test, for example) usually follow a thorough and expensive pre-test protocol for quality assurance. In less professional testing environments, the exercise selection depends on the subjective and uncontrolled judgment of individual persons.

Computational methods can provide support for exercise selection. In this thesis, we focus on text-completion exercises and the goal is two-fold:

- (1) to develop automatic methods for difficulty prediction based on objective measurable properties
- (2) to evaluate how difficulty prediction can contribute to automated manipulation of exercise difficulty.

## 1.1 Computer-Assisted Language Learning

In the last years, numerous language learning portals and apps have been developed for multiple languages and some of them reach a huge range of users worldwide. Popular commercial examples are *Rosetta Stone*<sup>®</sup>, *busuu* and *babbel.*<sup>2</sup> *Rosetta Stone*<sup>®</sup> promotes a "natural" learning approach inspired by L1 acquisition and is mainly based on picture association tasks. *Busuu* follows the idea of social networks and supports the interaction between language learners and native speakers. *Babbel* aims at reaching communicative goals and claims that their exercises provide interactive contextualized content. All three companies offer free demo versions of their systems, but the full product can only be accessed through paid accounts. The most established free platform is *Duolingo* which was founded after a research initiative at Carnegie Mellon University.<sup>3</sup> The business model behind *Duolingo* is based on the idea that language learners translate content and evaluate translations of their peer learners for free while learning the language.

The fast success of language learning technology in commercial applications is astonishing. However, all applications still mainly provide pre-fabricated exercises that are manually designed by humans. This static approach can quickly lead to boredom of advanced learners and is lagging behind the recent developments in natural language processing. Ideal language learning exercises should be adapted to the learners' abilities and preferences and make use of authentic material matching the learners' interests (Gilmore, 2011). In order to enable such a fine-grained matching, exercises should be automatically generated on the fly so that the difficulty can be adjusted. Research in natural language processing for learners is currently developing very fast (see section 2.2). Automatic methods for

<sup>&</sup>lt;sup>2</sup>http://www.rosettastone.de, http://www.busuu.com, http://www.babbel.com, all accessed: November 25, 2015

<sup>&</sup>lt;sup>3</sup>http://www.duolingo.com, accessed: November 25, 2015

content selection, exercise generation and feedback generation have improved significantly. However, most existing approaches for exercise generation exhibit two main shortcomings. They fail to exert appropriate quality control on the generated exercises and they cannot anticipate the difficulty level (Skory and Eskenazi, 2010). These two aspects are closely related: exercises that are rated to be of insufficient quality are almost always either too difficult or too easy. In order to overcome these weaknesses, this thesis presents a new model for predicting the difficulty of exercises. In a second step, the automatic difficulty estimate serves as a selection filter for generated exercises and guides difficulty manipulation.

As a general scenario for this thesis, we focus on fast language proficiency tests. At the beginning of a language learning phase, it is important to analyze the proficiency level of the learner. For this purpose, the exercises should be in a range of appropriate difficulty. Exercises that are too easy or too difficult are not discriminative for the proficiency level. The difficulty datasets that we examine in the thesis have all been created for this scenario. However, text-completion exercises are used for a wide range of scenarios and our findings for difficulty prediction can easily be extended to other use cases. The ultimate goal is the generation of adaptive exercises for self-directed learning. As a first step towards this goal, we evaluate how difficulty prediction estimates can be used for manipulating the difficulty of exercises.

## 1.2 Exercise Difficulty

Natural languages are complex and continually developing constructs that include many exceptions to the rules. Hence, the potential problems for foreign language learners are manifold and hard to anticipate. This thesis presents a new difficulty model that captures many aspects of language learning and is applicable to a range of text-based exercises. The thesis outline follows the structure of the model.

The developed difficulty model that constitutes the theoretical basis for this thesis is displayed in figure 1.1. In general, the difficulty of an exercise is determined by the *exercise content* and the *exercise format* (Sigott, 2006). The exercise content is often referred to as the prompt or the input in other related work. For the exercise types discussed in this thesis, the exercise content consists of a short text or sentence from a general domain. This text is transformed into an exercise according to the exercise format. The upper part of the model in figure 1.1 represents the format factors and the lower part the content factors. The model is further categorized into four dimensions: the *text difficulty*, the *word difficulty*, the *candidate ambiguity*, and the *item dependency*. The dimensions are exemplified here following an example for a cloze exercise displayed in figure 1.2. Other exercise types will be introduced in section 2.3.

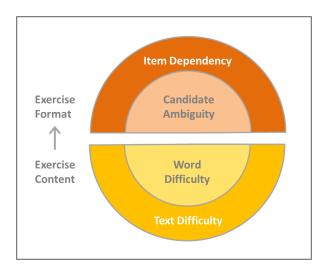
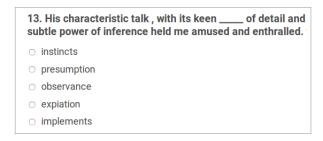


Figure 1.1: Thesis model for exercise difficulty

**Exercise content** The exercise content in the example is the sentence "His characteristic talk, with its keen <u>observance</u> of detail and subtle power of inference held me amused and enthralled". The difficulty of the content can be divided into the overall text difficulty and the particular word difficulty of the solution <u>observance</u>. The model needs to capture the readability of the text and the familiarity of the solution word for the learner.

**Exercise format** The difficulty of the format for this exercise depends on the candidate ambiguity of the answer options *instincts*, *presumption*, *observance*, *expiation*, and *implements*. The model needs to capture the difficulty of identifying the correct solution from the candidates. The item dependency is only relevant for exercises that contain more than one item such as C-tests and X-tests which will be introduced in section 2.3. Imagine the word *detail* in the example would also be replaced by a gap. Selecting the correct solution for the first gap would then be considerably more difficult. In this scenario, the model needs to capture whether the difficulty of the current item is influenced by the difficulty of the surrounding items.



**Figure 1.2:** Example for a cloze exercise, the solution is *observance*.

Content and format factors can both be further categorized into micro-level and macro-level processing. This term has been introduced to describe the solving strategies of language learners (Babaii and Ansary, 2001). For *micro-level processing*, the learners only take the gap itself and the directly neighboring words into account. For *macro-level processing*, they also consider whether the solution candidate fits into the larger sentential context. Psycholinguistic analyses discuss these strategies in detail and claim that both levels are required for successful solving (Sigott, 2006; Grotjahn and Stemmer, 2002). In the model, the outer circle represents macro-level processing and the inner circle represents micro-level processing. The distinction of the different difficulty dimensions is described in more detail in chapter 3. The subsequent chapters each focus on one particular dimension and their implementation as computational measures.

#### 1.3 Thesis Outline

The first two chapters of the thesis narrow down the thesis tasks. Chapter 2 delineates the field by identifying the educational requirements for language learning exercises and discussing the strengths and limits of educational natural language processing. Text-completion exercises are a popular choice for fast placement tests and combine positive educational properties with technical feasibility. Chapter 3 introduces the difficulty model that forms the theoretical basis for exercise difficulty prediction. In addition, it provides a specification of the thesis goals, a description of the difficulty datasets, and a comparison of two theories for measuring difficulty. The datasets have been obtained from authentic test settings with the goal of providing a rough estimate of the participants' language proficiency. Four datasets target language learners of English, German, and French, and one dataset has been collected in a language proficiency test with native and near-native speakers of German.

The following three chapters are based on the structure of the difficulty model. Chapter 4 describes the macro-level content dimension of *text difficulty*. The relationship between readability and difficulty is analyzed with a focus on the difference between native speakers and language learners. Chapter 5 comprises work for the micro-level content dimension of *word difficulty*. Words that are similar across languages (called *cognates*) facilitate word recognition, but complicate word production for learners. We address these two aspects of word difficulty in the novel tasks of *cognate production* and *spelling difficulty* prediction and develop new approaches to solve them. Chapter 6 deals with the two format dimensions of micro-level *candidate ambiguity* and macro-level *item dependency*. We analyze the candidate space and introduce new measures for candidate ambiguity that are rooted in methods for automatic solving of exercises. In addition, the existing approaches for the detection of item dependencies are evaluated on the datasets.

In chapter 7, the four difficulty dimensions are combined to predict the difficulty of exercises. The predictive power is evaluated and compared for three test variants and three

languages. The performance of human experts on the same task is determined by conducting an annotation study to provide a basis for comparison. The quality of the automatic prediction reaches the levels of human accuracy for the largest datasets. In chapter 8, we apply the findings for difficulty prediction to the task of *difficulty manipulation*. The proposed approaches are ready to be directly integrated into language teaching routines and can contribute to improved exercise generation. Chapter 9 summarizes the findings of the thesis and provides an outlook to future research directions.

#### 1.4 Contributions

The main contribution of this thesis can be summarized as follows:

A theoretical model and its computational implementation for the automatic prediction of exercise difficulty that is applicable to various languages and exercise types and reaches human performance levels.

The main contribution is evaluated in detail in chapter 7. As the task of difficulty prediction has rarely been tackled and has not been combined with recent advances in the field of natural language processing before, several sub-problems needed to be solved leading to the following contributions. They are distinguished into analyses, measures, resources, and applications (the corresponding chapter or section is indicated in brackets).

**Contributed analyses** The following analyses have been performed to build the theoretical basis for the main contribution.

- Analysis of the educational and technological requirements for high quality exercises for language learning (2).
- Development of a difficulty model that combines individual factors into an integrated concept of difficulty (3.1).
- Comparison of different approaches for measuring difficulty (3.4).
- Analysis of the psycholinguistic differences between L1 acquisition and L2 learning, and the computational consequences for L1 and L2 readability (4.3).
- Analysis of the candidate space for different exercises as an indicator of redundancy reduction (6.1).
- Analysis of item dependencies in text-completion exercises (6.3).

Contributed measures In order to address the difficulty dimensions word difficulty and candidate ambiguity, new computational measures were required. The new measures are described and evaluated on sub-tasks in the corresponding chapters and outperform existing measures.

- Implementation of existing measures for text and word difficulty (4.1, 5.1). All features that do not rely on proprietary resources are publicly available in the *DKPro TC* framework (Daxenberger et al., 2014).
- Development of an approach for cognate production to capture an important aspect of receptive word difficulty (5.2).
- Development of computational measures for predicting spelling difficulty to capture an important aspect of productive word difficulty (5.3).
- Development of computational measures for evaluating candidate ambiguity for language learners (6.2).

**Contributed resources** The following new resources and extensions of existing datasets are available for research purposes. An overview page with all links can be found at: www.ukp.tu-darmstadt.de/research/past-projects/exercise-difficulty-for-language-learning/

- Extension of a dataset of existing cloze exercises (Zweig and Burges, 2012) with difficulty information obtained from language learners (3.3.1).
- Collection of cognate resources for seven language pairs including the languages English, Czech, Farsi, German, Greek, Italian, Spanish, and Russian (5.2).
- Extraction of English, German, and Italian spelling errors from three corpora of annotated learner essays and determination of the spelling error probability of each word in the corpora (5.3).
- A new dataset of difficulty predictions for text-completion exercises by three human experts for 398 items (7.1).
- Extraction of 28,097 English-German synonyms from UBY (Gurevych et al., 2012) that are enriched with cognateness and spelling difficulty information (8.1.4).

**Contributed applications** In chapter 8, the difficulty prediction approach is used for the goal of automatic difficulty manipulation leading to the following applications.

- A web demo that automatically generates exercises from texts and provides a visualization of the predicted difficulty (8.1.1).
- An approach for automatic exercise generation and selection from a text corpus (8.1).
- An approach for distractor manipulation that is based on the measures for candidate ambiguity (8.2).

#### 1.5 Publication Record

Parts of this thesis have been previously published at peer-reviewed conferences and in internationally recognized journals in the fields of natural language processing and computational linguistics.<sup>4</sup> The concept of text difficulty for language learning (chapter 4) is described in Beinborn et al. (2012). Beinborn et al. (2014b) elaborates on the topic with a psycholinguistic comparison of L1 and L2 readability. The discussed consequences for L2 readability are an increased focus on word difficulty in general and on cognates in particular (chapter 5). A computational approach for cognate production (section 5.2) is introduced and evaluated in Beinborn et al. (2013). Cognates facilitate text comprehension, but trigger many spelling errors. We developed a computational approach for spelling difficulty prediction and evaluated it on spelling errors extracted from learner corpora in Beinborn et al. (2016). The difficulty prediction for English C-tests has first been published in Beinborn et al. (2014a). In Beinborn et al. (2015a) the approach is extended to other exercise types and languages (chapter 7). As this extension requires a re-evaluation of the candidate space, new measures for candidate ambiguity have been introduced (section 6.2).

The main thesis content has been presented at an invited talk at the *FEAST* series at Saarland University (*Forum Entwicklung und Anwendung von Sprach-Technologien*). Inter-disciplinary exchange has been supported by presenting the work at the workshop *Virtualisierung von Bildungsprozessen* organized by the *Forum Interdisziplinäre Forschung*, in the academic network *Individualisiertes Sprachenlernen*, and at the international conference *Language.Learning.Technology. 2015* (Beinborn et al., 2015b). These events convene researchers from computer science, linguistics, psychology, and education and produce synergies by connecting knowledge from different strands of research.

<sup>&</sup>lt;sup>4</sup>A detailed list can be found on page 226.

#### CHAPTER 2

# **Exercises for Language Learning**

"Learning another language is not only learning different words for the same things, but learning another way to think about things."

Flora Lewis

This chapter analyses the expectations for exercises in the area of language learning. The first section discusses the requirements for high quality exercises. The second section provides an overview of the possibilities and limitations of educational natural language processing. The third section introduces text-completion exercises which are a popular choice for language proficiency tests.

## 2.1 Educational Criteria for Language Learning Exercises

In 2001, the European Union published the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) which is considered an important milestone towards international comparability of language skills (Council of Europe: Language Policy Division, 2011). The CEFR categorizes language proficiency into 6 levels: A1, A2, B1, B2, C1, C2. A1 indicates beginner knowledge and C2 corresponds to near-native proficiency. The proficiency levels are expressed as a combination of the four skills reading, writing, listening, and speaking. Listening and reading are considered to be receptive activities (understanding), whereas speaking and writing are productive activities. These activities are often trained separately and thus require different types of exercises. We will focus on textual activities; exercises requiring speech technology such as pronunciation or dialogue training are beyond the scope of the thesis.

#### 2.1.1 Exercise Quality

The quality of exercises has been frequently discussed in the context of language testing (Hughes, 1989; Grotjahn, 2009). Language testing can be categorized into *summative assessment* and *formative assessment*. Summative assessment refers to diagnostic tests which usually take place after a learning period and categorize the learner's proficiency at this specific point in time. In contrast, formative assessment is defined as performance feedback during the learning process to assure continuous evaluation (Fulcher and Davidson, 2007). The issue of exercise quality has been thoroughly analyzed for summative assessment which includes placement tests, achievement tests and proficiency tests. These tests are also called *high-stakes tests* because the outcome determines whether the learner receives a required certificate and might even have financial consequences.

A first step towards improving fairness and comparability for language testing was the introduction of *criterion-referenced testing* (Glaboniat, 2010). In criterion-referenced testing, the learner's result is compared to an evaluation criterion that has been determined prior to the test. This procedure assures that the result of an individual is evaluated independent of the results of his peers as in norm-referenced testing. If an individual's performance is only judged with respect to an expected performance on a standardized scale, results can be compared across samples and also across languages. The CEFR has been established as the standard reference system in Europe and all major language testing institutions have linked their language certificates to the benchmark descriptions of the CEFR levels.

The exercises in language tests need to fulfill quality criteria such as objectivity, reliability and validity which are explained below. In summative assessment, the learner performance is compared to a desired standard independent of prior knowledge. In formative assessment, the learning curve of the individual is more relevant and the performance is compared with the expected performance based on the individual's previous learning progress. In the classical setting, a teacher monitors the learning progress and provides continuous formative assessment by correcting and evaluating the learner's activities and adjusting the difficulty of the learning tasks. Since the introduction of computer-based language learning, formative assessment has received increasing attention.

Quality criteria The three most important quality criteria for assessment are objectivity, reliability and validity (American Educational Research Association et al., 1999). The following explanations are based on overviews by Hughes (1989) and Grotjahn (2009) who review these criteria in the light of language testing.

Objectivity of a test can be accomplished if the evaluation of a test is independent from the test organizer and the corrector. The test conditions should be comparable for all participants and the evaluation should be consistent. This means that the same test outcome should always receive the same grade independent of the corrector. In order to assure objectivity, detailed evaluation guidelines are required to specify which performance is expected

for which grade. The CEFR has provided guidelines that describe the expectations for the six proficiency levels. A CEFR manual (Council of Europe: Language Policy Division, 2009) and the recommendations of the Association of Language Testers in Europe (ALTE)<sup>5</sup> help to link these scales more directly to language tests. Exercises with a closed answer set usually exhibit higher objectivity because they specify pre-determined answers and leave no room for interpretation in the evaluation.

Reliability refers to the accuracy and the reproducibility of test results. Reliable tests rate the language proficiency of a participant consistently and exhibit only a minimal level of measurement error. If the same participant completes two comparable versions of a test, the results should be highly correlated. Reliability is thus strongly connected with objectivity. With the introduction of the CEFR, it has become common practice to analyze the reliability of high-stakes tests in extensive pilot studies.

Validity is the most complex quality criterion for language tests. According to Brown (1989, p. 59), "the validity of a measure may be defined as the degree to which it is measuring what it claims to be measuring". Success in an exercise that is supposed to measure language proficiency should thus not depend on general intelligence or world knowledge. As this concept is rather vague and difficult to operationalize, validity is often further broken down into criterion validity, content validity, construct validity and face validity. Criterion validity indicates that the test correlates with an expected outcome. Good placement tests, for example, should predict the learning success of a participant in a specific course. Content validity captures the representativeness of the test for the measured construct. For language testing, content validity can be improved if the exercises are based on a representative sample of authentic language usage. A test is considered to have high construct validity if it measures the intended construct. As this is hard to determine, construct validity is approximated by measuring the correlation of the test with already established measures of the construct (also called *convergent validity*). A language proficiency test should correlate with the participant's school grades for this language and with other language proficiency tests. It should not correlate with variables that are considered to be irrelevant for the construct such as age or sex (divergent validity). Face validity refers to the transparency of the test and considers the perspective of the test participants. The exercises of a test with high face validity should be intuitive for the participants and generally be considered as a good and relevant measure for language proficiency.

Objectivity, reliability and validity are the most important quality criteria for language tests and exercises. Additional aspects are often summarized as the *usability* or the *economy* of a test. A test that takes ten hours and requires complex equipment is considered to be less useful than a short and inexpensive test. These criteria have been established for traditional assessment practice. In the last decade, computer-assisted language learning and testing started to replace pencil-and-paper tests. As the exercise types and the testing

<sup>&</sup>lt;sup>5</sup>http://www.alte.org, accessed: January 21, 2016

procedures are being adapted to the change of the medium, new requirements for exercise quality are emerging which are discussed below.

Exercises for computer-assisted language learning The practical developments in computer-assisted language learning (CALL) started rather slowly, although the topic attracted many theoretical debates (Blake, 2011; Garrett, 2009). In an overview of CALL tools that make use of natural language processing technology, Amaral and Meurers (2011) describe only three stand-alone systems (one for Japanese, German, and Portuguese each). In the last years, the trend is moving from computer-assisted to mobile-assisted language learning (MALL) and we see a wave of newly emerging applications for language instruction. While overviews of computer-assisted language learning by Levy (2009), Mitschian (2010), and Blake (2011) focus mainly on tools serving a single purpose like vocabulary or grammar training that can be used to complement traditional courses, more recent language learning applications aim at covering a wider range of language proficiency training. The most popular examples like Duolingo, busuu, and babbel attract thousands of users who cherish the non-institutional setup that allows them to integrate language learning into their free time (Hockly, 2015). Academic research that embeds the mobile technologies into theories of language learning and evaluates them based on empirical experiments is lagging behind the fast developments on the application side (Viberg and Grönlund, 2012).

Most research overviews focus instead on institutionalized computer-based language testing. The most popular language certificates can now be obtained via computer-based tests (e.g. *TOEFL* and *IELTS* for English, *TestDaF* for German).<sup>6</sup> Pathan (2012) lists many advantages of computer-based testing such as the independence of administrative and logistic burdens and guaranteed uniform testing conditions. In addition to these practical advantages, computer-based tests are often praised for achieving a greater authenticity due to the integration of multi-media exercises (Joseph and Uther, 2009). On the other hand, electronic tests also pose limitations on the exercise types, for example, due to the size of the screen (Chinnery, 2006).

Despite the flexibility of the medium, the exercises in language tests and in mobile language learning applications are quite static and often simply resemble the exercises in traditional pencil-and-paper tests. In this case, the main advantage of computer-based language learning is the availability of additional data. Keystroke logging and semi-automatic analyses of processing times and errors allow insights into test-taking strategies from hundreds of participants that were not available in traditional testing. These datasets form a promising basis for language testing research and will lead to further improvements. The approach in this thesis shows that the analysis of existing results makes it possible to predict the difficulty of new tests. This is an important step towards more adaptive exercise generation.

<sup>&</sup>lt;sup>6</sup>www.ets.org/toefl, www.ielts.org, www.testdaf.de, all accessed: January 21, 2016

The most promising progress in *computer-based testing* is expected from advances in automatic scoring for open exercise formats that require free text answers (Douglas and Hegelheimer, 2007). The delivery of immediate feedback even for complex tasks would open up many new possibilities. The technology for automatic scoring has seen a lot of progress (see section 2.2.3). The *Educational Testing Service* (the leading institution of computer-based testing in the US) has adopted the policy to score essays by combining the output of their automatic evaluation system *e-rater* with the evaluation of a human rater (Attali and Burstein, 2006). This has led to an improvement over the policy of using two human raters, but still requires a large amount of human effort.

Computer-based learning and testing is often advertised as a means to support *self-directed learning* because learners can follow their own pace. However, in order to adapt the exercises to the needs of the learner, the difficulty of the items needs to match the learner's ability. This goal is the underlying paradigm for *computer-adaptive testing*. The first computer-adaptive test was introduced already in 1985 and addresses the problem that most language tests are only informative about high- and low-scoring participants and do not discriminate well for intermediate levels (Larson and Madsen, 1985). All items of this test had been evaluated in extensive pre-tests and the difficulty was calibrated using models from item response theory. During the computer-adaptive test, the items are administered to the participants depending on their performance on previous items. Participants who have solved an item are presented with a more difficult item, those who have failed continue with an easier item. This procedure leads to more individualized and considerably shorter tests because the participants only need to answer those items that are discriminative for their ability level (Meunier, 2013). This approach requires a model for the difficulty of all available items.

After a period that praised computer-adaptive testing as a breakthrough in language assessment, the interest quickly declined again because of the unfeasible demands. Douglas and Hegelheimer (2007) and Meunier (2013) highlight that item pools need to grow tremendously in order to satisfy the needs for computer-adaptive testing and still assure high item variation to undermine cheating attempts. Therefore, most language testing institutions are hesitating to establish the computer-adaptive testing paradigm. The popular *TOEFL* test, for example, does currently not contain any computer-adaptive sections.<sup>7</sup> A notable exception is the computer-adaptive *Business Language Testing Service (BULATS)* that assesses the proficiency of English, French, Spanish and German in the business domain (Perlmann-Balme, 2010).<sup>8</sup>

Computer-adaptive testing is particularly attractive for the scenario of formative assessment. The idea of providing direct feedback and to adapt the exercises to the learner's

<sup>&</sup>lt;sup>7</sup>Information from the *Educational Testing Service* by Aiofe Cahill, Managing Research Scientist, and Teresa Axe, Associate Director Global Education, January 20, 2016.

<sup>8</sup>http://www.bulats.org, accessed: January 19, 2016

abilities is in line with the goal of using formative assessment to trigger learning progress. The leading online course *Duolingo* already claims to use adaptive learning modules that take the past performance of the user into account (Settles, 2013). However, *Duolingo* has also been criticized for presenting the user with decontextualized sentences (Hockly, 2015) due to its commercial model of collaborative translation. In order to make computer-adaptive testing more feasible for a wider range of exercises, technical improvements in automatic scoring and exercise generation are required to overcome the laborious composition of manually created item banks. Instead, exercises could be generated on the fly to closely match the individual requirements of the learner. In section 2.2.2, we see that the automatic generation of exercises is already possible for some exercise types. However, predicting the quality and the difficulty of a generated exercise has not yet received enough attention.

Conclusions We conclude that good exercises should follow the quality criteria objectivity, reliability and validity and should be usable for computer-adaptive testing. The last condition demands automatic scoring and difficulty estimation of the exercise. A common obstacle for computer-adaptive testing is the need for large calibrated item pools. If the difficulty of exercises could be manipulated automatically, this hard constraint would be alleviated. The following section provides an overview of existing exercise types for text-based tasks.

#### 2.1.2 Exercise Types

In previous work, text-based exercises have been categorized into hierarchies depending on the educational knowledge level they address. These hierarchies differ slightly depending on the purpose of the exercise. Generally, two purposes are distinguished: *comprehension exercises* make a presented content more comprehensible and *linguistic exercises* train a particular phenomenon to improve the learner's vocabulary or grammar skills. In the first case, the learner is encouraged to reflect upon the input to reach a deeper understanding. The most popular hierarchy for comprehension exercises is *Bloom's categorization of educational objectives* (Bloom et al., 1956). In the second case, the learner focus is directed towards a certain linguistic phenomenon by highlighting different occurrences, pointing out exceptions and encouraging repeated usage. A hierarchy of text-based linguistic exercises has been provided by Wesche (1996).

Table 2.1 provides an overview of these hierarchies including examples of the corresponding exercise types. The descriptions above refer to the formative goals of the exercise, but both exercise types can also easily be applied in summative testing to evaluate the learner's abilities.

<sup>&</sup>lt;sup>9</sup>Co-founder Luis van Ahn explains *Duolingo*'s commercial model in a TED talk: <a href="http://www.ted.com/talks/luis\_von\_ahn\_massive\_scale\_online\_collaboration">http://www.ted.com/talks/luis\_von\_ahn\_massive\_scale\_online\_collaboration</a>, accessed: November 25, 2015

	Objective	Typical exercises
	Knowledge Comprehension Application	Who, what, when, where, why, how? Describe X. Summarize X. How is X an example of Y? How is X related to Y? Why is X significant?
Comprehension Exercises (Bloom et al., 1956)	Analysis Synthesis	What are the parts or features of X? Classify X. What would you infer from X? What ideas can you add to X? How would you design a new X? What would happen if you combined X and Y? What solutions would you suggest for X?
	Evaluation	Do you agree that X? What do you think about X? What is the most important X? Place the following X in order of priority. How would you decide about X? What criteria would you use to assess X?
Linguistic Exercises (Wesche, 1996)	Selective Attention Recognition Manipulation Interpretation Production	Identification of highlighted words Matching exercises Morphological exercises, shuffle exercises Odd-one-out exercises, tribond exercises Translation, Cloze exercises

**Table 2.1:** Overview of text-based exercises and the corresponding educational objectives for language learning

**Comprehension exercises** Comprehension exercises encourage the learner to reflect upon a topic and gain a better understanding. They are often expressed as classical questions starting with an interrogative pronoun and ending with a question mark.

Bloom classified exercises according to the associated cognitive educational objective. These objectives can be mapped to classical questions as in the upper part of table 2.1.<sup>10</sup> The educational objectives are ordered by increasing complexity and are hierarchically organized, i.e. higher learning objectives like *synthesis* and *evaluation* subsume lower ones like *knowledge* and *comprehension*. The exercise types range from simple factual questions up to complex essay writing. The evaluation level also encompasses transfer tasks that encourage the learner to apply the new knowledge and skills in a slightly different context.

The questions described above are targeted at the content of the educational material. In this setting, the language is only the medium of teaching and the exercises aim at the comprehension of the content. In language teaching, textual content often deals with cultural differences, traditions or historical events in the countries of the target language. However, the main educational content in language learning is the language itself and although language proficiency is also implicitly contained in comprehension questions, linguistic exercises that are explicitly targeted at mastering the language are required as well.

<sup>&</sup>lt;sup>10</sup>Compare https://en.wikipedia.org/wiki/Question, accessed: December 9, 2015

**Linguistic exercises** Linguistic exercises are applied to train a particular linguistic phenomenon by forcing the learner to use it repeatedly, highlighting different use cases and pointing out exceptions. This includes grammar exercises which train, for example, the usage of verb tenses, but also vocabulary exercises which train the usage of new words.

As Bloom's taxonomy can only vaguely be matched to these exercise types, Wesche (1996) proposes a slightly different taxonomy to classify text-related vocabulary exercises according to the required language competence (see the lower part of table 2.1). The taxonomy follows again an ascending order of complexity. In exercises for selective attention, the words of interest are visually highlighted to guide the learner. The task for the learner is rather simple and usually consists in reading or noting down these words. In recognition exercises, the focus elements are provided and only partial knowledge is required from the learner. A typical example are matching exercises that require the learner to match a target word with its definition, its translation, a sample context, a picture, or a synonym. In manipulation exercises, the learner applies morphological operations on a word stem or grammatical transformations on a group of words. A typical example are shuffle exercises that present a group of words (or word stems) in irregular order and ask the learner to produce the corresponding sentence (Perlmann-Balme, 2010). Interpretation exercises require deeper semantic analysis of the relationships between words. In odd-one-out questions, for example, the learner needs to identify one word from a group that does not fit with the others (Colton, 2002). For example, the word sit is the odd-one-out in the group [walk, sit, run, joq/ because it does not denote a movement of the legs. Tribond questions are similar; in this case the learner is asked to identify the shared property of a group of associated words. Wesche (1996) classifies production exercises as the most demanding type because they require recall and reconstruction. The learner needs to actually produce the item, for example, as a translation or to fill the gap in a sentence (the latter is commonly known as a cloze exercise).

Due to the over-use of structural pattern drills in language education in the 1960s (Paulston, 1970), linguistic exercises (and in particular grammar exercises) have been stigmatized as mechanical means of instruction that only train isolated phenomena and do not contribute to integrative language knowledge (Gartland and Smolkin, 2015). Modern language teaching curricular recommend to introduce grammar exercises with respect to communicative needs and not as an independent learning objective (Garrett, 2009). However, many advocates of computer-assisted language learning have analyzed that extracurricular training of isolated linguistic phenomena has a positive effect on language proficiency (Liou, 2013). Warschauer and Healey (1998) propose a shift towards "integrated CALL", i.e. to integrate linguistic exercises into more authentic contexts. In practice, this could be obtained by combining comprehension exercises and linguistic exercises.

We have seen that comprehension and linguistic exercises can both be mapped to taxonomies of cognitive objectives. It is generally assumed that higher cognitive objectives correspond to higher exercise difficulty. In the educational literature, the main distinction is commonly made between *recognition* and *production* exercises. In both taxonomies, productive exercises are considered to be more demanding than recognition exercises. For many exercises, the answer format determines whether an exercise requires recognition or production skills.

Answer formats The complexity of an exercise can be manipulated by the *answer format* which can be distinguished into *open*, *half-open* and *closed* (Grotjahn, 2009). In the case of open exercises, the set of potentially correct answers is large and the corrector has to interpret the answer provided by the learner. Exercises with open answer formats correspond mainly to the higher levels of Bloom's taxonomy of educational objectives and require very detailed guidelines for the learner to understand the demands of the task and for the corrector to assure reliable scoring. Typical examples for open exercises are free text answers that express an argument or an opinion.

Most of the linguistic exercises described above are either half-open or closed. In half-open exercises, only a single answer is correct and it is known to the corrector. In closed exercises, the correct answer is also available for the test taker who only needs to select it from a set of options. The most popular closed answer format is the *multiple choice* option that presents the correct answer along with several distractors (see the cloze exercise in Figure 1.2 for an example). For so-called *multiple response* exercises, more than one option is correct and the learner should identify all of them. Matching exercises consist of multiple items and a word bank with all solutions. The learner then needs to match each item with the corresponding solution. In a simpler variant of closed formats, the learner simply needs to make a binary decision whether a statement is true or false.

Exercises corresponding to the production level in the taxonomy by Wesche (1996) are usually designed as half-open or open formats, whereas recognition exercises are per definition closed formats. In standardized tests, half-open and closed formats are often preferred over open formats because they ensure higher reliability (Glaboniat, 2010). Comprehension exercises are usually either closed or open formats. Half-open comprehension exercises are only reasonable for factual questions.

#### 2.1.3 Conclusions

We have seen that there exists a wide range of text-based exercise types. Each exercise type serves a slightly different educational objective and requires different skills. The choice of exercises depends on the learning scenario. From the above analyses, we conclude that the following desiderata are important indicators for high-quality exercises that can be used for fast language proficiency tests:

(1) The exercises should fulfill the quality criteria objectivity, reliability, and validity.

- (2) It should be possible to evaluate the exercises automatically.
- (3) It should be possible to manipulate the content and the difficulty of the exercise automatically to support formative assessment.
- (4) The exercises should address receptive and productive skills.
- (5) Ideally, the exercises should integrate comprehension of authentic texts and training of linguistic phenomena.

In the following section, we discuss how the requirements correspond with the state of the art in natural language processing.

## 2.2 NLP for Language Learning Exercises

Computer-assisted language learning employs a wide range of technologies for language instruction (see Golonka et al. (2014) for an overview). The most interesting developments in computer-assisted language learning from a computational perspective are subsumed under the term *intelligent computer-assisted language learning (ICALL)*. The *Computer-Assisted Language Instruction Consortium* defines ICALL as "an interdisciplinary research field integrating insights from computational linguistics and artificial intelligence into computer-aided language learning." The technology aspect of computational linguistics is usually referred to as *natural language processing (NLP)*. In a detailed overview, Meurers (2012) lists only three NLP-based language tutoring systems that aim at autonomous and complete tutoring of language learners, namely *E-Tutor* (Heift, 2010), *Robo-Sensei* (Nagata, 2009), and *TAGARELA* (Amaral and Meurers, 2011). However, he describes many tasks in computer-assisted language learning that can benefit from the application of natural language processing. This section provides an overview of these tasks and groups them into different stages of language instruction: content selection, exercise generation and learner evaluation.

#### 2.2.1 Exercise Content Selection and Manipulation

One important aspect of preparing successful learning and teaching is the selection of learning material. In natural language processing, the selection of suitable texts has received considerable attention, while the selection of other educational resources such as images, figures and audio/video samples is mainly discussed in pedagogical and inter-cultural studies.

Especially for language learning, the use of authentic material is encouraged to confront the learner with real-world language use (Gilmore, 2011). However, authentic material is often too complex for learners because of the uncontrolled occurrences of unknown words and constructions. Finding suitable textual resources for educational purposes that fit the

<sup>&</sup>lt;sup>11</sup>https://calico.org/page.php?id=363, accessed: October 28, 2015

learners' level is a challenging and tedious task. There exist two approaches for automatically determining appropriate content. *Readability* measures can evaluate the text difficulty of a large body of texts and select the ones that best fit the intended readability level. *Simplification* approaches work on a single text instead and modify the most difficult elements to lower the difficulty of the text. In her thesis, Vajjala Balakrishna (2015) has examined both approaches in detail.

Readability Research on the development of automatic measures for assessing the difficulty of texts, also referred to as readability, has a long tradition. Early measures for readability like the Flesh-Kincaid Reading Ease (Kincaid et al., 1975) approximate readability only based on the average word and sentence length of a text. This approach is still quite popular, although it is a very simplistic approximation of text complexity. More sophisticated approaches consider a large amount of linguistic features covering morphological, lexicalsemantic, syntactic, and discourse aspects of readability and reach very high prediction quality (Feng et al., 2009; Vajjala and Meurers, 2012). A detailed description of readability research for language learners is provided in chapter 4. The automatic measures allow to instantaneously estimate the difficulty of a large number of texts. The REAP engine (Heilman and Eskenazi, 2006) and the CohMetrix system (Graesser and McNamara, 2004) provide web demos to estimate the readability for any text. <sup>12</sup> For REAP, the output simply consists of a US school grade for which the text is considered to be appropriate. CohMetrix provides a very detailed analysis of multiple readability features. However, both systems are trained for native speakers as target readers and not language learners. <sup>13</sup> Unfortunately, even a text that has been evaluated to exhibit a moderate difficulty on average, might still be considered as impractical by the teacher because of the presence of constructions that are not yet mastered by the targeted learners (e.g. passive).

Text simplification In practical scenarios, teachers often manipulate authentic material by deleting, substituting and re-ordering words or phrases to reduce the lexical and syntactic complexity. The research area of text simplification aims at automatizing this process. The goal is to transform a complex text into a simpler one using various simplification operations such as splitting a long sentence into two shorter ones. Siddharthan (2014) provides a good overview of simplification approaches. The first approaches to text simplification focused on syntactic simplification based on parse trees and pattern rules for simplifying the trees. Syntactic simplification of sentences is often also called sentence compression and can be an important pre-processing step for other tasks like automatic summarization. More recently, simplification has often been targeted as a monolingual translation

 $<sup>^{12} \</sup>rm http://cohmetrix.com,\, http://reap.cs.cmu.edu,\, both\, accessed:$  November 25, 2015

<sup>&</sup>lt;sup>13</sup>Note that (Heilman et al., 2007) aims at determining readability for language learners. However, the findings do not seem to be incorporated in the web demo.

task from sentences in the standard *Wikipedia* to the corresponding sentence in the *Simple English Wikipedia*. Phrase-based machine translation systems like *Moses* (Koehn et al., 2007) have been quite successful in particular for lexical simplification tasks. Unfortunately, a simplified version of *Wikipedia* exists only for English and comparable resources for other languages are hard to find. The quality of simplification is difficult to measure and there does not yet exist a good method to reliably compare the state-of-the-art systems. Recent experiments by Vajjala Balakrishna (2015) show that simplification approaches are not yet robust enough when applied to different corpora.

Simplification approaches originally targeted younger or language-impaired readers reading in their mother tongue. Petersen and Ostendorf (2007) and Crossley et al. (2007) performed semi-automatic corpus analyses of simplified texts and started focusing on the needs of language learners. However, the usability of simplified texts for language learning has also been subject to strong debates. Siddharthan (2014) summarizes that the most frequently expressed concern is that text simplification denies learners the opportunity to learn from authentic and natural input.

Readability and simplification can be seen as two aspects of text difficulty: measuring difficulty and manipulating difficulty. Aluisio et al. (2010) aim at identifying the readability of texts to identify simplification needs for poor readers. Pilán et al. (2014) and Vajjala and Meurers (2014) proceed similarly, but they focus on the readability of individual sentences and target language learners. In analogy to the work on readability, this thesis aims at measuring and predicting exercise difficulty to leverage the task of automatically manipulating exercise difficulty. Text simplification explicitly targets only one direction of difficulty manipulation, but for adaptive language learning both directions are important. We aim at making exercises more difficult for advanced learners and less difficult for beginners.

Collins-Thompson (2014) points out that text difficulty is not only determined by the text, but also by the characteristics of the reader. The same holds for exercise difficulty. To determine the absolute difficulty of an exercise for a specific reader, her background knowledge, interests, and learning motivation need to be taken into account. In this thesis, we approximate the difficulty of exercises relative to a sample of learners. This concept of difficulty is discussed in more detail in chapter 3.

#### 2.2.2 Exercise Generation

In order to successfully support language learning, it is important to generate exercises that fit the learner's background and proficiency level. To facilitate this task, NLP researchers aim at automating the process at least partially so that the human effort can be minimized to a final review and selection step. This section provides an overview of existing *exercise gen*-

<sup>&</sup>lt;sup>14</sup>https://simple.wikipedia.org

*eration* approaches for comprehension exercises and linguistic exercises. Text-completion exercises integrate language comprehension and linguistic skills and are discussed separately.

Generating comprehension exercises Comprehension exercises are also known as "classical questions". Almost all automatic approaches for the generation of comprehension exercises work sentence-based, i.e. they transform a sentence into the question form by replacing one of its elements with an interrogative pronoun and reordering the syntactic structure. Not all sentences of a text are good candidates for a question. Most approaches first apply term extraction and summarization methods on the document to select the sentences that contain salient information. For the actual transformation of the selected sentences into questions, different methodologies can be applied: pattern-based, syntax-based, and semantics-based approaches. Pattern-based approaches directly work with surface realizations and are usually combined with a POS-tagger and a named entity recognizer. If manually designed patterns are detected in the input sentence, a corresponding question is generated according to predefined templates and rules (Mitkov et al., 2006; Chen et al., 2009; Rus et al., 2007; Curto and Mendes, 2012). Syntax-based approaches operate on parse trees. They transform a declarative sentence into an interrogative one by applying grammar rules such as wh-movement and subject-auxiliary inversion (Heilman and Smith, 2010; Wyse and Piwek, 2009; Kalady, 2010; Agarwal et al., 2011; Bernhard et al., 2012). Semantics-based approaches build a semantic representation of the sentence and take this representation as input for a generation grammar. Yao et al. (2012) use a combination of minimal recursion semantics and an HPSG generation grammar and use transformations to get from a declarative to an interrogative representation. Olney et al. (2012) maintain a situational model by building up conceptual maps from the input. They then use templates to generate questions from the conceptual maps. As the generation mechanisms are quite error-prone, several systems follow an overgenerate-and-rank approach (Heilman and Smith, 2010; Chali and Hasan, 2015). They generate multiple questions and apply ranking and filtering techniques to eliminate the questions that are ill-formed. This leads to a strong improvement of the quality, but according to the results reported by Heilman and Smith (2010) only 52% of the top-ranked questions were considered to be acceptable by human judges. In a more recent evaluation, Chali and Hasan (2015) compared their system, which employed a more sophisticated ranking algorithm, with the system by Heilman and Smith (2010) and reported an even lower ratio of acceptable questions (below 50% for the 15% top-ranked questions for both approaches).

The answer for the automatically generated questions of the described approaches can always be found directly in the text. Bloom's higher educational objectives require more complex reflection on the topic and deeper evaluation of arguments and are therefore not met by the current question generation approaches. One of the first attempts towards

deeper assessment of reading comprehension is the approach by Ai et al. (2015). They use relation extraction techniques and paraphrasing patterns to generate multiple-choice inference questions. The learner is presented with a text and with multiple statements that express a relation between two entities. The task is to select the statement that expresses a relation which can be inferred from the information provided in the text.

All of the mentioned approaches aim at improving general understanding of texts and do not specifically target language learners. In 2010, a shared task on automatic *question generation* (*QG2010*) attracted the participation of four research groups. The task consisted in generating questions for a short text according to a specified question type. Rus et al. (2012) discuss the results of the subsequent human evaluation and report high average scores for the quality of the systems. Unfortunately, the majority of generated questions contain syntactic inconsistencies. While these syntactic flaws might be negligible in tutoring systems for native speakers that focus on the question content, the input needs to be impeccable for language learners. Flawed questions are more likely to have a negative effect on the learning progress and on the motivation of language learners. We have seen that generation approaches for comprehension exercises struggle with syntactic inconsistencies when applied to new content. Linguistic exercises are usually more robust because the same exercise format can be applied to a large group of words in a repetitive way without adjustments.

Generating linguistic exercises Linguistic exercises can be roughly distinguished into *vocabulary* and *grammar exercises*. The existing approaches to automatic generation of vocabulary exercises all rely on additional resources that provide information about words and their relationships with other words. Brown et al. (2005) develop automatic approaches for six different vocabulary exercises. The tasks include mapping words to a corresponding definition, to a synonym, antonym, hypernym, or hyponym of the word, or to a suitable example context. The exercises are presented in two closed answer formats. In the word bank format, answers and exercises need to be matched and in the multiple-choice format, the answer needs to be selected from a set of distractors. The authors find that student performance on the automatically generated questions correlates well with the performance on human-generated questions and on a standardized vocabulary test. In more recent work, Susanti et al. (2015) follow a similar approach to generate synonym questions but the target word is presented in an authentic context retrieved from the web. The contextual evaluation requires additional word sense disambiguation to select a suitable synonym.

Both approaches above are based on semantic relations from *WordNet* (Fellbaum, 1998). *WordNet* is a very rich resource, but it necessarily has a limited coverage. To avoid the dependence on limited resources, Heilman and Eskenazi (2007) base their generation approach for vocabulary exercises on a thesaurus which is automatically extracted from a text corpus. See below an example of their questions (the solution is C):

Which set of words are most related in meaning to reject?

- A. pray, forget, remember
- B. invest, total, owe
- C. accept, oppose, approve
- D. persuade, convince, anger

Although the question format is quite static, the exercise can be classified as an interpretation task and can be generated for many content words. The challenge lies in the generation of suitable answer sets. The authors use a dependency parser and extract relations between words based on the measure of mutual information. Words that occur in the same dependency relation with another word are then considered to be related. Heilman and Eskenazi (2007) asked a teacher to evaluate the generated questions and found that 68% of the questions were usable. For the unusable questions, the solution was either too difficult due to ambiguous word relations or too easy. It can be concluded that the *quality* of automatically generated vocabulary exercises is higher than for comprehension exercises, but anticipating and manipulating the difficulty of these exercises appropriately is challenging. Sun et al. (2011) make a first step towards applying psycholinguistic findings for the manipulation of exercise difficulty. They create simple wordbank exercises that require the learner to match words with their definition, but they include an additional difficulty factor to the generation of the wordbank. Based on the observation that Chinese learners struggle with phoneme-to-grapheme mapping in English, the authors aim at grouping words that look similar and are closely related in meaning (e.g. transform and transfer) to raise awareness for the differences.

Highlighting the contrast between different word forms is also an important aspect of grammar exercises. Schmidt (1990) analyze the important role of consciousness in second language learning and argue for a focus on form. Based on this idea, Meurers et al. (2010) enhance web pages with visual input for language learners. In noticing exercises, they focus on specific grammatical phenomena (e.g. the difference between gerunds and to-infinitives) and highlight the different use cases in authentic texts. In addition, they also provide practice exercises to train the phenomena. Shuffle exercises, for example, present a sentence or a question in random order and require the learner to re-arrange the words correctly. Perez-Beltrachini et al. (2012) generate similar shuffle exercises for French. As additional difficulty, they present the shuffled words as lemmas forcing the learners to pay attention to word order and inflection at the same time. Aldabe et al. (2006) generate exercises for training the complex morphology of Basque. They create error correction exercises which require the learner to recognize the error in a wrong word form, and word formation exercises that provide the lemma and ask the learner to produce the correct inflection for a given pattern. Bick (2004) uses natural language processing to develop games that ask learners to label words with classes and sentences with morphological and syntactic features to pass levels. Almost all of the approaches towards vocabulary and grammar exercises described

above, additionally implement approaches for generating text-completion exercises which are explained in more detail below.

Generating text-completion exercises *Text-completion exercises* combine the training of vocabulary with the comprehension of context. This is considered to be a more authentic task than determining semantic relations of isolated words (Smith et al., 2009). The most popular type of text-completion exercises in natural language processing is the *cloze exercise* (see figure 1.2). A cloze exercise consists of a sentence in which one word (the *key*) has been replaced with a gap and the learner is asked to fill in the gap. In order to simplify the task, cloze exercises are usually presented in multiple-choice answer format. Cloze exercises can be generated for a variety of learning goals. Most automatic approaches for cloze exercises focus on factual comprehension or vocabulary knowledge, but they can also be used for directly testing linguistic phenomena. Cloze exercises are not the only option for text-completion exercises. In psychological and educational research, other text-completion exercise types like the *C-test* and the *X-test* have been discussed, but they have not yet attracted interest in natural language processing research. These exercises are introduced in more detail in section 2.3 as part of the text-completion principle.

Generating a cloze exercise first consists of choosing a key and a corresponding context. In a second step, the wrong answer options (also called the *distractors*) are generated. In most approaches, either the key or the context are already provided depending on the application scenario.

If the learner should practice specific key terms, corresponding contexts need to be created for them. Brown et al. (2005) extract the example contexts for a given key from *WordNet* and Gates (2011) use definitions from a learner dictionary. Pino et al. (2008) and Smith et al. (2009) aim at finding suitable contexts for given keys in unstructured text corpora based on pre-defined criteria. Instead of selecting existing sentences, Liu et al. (2013) explicitly generate typical example context from the *Web1T* n-grams (Brants and Franz, 2006).

In another scenario, the focus lies on reading comprehension of specific texts. This means that the context is already provided and the challenge lies in identifying a useful key, i.e. to determine a good placement for the gap. In this case, a common approach is to over-generate potential cloze exercises from the given input data and use a classifier trained on existing human-generated cloze exercises to filter out the flawed instances. Becker et al. (2012) used a large set of linguistic features to distinguish between high-quality and low-quality cloze exercises. Niraula and Rus (2015) build on this approach and add an active learning step that requires human judges to label additional instances to improve the classifier. These approaches are comparable to the previously introduced *overgenerate-and-rank* approaches for comprehension questions (Heilman and Smith, 2010; Chali and Hasan, 2015).

Once the key and the context are determined, the distractors need to be generated. For grammatical cloze exercises, this task is relatively easy because they either target closed classes like prepositions (Lee and Seneff, 2007) or they test regular grammatical processes like different verb tenses (Meurers et al., 2010; Chen et al., 2006) or morphological inflection of noun phrases (Aldabe et al., 2006). The distractors are therefore simply ungrammatical versions of the correct answer. For word knowledge and comprehension exercises, the selection of suitable distractors is more difficult because the candidate set is larger. Suitable distractors need to be close enough to the correct answer to be a tempting option for the learner, but may not be a proper solution (Lee and Seneff, 2007). In order to increase the difficulty of a multiple choice question, the goal is to maximize this "closeness" without selecting actual solutions. Good distractors are usually from the same word class as the solution and exhibit comparable usage statistics such as frequency and collocation behavior (Hoshino and Nakagawa, 2007). Brown et al. (2005) and Liu et al. (2005) access the standard WordNet relations (e.g. hyponymy, hypernomy, synonymy) to determine close distractors. Heilman et al. (2007), Mitkov et al. (2006), and Sumita et al. (2005) determine the similarity of words using a thesaurus. Agarwal and Mannem (2011) and Moser et al. (2012) focus on contextual similarity of the target word and the distractors in a corpus. Similarly, Zesch and Melamud (2014) select distractors that are near-synonyms in one context, e.g. purchase and acquire, but cannot be substituted in another context (e.g. children can only acquire skills and not \*purchase skills). For this approach, a subsequent reliability check based on context-sensitive lexical inference rules that filters distractors which are valid solutions is particularly important.

Pino and Eskenazi (2009) develop one of the first approaches to *distractor generation* that explicitly targets language learners. They generate distractors that are similar with respect to morphology, orthography, or phonetic representation (e.g. *shared* and *shredded*) and also consider the mother tongue of the learner. Yin et al. (2012) and Sakaguchi et al. (2013) generate exercises based on a corpus of manually annotated learner errors in order to specifically target learner problems.

Skory and Eskenazi (2010) claim that the *quality of exercises* is directly related to the *difficulty*. They find that the difficulty of open cloze exercises correlates with the readability of the sentence. The variety of approaches to distractor generation for closed cloze exercises show that the choice of distractors determines the educational goal and has an influence on the difficulty and the quality of the exercise. Pino et al. (2008) evaluate their generated exercises with human judges and find that 67% of the generated questions are directly usable. For grammatical exercises, the quality is even better. Chen et al. (2006) report that 80% of their exercises are "worthy" and Perez-Beltrachini et al. (2012) find that even more than 90% of the generated exercises are correct. The evaluation guidelines vary across the approaches and a direct comparison is not possible. In order to support the usage of automatic exercise generation approaches, several researchers implemented authoring

tools to include a human quality check for the generated questions (Liu et al., 2005; Mitkov et al., 2006; Aldabe et al., 2006). Based on their evaluations, a common source of errors for inappropriate exercises can easily be identified: exercises are either too easy (for example due to implausible distractors) or too difficult to solve (due to unresolvable ambiguity).

Conclusions We have seen that automatic exercise generation has been tackled for almost all exercise types. Text-completion exercises have been particularly attractive for several reasons. They are usually designed as closed or half-open exercises that are easy to score automatically. They integrate comprehension exercises with linguistic exercises and can therefore be used for a fast estimate of language proficiency. Another important aspect of text-completion exercises is the ability to manipulate the exercise content and the exercise format independently. This makes it possible to adjust the exercises to different learning goals. For example, the range of required comprehension skills could be varied by using different test types and linguistic skills could be determined by the gap placement and the choice of answer options (e.g. semantic, syntactic, or morphological distractors). Most computational research has been restricted to cloze exercises. From an educational perspective, other text-completion exercise types have been found to be advantageous over cloze exercises (see section 2.3).

The quality of generated exercises is strongly related to their difficulty. However, generation approaches cannot yet explicitly manipulate or measure the difficulty due to lack of objective measures for this task. This thesis aims at modeling the difficulty of exercises for language learners to provide an objective basis for improving and manipulating exercises automatically. We have seen that the choice of distractors can have an influence on the exercise difficulty and further explore this aspect in chapter 6.

#### 2.2.3 Exercise Evaluation

Selecting suitable material and generating useful exercises are important tasks to foster learner progress. In order to measure the progress and the learning effect of different exercises, the performance of the learner needs to be evaluated. Natural language processing techniques can be applied to support three very related tasks: analyzing learner language, detecting and correcting errors and automated scoring.

**Analyzing learner language** Exploiting learner corpora to analyze and process learner language has recently experienced a huge rise. The *Learner Corpus Bibliography* lists more than 1,100 bibliographical references, although it is limited to English publications and only dates back until 1990.<sup>15</sup> Most learner corpora contain written or spoken productions from a group of learners within a specific task setting. Good examples for big learner corpora

<sup>&</sup>lt;sup>15</sup>http://www.uclouvain.be/en-cecl-lcbiblio.html, accessed: December 15, 2015

are the *EF-Cambridge Open Language Database* (Geertzen et al., 2012) and the *MERLIN* corpus (Boyd et al., 2014), which are introduced in more detail in section 5.3. Learner corpora provide empirical evidence for second language acquisition hypotheses and can provide insights into the practical effects of the different paradigms in foreign language teaching. Natural language processing tools can be very helpful in automatizing the analysis of learner corpora. Automatic linguistic pre-processing enriches the input with layers of linguistic annotation (e.g. tagging, morphological analysis, parsing) to prepare it for further analysis (see Meurers (2015) for a good overview). The main challenge for linguistic pre-processing is the flawed input produced by the learners. Most tools are trained on standardized and correct input and are not robust enough to deal with the many errors in learner language (Ott and Ziai, 2010). More robust approaches are less sensitive to errors in the text, but come with the disadvantage that the subtleties of learner language might get lost in the analysis. Depending on the target application, researchers thus have to find a trade-off between increasing robustness and maintaining a sufficient level of detail for useful analysis of learner language.

Psycholinguistic research analyzes learner language with respect to acquisition theories, and pedagogical research focuses on the effect of teaching conditions. In contrast, approaches towards automatic analyses usually focus on fulfilling a very concrete task. A typical example is the task of natural language identification. Natural language identification aims at identifying the mother tongue (L1) of a learner based on a text written in a foreign language (L2). In 2013, a dataset of English essays written by native speakers of Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugua, or Turkish was compiled for a shared task on natural language identification which attracted the participation of 29 teams (Tetreault et al., 2013). The quality of the classification systems was high; 13 teams reached an accuracy of more than 80%. Malmasi et al. (2015) show that the automatic approaches perform highly above the human upper bound for this task. The majority of systems relied on n-gram features calculated over words, POS-tags and characters. In addition, several teams included syntactic and spelling features. It can be concluded, that the L1 has a measurable effect on the writing style of a learner in a foreign language. This highlights the importance of considering transfer effects in language learning. This aspect will be discussed in more detail in the chapters 4 and 5.

Error detection and correction It is important that learners receive feedback on their performance in exercises to support continuous learning progress. The direct *detection* and *correction* of errors can prevent the development of language fossilization and structural problems (Ellis, 1994). For half-open and closed exercise formats, error detection and correction only involves a comparison of the learner answer with the expected solution. Due to the direct comparison, structural errors can also be detected and reported if necessary (e.g. if the learner always provides adjectives instead of adverbs, additional training of ad-

verb inflection can be recommended). Reliable error correction of open exercises is more challenging.

While the automatic correction of spelling errors has become an indispensable tool for most writers, the correction of grammatical and stylistic errors is still challenging. In the last years, several shared tasks aimed at grammatical error detection and correction. In the first Helping Our Own (HOO) challenge in 2011, the focus was on correcting scientific papers by non-native speakers. As this task turned out to be too vaguely defined, the performance of the submitted systems was rather low (Dale and Kilgarriff, 2011). The main problem of this task was rooted in the dependencies between multiple sources of error that were not properly reflected in the evaluation. This problem is related to the task of determining dependencies between exercise items that is discussed in section 6.3. In the following year, the organizers thus turned to the correction of isolated errors. The submitted systems should detect preposition and determiner errors in learner essays from the FCE corpus (Yannakoudakis et al., 2011). This corpus is introduced in more detail in section 5.3. The results were slightly better than in the previous task, but even the best systems detected and corrected less than half of the errors in the test set properly (Dale et al., 2012). The task of error correction was continued as a shared task at the CoNLL conference in 2013 and 2014 and extended to a wider range of error types (Ng et al., 2013, 2014). As the task became more complex, the reported performance results for error detection and correction did not improve much over the years.<sup>16</sup> The main strategy for error correction are hybrid systems that combine rule-based approaches for specific error types with language models that compare the probability of a learner phrase with the probability of a correction candidate. Language models for candidate evaluation are also used in section 6.2 in a slightly different setting. Another popular approach is the usage of machine translation systems that "translate" learner English into correct English. In 5.2, we use machine translation technology for the task of *cognate production*. This shows that methods from natural language processing that have been developed for one task, can also prove useful for other tasks with a parallel problem structure.

The main problem with the analysis of learner language is that sentences often contain several errors which cannot easily be disentangled. Lüdeling (2008) claims that errors in learner language can only be corrected with respect to a specific *target hypothesis* and that this target hypothesis is hard to identify even for experienced linguists.

The multi-faceted research interest in the shared tasks shows that there has been considerable research development in the area of grammatical error correction. However, the quality is not yet satisfactory enough to apply automatic correction systems in self-directed learning environments. Native speakers can tolerate quality deficits and still benefit from a tool. In contrast, language learners rely on impeccable feedback because they do not have the expertise to re-evaluate the system output.

 $<sup>\</sup>overline{^{16}}$ The best-performing system at the CoNLL 2014 shared task reached an  $F_{0.5}$  score of 37.33 (Ng et al., 2014).

**Automatic scoring** As the automatic correction of individual errors is very challenging, researchers have focused on providing more general quality feedback for the tasks of proficiency classification, essay grading, and short answer scoring.

Proficiency classification aims at identifying the proficiency level of a language learner based on a small sample. The ability to quickly judge the current status of the learner helps to assign learners to homogeneous groups and to adapt the difficulty of the learning material appropriately. Proficiency classification can be performed based on a lexical decision task that requires the learner to distinguish words from artificial non-words (Lemhöfer and Broersma, 2012) and based on text-completion tests (see section 2.3). Both tests are closed formats which can easily be evaluated automatically by comparing the learner answer with the reference answer. Another typical exercise for proficiency classification are argumentative essays. In this open format, learners are asked to write a free text answer to a given prompt. Crossley et al. (2012) use lexical features to perform proficiency classification of learner essays. Their model decides whether an essay was written by a beginning, intermediate, or an advanced learner and classifies 23 out of 33 essays in the test set correctly. Hancke and Meurers (2013) work with a larger essay set by learners of German and also a wider range of linguistic features and reach an accuracy of 62.7%. It should be noted that they classify the essays on the more fine-grained CEFR scale with five categories. For this task, the relative linguistic quality of the learner essay is compared with essays from a known proficiency class without judging the absolute quality or the topic of the essay.

A strongly related and more popular task is direct *essay grading*. For this task, the linguistic quality of the essay has to be evaluated with respect to the topic suitability of the essay for a given *prompt*. An example for a prompt can be seen below:<sup>17</sup>

The well-being of a society is enhanced when many of its people question authority. Write a response in which you discuss the extent to which you agree or disagree with the statement and explain your reasoning for the position you take. In developing and supporting your position, you should consider ways in which the statement might or might not hold true and explain how these considerations shape your position.

In related work, prompts are also called tasks, issues or topics. We limit ourselves to the term prompt to avoid confusion. Dikli (2006) provides an overview of systems for automatic essay grading. The systems determine the grade for an essay by comparing it to manually scored essays based on a wide range of linguistic features. The most popular system is the *e-rater*<sup>®</sup> system that was developed by the *Educational Testing Service* (Attali and Burstein, 2006). It is used in many testing scenarios to grade the essays of native speakers and language learners. However, it is not used as a stand-alone scorer but the output is combined with human evaluations. In a competition on automated essay grading that was organized

 $<sup>^{17}</sup> Taken \ from: \ https://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/pool, accessed: January 19, 2016.$ 

by the *Hewlett Packard Foundation* (the *Automated Student Assessment Prize*), the best systems were able to consistently replicate human scores (Shermis, 2014). The drawback of these systems is that they require manually labeled essays as training data for each prompt. In order to be able to grade essays for new prompts without hand-labeled data, Zesch et al. (2015) attempt to grade essays without relying on prompt-specific features . They find that the prompt-agnostic models perform better on unseen prompts than prompt-specific models, but the performance losses compared to the performance on known prompts are still extremely high.

More recently, the focus shifted towards *short answer scoring*. This task is even more challenging because the student answer contains less textual data and the grading system needs to evaluate factual knowledge and relationships more precisely (e.g. word order, negations, and agent/patient relations) to compare the answer to a reference answer. Burrows et al. (2014) provide a detailed overview for this task. The results for short answer scoring vary strongly depending on the dataset.<sup>18</sup> Horbach et al. (2013) and Meurers et al. (2011) present results for scoring answers to reading comprehension written by language learners and report accuracy values of up to 84%. They argue that their systems can facilitate the work of human scorers, but should not yet be used as stand-alone evaluation.

We have seen that the field of processing learner language and evaluating learner performance has advanced tremendously in the last years. Automatic natural language identification outperforms human experts by far and automatic essay grading systems have already been introduced into high-stakes testing procedures. They are very useful for large-scale assessments, but cannot easily be adapted to new tasks yet as they rely on high-quality training data. For evaluation on more fine-grained levels such as short answer scoring and grammatical error correction, the automatic systems are steadily improving, but cannot yet compete with human raters. Closed or half-open exercise formats are therefore the safest option if fast and reliable scoring is required.

### 2.2.4 Conclusions

We have seen that natural language processing techniques can contribute to a wide range of application scenarios in language learning. The automation and dynamic adaptation of educational tasks is a first step towards formative computer-adaptive testing and individualized learning support (see condition 2 and 3 for high-quality exercises in section 2.1.2). Based on the overview of the current state of the art, we decide to focus on half-open and closed exercise formats because they can be generated and evaluated automatically without reference answers. We aim at exercises that allow independent manipulation of exercise content and exercise format to facilitate difficulty manipulation.

<sup>&</sup>lt;sup>18</sup>The interested reader is referred to the results of the second phase of the *ASAP* challenge for short answer grading (https://www.kaggle.com/c/asap-sas/leaderboard) and to the *SemEval* shared task on student response analysis (Dzikovska et al., 2013).

In section 2.2.2, we have seen that cloze exercises are a popular format for automatic approaches because they combine technical practicality with educational quality expectations for exercises. In the cloze paradigm, linguistic exercises are embedded into authentic comprehension tasks and can target a variety of skills depending on the choice of the context and the distractors. In the last years, many variants of cloze exercises have been developed. All of these exercises ask the learner to complete the missing elements of a text. For a detailed and fine-grained analysis of language skills, text-completion exercises are usually not sufficient. They need to be coupled with open formats and exercises that address other skills such as speaking and listening. However, for a fast estimate of language proficiency, the use of text-completion exercises has become a commonly used procedure in practice. The range of text-completion exercises is introduced in more detail in the following section.

# 2.3 Text-Completion Exercises

In this section, we analyze *text-completion exercises* in more detail. We first introduce the *reduced redundancy principle* which provides the theoretical basis for all text-completion exercises. We then describe several exercise types that have been developed based on this principle. Text-completion exercises have been analyzed from many angles. We summarize the findings with respect to the properties and the difficulty of text-completion exercises.

## 2.3.1 Reduced Redundancy Principle

In the long history of second language education and language testing, a central question has been the definition of the concept of *language proficiency*. In an attempt to characterize the "notion of knowing a language", Spolsky (1969) uses an analogy from communication theory (Shannon and Weaver, 1949). He describes the concept of *redundancy* as the observation that many elements of a message do not directly contribute to the encoded information value, or, to phrase it the other way around, the information value of the message is not affected if redundant elements are removed. While redundancy is considered as useless in data compression, it is central for communication. Spolsky (1969, p. 8) states:

When one considers all the interference that occurs when natural language is used for communication, it is clear that only a redundant system would work.

In his study, Spolsky (1969) analyzed how the participants could cope with conditions of intentionally created interference, which are more commonly termed as *noise*. He found that the proficiency level of language learners can be estimated by their ability to deal with text-completion. While the performance of native speakers is quite robust in situations of reduced redundancy, language learners have difficulties to restore the information. An illustrative example for a real-life situation of reduced redundancy is a telephone call. In

this scenario, the transmission of additional cues such as gestures and facial expressions is inhibited, making communication a challenging task for language learners.

The observation that the reduced redundancy theory provides a means to measure language proficiency corresponded well with a research theory called the *unitary trait hypothesis*. Oller (1976) claims that second language proficiency should be understood as a single general trait of cognitive processing underlying all of the four skills reading, writing, speaking, and listening. These theoretical developments were soon reflected in the design of corresponding language tests. Klein-Braley (1997) provides a good overview of the operationalization of the theory and compares different reduced redundancy tests. She distinguishes between tests using the acoustic channel (dictation, partial dictation, noise test) and the visual channel (different variants of text-completion exercises). The tests differ mainly in the strategy for introducing noise to the message. For the acoustic tests, additional signals are added to the audio stream. The tests affecting the visual channel are based on texts with partial deletions. The deleted elements (phrases, words, or partial words) are replaced with a gap and the learner is asked to fill in the gaps, i.e. to complete the missing words. We refer to the visual exercises as text-completion exercises.

### 2.3.2 Types of Text-Completion Exercises

The euphoria for the theoretical model of the unitary trait soon cooled down and the focus shifted towards more communication-oriented language learning theories (Bachman, 2000). However, the appreciation of the functional test design of text-completion tests and their reliable and informative results persisted. Today, text-completion exercises are still heavily used, in particular as placement tests and linguistic exercises. In the following, we introduce the three test variants *cloze test*, *C-test*, and *X-test* in more detail.

Cloze test Cloze tests have been originally introduced by Taylor (1953) as a measure for the readability of a text. To produce a cloze test, single words are deleted from a text and the learner is expected to re-produce them. Instead of manually selecting the words to be deleted, a random selection based on a fixed *deletion rate* is recommended. A deletion rate of seven indicates that every seventh word in the text is replaced with a gap. The suggestions for the optimal deletion rate vary in the literature, but it is generally agreed that it is important to provide enough context. Oller (1973) recommends the deletion rate should be higher than five, Brown (1989) even uses a deletion rate of twelve. Taylor (1953) argues that a cloze test from a highly readable text should be easier to complete than a cloze test from a less readable text. While the cloze procedure was never established as a readability measure, the simple and convenient procedure to transform a text to a test was soon discovered for language proficiency testing in the context of the text-completion principle (Oller, 1973; Alderson, 1979).

Unfortunately, cloze gaps are usually highly ambiguous and the set of potential solutions cannot be exactly anticipated (Horsmann and Zesch, 2014). Educational researchers have proposed two ways of dealing with this ambiguity: the application of relaxed scoring schemes and the use of distractors. In *relaxed scoring*, all acceptable candidates for a gap are considered as correct solutions and not only the original word as in exact scoring (Alderson, 1979). Unfortunately, this scoring method turned out to be quite subjective and time-consuming as it is not possible to anticipate all acceptable solutions (Raatz and Klein-Braley, 2002). To circumvent the open solution space, most cloze tests are designed as closed formats by providing a fixed set of candidates from which the solution needs to be picked (Bensoussan and Ramraz (1984), see Figure 1.2 for an example). Closed cloze tests are particularly popular for vocabulary exercises (Skory and Eskenazi, 2010; Dela Rosa and Eskenazi, 2011). The automatic generation of cloze tests and in particular the selection of good distractors has been attempted repeatedly in natural language processing (see section 2.2.2). However, the focus was restricted to creating correct and solvable cloze items; the difficulty of the created items was rarely discussed or evaluated.

C-test The *C* in *C-test* stands for its origin in the cloze test. Although the cloze test is widely used, the setup contains several weaknesses such as the small number of gaps and the ambiguity of the solution. Klein-Braley and Raatz (1984) systematically analyze the shortcomings of the cloze test in detail. They developed the C-test as an alternative and claim that it addresses most of the weaknesses of the cloze test (Klein-Braley and Raatz, 1982, 1984). The C-test construction principle produces a higher number of gaps because every second word of a short paragraph is transformed into a gap. Tests with smaller deletion rates are preferable because they provide more empirical evidence for the students' abilities on less text. However, they also lead to an unfeasible degree of redundancy reduction. To account for this increased difficulty, Klein-Braley and Raatz (1984) propose to delete only the end of the word. The remaining prefix consists of the smaller "half" of the word. If three characters are provided, the correct solution has a length of six or seven characters. A C-test commonly contains 20 gaps and starts with an introductory sentence to provide context as in the following example:<sup>19</sup>

```
The roots of humanity can be traced back to millions of years ago. T__ primary evid__ comes fr__ fossils - skulls, skel__ and bo__ fragments. Scien__ have ma__ tools th__ allow th__ to ext__ subtle infor__ from anc__ bones a__ their enviro__ settings. Mod__ forensic wo__ in t__ field a__ in labora__ can provide a rich understanding of how our ancestors lived.
```

The given prefix and the length constraint restrict the solution space to a single solution (in almost all cases) which enables automatic scoring. However, the C-test is a half-open test

<sup>&</sup>lt;sup>19</sup>The solutions are: The, evidence, from, skeletons, bone, Scientists, made, that, them, extract, information, ancient, and, environmental, Modern, work, the, and, laboratories, now.

that requires productive skills and cannot be solved by guessing (as opposed to the closed cloze test).

Since its introduction, the C-test has been researched from many angles (see Grotjahn et al. (2002) for an overview) and has been tested for many other languages (e.g. German, French, Italian, Portuguese, Hungarian, Hebrew, Turkish). Thorough analyses indicate advantages of the C-test over the cloze test regarding empirical validity, reliability, and correlation with other language tests (Khodadady and Hashemi, 2011; Babaii and Ansary, 2001; Klein-Braley, 1997; Jafarpur, 1995). Reichert et al. (2010) show that C-test performance can be directly linked to the levels of the CEFR which is the European standard guideline for language courses. This property has made the C-test particularly popular for placement tests. Farhady and Jamali (2006) experiment with other deletion rates for the C-test (every 3rd, 4th, 5th, 6th word) and found that these variants do not exhibit the same stable characteristics as the C-test. Tavakoli et al. (2011) examine the effect of genre familiarity for C-tests and cloze tests and find that language learners have an advantage in solving the C-test when they are familiar with the topic. To avoid this topic bias, a C-test session usually consists of five individual C-tests with varying topics.

X-test In order to improve the discriminative power of the C-test, Cleary (1988) introduced a more difficult variant which he called left-hand deletion. The variant is similar to the C-test, but the first "half" of the word is deleted instead of the end of the word. Later, Köberl and Sigott (1994) refer to this variant as the *X-test*.<sup>20</sup> The previous example is transformed into X-test below.

```
The roots of humanity can be traced back to millions of years ago. __e primary __ence comes __om fossils - skulls, __tons and __ne fragments. __tists have __ny tools __at allow __em to __act subtle __ation from __ent bones __d their __mental settings. __ern forensic __rk in __e field __d in __tories can __w provide a rich understanding of how our ancestors lived.
```

In standard C-tests, one of the main challenges consists in selecting the correct inflection of the solution, especially for languages with a rich morphology. In X-tests, the inflected ending of the word is provided and thus the focus is shifted towards lexical-semantic challenges (Scholten-Akoun et al., 2014). Cleary (1988) and Köberl and Sigott (1994) find that the X-test is considerably more difficult than the C-test and discriminates better between the participants. This could be explained by the results of psycholinguistic experiments that have shown that the information value of the initial part of a word is higher than the final part (Broerse and Zwaan, 1966) and that words are easier to recall based on their onset (Kinoshita, 2000). Another explanation could be the ambiguity of the solution. As word

<sup>&</sup>lt;sup>20</sup>In Beinborn et al. (2015a), we used the term *prefix deletion test*, but *X-test* is used throughout the thesis to be more consistent with related work.

endings vary less than word onsets (at least for the languages under study), many X-test gaps allow multiple solutions that are equally valid. In a detailed item analysis of cloze tests, Kobayashi (2002) found that gaps which allow multiple solutions are more difficult to solve than non-ambiguous gaps. He argued that these items require more cognitive ability to evaluate the wide range of possible answers. The X-test is thus a good variant to assess more advanced students up to native speakers.

**Terminological conventions** In this thesis, we do not distinguish between exercises for formative assessment and tests for summative assessment as almost all exercises can be used for both scenarios. The two terms *test* and *exercise* will thus be used interchangeably. The following terms are used to describe text-completion exercises:

- A gap is one item in a text-completion exercise. The first gap in the C-test above is
   T\_\_\_ and will serve as an example here.
- A *solution* for a gap is a word that solves the gap correctly. For some gaps, multiple solutions are valid. The solution for the example gap is *The*.
- An *answer* is a word that has been provided by a participant to fill the gap. For the example gap, most participants provided the solution *The* as answer, but some provided *To*. Answers that are non-words are marked with an asterisk, e.g. \**Teh*.
- The *hint* of a gap is the part of the word that is already provided. In C-tests, it is the first "half" of the word (the *prefix*, T in the example) and in X-tests the second "half" (e for the example gap). Cloze tests usually do not display a hint.
- A *candidate* is a word that fulfills the formal constraints of the gap, i.e. it is part of the vocabulary, it contains the hint and has the correct length. Candidates for the example are *Toe*, *Tip*, *To*, *The*, *Tan*, ...
- A *distractor* is a candidate for the gap that is not a solution. In closed exercises, each gap is presented along with a fixed set of candidates. To solve the gap, the participant needs to distinguish the solution from the distractors. Figure 1.2 displays a cloze gap with five candidates: the solution *observance* and the distractors *instincts*, *presumption*, *expiation*, and *implements*.

### 2.3.3 Properties of Text-Completion Exercises

The strict construction principles for text-completion exercises reduce the influence of the educator and increase the objectivity. The *reliability* and *validity* of text-completion exercises have been subject to many psychological analyses and have shown positive results, particularly for the C-test. While the reliability of the C-test has been widely acknowledged (most recently by Khodadady and Hashemi (2011)), the question "what the test actually measures", also called the *construct validity*, has been more conversely debated. Klein-

Braley (1985, p. 79) argues that the C-test measures general language proficiency, involving all levels of language:

If the learner's competence is fully developed he or she will be able to use all levels of language to restore/reconstruct the text – there will be grammatical, syntactical, lexical, semantic, collocational, contextual, pragmatic, logical, situational clues (and no doubt many others).

Sigott (1995) elaborates that C-test integrate vocabulary knowledge with aspects of word-class specific syntactic competence and sentence-level syntactic competence. Chapelle (1994) and Singleton and Little (1991), on the contrary, consider the C-test mainly as a tool for vocabulary research, but their definition of vocabulary is very wide and also comprises morphological and syntactic aspects. More recently, Eckes and Grotjahn (2006) show that the C-test comprises all dimensions of general language proficiency, namely reading, listening, writing and speaking.

Micro- and macro-level processing In order to find a more common terminology, the strategies a learner applies for solving a C-test have been categorized as micro-level processing and macro-level processing strategies (Babaii and Ansary, 2001). Learners who apply micro-level processing strategies only consider the gap itself and its direct context for solving it, while the full sentence or even the full text is taken into account for macro-level processing. Babaii and Ansary (2001), Babaii and Moghaddam (2006) and Salehi and Sanjareh (2013) examine the test taking strategies for C-tests and cloze tests by conducting so-called think-aloud protocols with the participants. They encouraged the participants to verbalize their mental processes during the actual solving attempt. They found that micro-and macro-level processing strategies are both required for C-test solving to approximately the same extent. Their results indicate a positive relation between test difficulty and processing strategies: more difficult tests trigger more macro-level processing. Babaii and Moghaddam (2006) report the impression, that the participants followed a hierarchical "hypothesis testing" approach for solving. The participants start guessing the solution based on micro-level cues and then refine their guesses by insights from macro-level processing strategies. Grotjahn and Stemmer (2002) further examine this aspect and find evidence that C-test solving primarily involves micro-level processing and that macro-level processing only plays a minor role. Sigott (2006) refines this conclusion by adding the finding that more proficient learners apply both micro-level and macro-level processing strategies when solving a C-test, whereas novice learners only focus on micro-level cues. This is in line with Klein-Braley's claim that proficient learners use knowledge from all levels of language.

**Recognition vs production** In a related debate, it has been discussed whether C-test solving is rather a *recognition* or a *production* task. Cohen (1984) argues that the given prefixes

reduce the extent to which productive skills are required. He therefore considers the C-test to be only a test of reading ability. Harsch and Hartig (2015) show that the C-test results of students correlate highly with their results for listening and reading tests. They conclude that the C-test is a very good predictor for receptive language skills. They did not include any tests of productive language skills in their analyses. Jakschik et al. (2010) transform the C-test into a true recognition test by providing multiple choice options and find that this variant is significantly easier than open C-test gaps. This indicates that C-test solving requires both, receptive and productive skills.

**Item dependencies** In most text-completion exercises, several gaps occur in one sentence. As a consequence, many gaps can only be solved if the context has been restored correctly by solving the preceding gaps. This phenomenon of *item dependencies* is particularly relevant for psychological analyses based on the *item response theory* (see section 3.4.2 for an explanation). The most common models in item response theory assume local independence between items. As this is a counter-intuitive assumption for C-tests, Harsch and Hartig (2010) and Krampen (2014) analyze the occurrences of dependencies between items. Eckes and Baghaei (2015) and Schroeders et al. (2014) aim at determining better models that incorporate local dependence with mixed results. This aspect will be discussed in more detail in chapter 6.

It can be concluded that text-completion exercises fulfill the educational criteria to a satisfactory extent and that they tap both recognition and production skills. The difference between micro-level and macro-level processing strategies and the effect of item dependencies are important aspects for modeling the difficulty of text-completion exercises.

### 2.3.4 Difficulty of Text-Completion Exercises

After the introduction of the text-completion principle, several linguistic researchers analyzed the difficulty of the resulting exercises. Sigott (1995) distinguishes between *content* and *format* aspects of difficulty.

**Exercise content** The content of a text-completion exercise consists of a short text. Taylor (1953) explicitly introduced the cloze test as a measure for readability. If a text is transformed into a cloze test and this cloze test is easy to solve, Taylor claims that the underlying text exhibits high readability. This claim has found many advocates because it provided a means to evaluate text comprehension. It has also been strongly criticized because completing a cloze test requires productive skills that are not necessary for reading

<sup>&</sup>lt;sup>21</sup>The term readability is commonly used in related work and can be interpreted as the inverse of text difficulty. It is discussed in more detail in chapter 4.

comprehension.<sup>22</sup> Recently, the research on readability has made significant progress (see chapter 4) and the cloze procedure is only known as an educational exercise. The relationship between readability and exercise difficulty has been exploited by Klein-Braley (1984) to predict the mean difficulty for English and German C-tests. She uses a linear regression equation with only two difficulty indicators: average sentence length and type-token ratio. She reports good results for various target groups. Dörnyei and Katona (1992) question her findings because they analyze that the relative mean difficulty of tests is not maintained across samples. Skory and Eskenazi (2010) analyze the correlation between cloze easiness with readability scores and with co-occurrence scores. The co-occurrence score is obtained from a corpus and indicates how often the solution co-occurs with the neighboring content words in the sentence. They report very low correlation that is not even significant for the readability scores. Karimi (2011), on the other hand, measure lexical richness of a text as an indicator of readability and find that lower lexical richness leads to lower C-test difficulty for Iranian learners.

Readability is averaged over a full text and thus can only predict the overall text difficulty. However, most text-completion exercises consist of several items that exhibit high variance in the difficulty. Therefore, the focus soon shifted from macro-level to micro-level difficulty.

Brown (1989) analyzes the difficulty of individual items in English cloze tests and identifies the word class and the local word frequency as factors correlating with cloze gap difficulty. Abraham and Chapelle (1992) analyze item properties for three variants of cloze test and find that the effects on difficulty vary for open cloze test and multiple-choice cloze tests. Unfortunately, they examine only a single test, which makes it impossible to generalize their results. Dörnyei and Katona (1992) analyze the performance of Hungarian learners on English C-tests and find that they can solve gaps with function words more easily than gaps with content words. Sigott (1995) examines the word frequency, the word class, and the constituent type of the gap for English C-tests and finds high correlation with the difficulty only for the word frequency. Klein-Braley (1996) identifies additional error patterns in English C-tests. She focuses on production problems (right word stem in wrong form) and the phenomenon of early closure, i.e. the solution works locally, but not in the larger context. Her findings are based on manual data analysis. Similarly, Kobayashi (2002) performs detailed item analysis of cloze tests for Japanese learners of English and finds that the word familiarity, the word class and the morphological inflection contribute to the item difficulty.

<sup>&</sup>lt;sup>22</sup>Christine Klein-Braley has discussed the relationship of readability and text-completion exercises in detail in the chapters 11 and 14 of her habilitation *Language Testing with the C-Test* which she submitted in 1994. Unfortunately, she passed away before her habilitation was published. We managed to get hold of a preliminary version and profited from the strong linguistic base. Fortunately, her main contributions have previously been published in several articles and we cite these instead.

The first automatic approach for *item-level difficulty prediction* was performed for cloze exercises. Hoshino and Nakagawa (2008) focus on grammatical distractors and developed path features that indicate the morphological distance from the distractors to the correct solution according to pre-defined patterns. In addition, they consider basic word frequency and length features. They attempt a binary decision between easy and hard items and ignore items with medium difficulty. Their approach classified only 60% of the questions correctly. They concluded that better word difficulty features are necessary, but did not follow up on this issue. Their dataset consists of questions extracted from preparation books for the *TOEIC*<sup>23</sup> test that are annotated with difficulty values. Unfortunately, the origin of the difficulty values cannot be determined. We therefore do not use the dataset in the thesis. For the sake of completeness, we attempt a comparison to the results by Hoshino and Nakagawa (2008) in section 7.4.4. An approach for C-test difficulty prediction that complements the work in this thesis has been performed in Svetashova (2015) and is described in the sections 3.4 and 7.2.6.

**Exercise format** Most analyses of exercise difficulty only focus on one specific exercise type and the *exercise format* is not considered as an additional variable. However, the exercise format has a strong effect on the difficulty. The difficulty of different types has been compared empirically by Köberl and Sigott (1994), Sigott and Köberl (1996), and Jakschik et al. (2010) indicating that X-tests are more difficult than C-tests and that closed formats are easier than half-open formats.

For text-completion exercises, three design choices can be distinguished: the *answer format*, the *gap type*, and the *deletion rate*. The answer format has already been discussed in section 2.1.2. For closed formats, the answer can be selected from a set of options; for half-open formats it needs to be produced by the learner. Jakschik et al. (2010) compare a standard C-test with a closed C-test (multiple choice with five candidates) and find that the closed version is significantly easier. The absolute difficulty of closed formats strongly depends on the selection of the distractors. If the distractors can easily be discarded, the learner can guess the right solution without even knowing it. Accordingly, distractors that strongly compete with the solution can easily mislead the learner. Distractor selection has already been discussed in more detail in section 2.2.2.

The gap type determines which portion of the word is deleted. The gap type is usually kept constant for all gaps of a test to avoid confusing the participants.<sup>24</sup> Köberl and Sigott (1994) analyze four tests each with a different gap type and examine the influence of the redundancy reduction on the test difficulty for German. In three cases, the end of the word is deleted (word-final deletion). They experiment with deleting half of the word (the stan-

<sup>&</sup>lt;sup>23</sup>http://www.ets.org/toeic, accessed: December 9, 2015

<sup>&</sup>lt;sup>24</sup>This holds at least for all publications about text-completion exercises that are cited in this thesis. A mix of different gap types would probably have a negative effect on face validity.

dard C-test gap), two thirds, and deleting everything except for the first character. In the fourth case, the first half of the word is deleted (the X-test gap). The authors show that for word-final deletion, the difficulty increases with a higher degree of redundancy reduction. Deleting half of the word results in easier exercises than deleting two thirds which in turn results in easier exercises than deleting everything but the first character. Interestingly, they also find that word-initial deletion is more difficult than word-final deletion (deleting the first half of the word is as difficult as deleting everything but the first character). The same tendencies can be observed for English tests, but the effects of the gap type are not as strong as for German (Sigott and Köberl, 1996). These results are consistent with experiments showing that the information value of the initial part of a word is higher than the final part (Broerse and Zwaan, 1966) and that word-initial priming facilitates production (Kinoshita, 2000).

On the global test level, the deletion rate determines the distribution of gaps in the text. The deletion rate has been expressed as a positive integer in previous work.<sup>25</sup> A deletion rate of n signals that every  $n^{th}$  word should be transformed into a gap. Higher deletion rates thus indicate a smaller number of gaps. For cloze tests, Oller (1973) proposes high deletion rates to provide a sufficient amount of context:

Typically every fifth, sixth, or seventh word is deleted either from a written or spoken passage. It has been determined that with native speakers deleting words more frequently than one out of five creates a test of such difficulty that much discriminatory power is lost.

For non-native speakers even higher deletion rates such as deleting every 12<sup>th</sup> word have been used (Brown, 1989). In our data, the cloze test items each consist of one sentence with a single gap. The C-test and the X-test are usually designed with a deletion rate of two, i.e. every second word is transformed into a gap. This lower deletion rate leads to a higher number of gaps per sentence and increases the dependency between gaps because the damaged context of a single gap can only be recreated by solving the surrounding gaps.

An additional aspect for difficulty can be the visual design of the items. Bresnihan and Ray (1992) and Meißner-Stiffel and Raatz (1996) experiment with different *item designs* that indicate the length of the solution with dashes or dots. In C-tests and X-tests, the participants can infer the interval for the length of the solution word by the length of the provided part p as [2 |p|, 2 |p| + 1]. However, many participants seem to forget this *length principle* while solving the test and provide solutions that are either too long or too short (see section 6.1). The works mentioned above show that directly visualizing the length principle facilitates the test for the participants.

<sup>&</sup>lt;sup>25</sup>The word "rate" is thus slightly misleading here.

### 2.3.5 Conclusion

We have seen that numerous manual analyses have been performed focusing on specific aspects of difficulty. In real learning scenarios, individual difficulty effects can usually not be observed independently of other influences. It is therefore important to combine all aspects into an integrated concept of difficulty. The inferences about difficulty have usually been drawn from existing results. The next step consists in predicting the difficulty for unseen exercises.

# 2.4 Chapter Summary

This chapter has described a wide range of text-based exercises that can be applied in different learning scenarios. Based on the analyses in this chapter, we conclude that text-completion exercises are a reasonable choice for fast language proficiency tests. They meet the educational quality criteria and can be generated and manipulated automatically because exercise content and exercise format are separable. Half-open text-completion exercises like C-tests and X-tests provide a trade-off between recognition and production exercises and function as an integrative measure of linguistic knowledge and comprehension abilities. We have seen that approaches to automatic exercise generation have already provided a wide range of methods for different exercise types. Improving the quality and the adaptivity of these generated exercises requires a concept for measuring and adapting the difficulty. Related work has provided analyses of the influence of content and format factors on the exercise difficulty. In the following chapter, we contribute a model which consolidates the different findings into an integrated concept of difficulty.

### Chapter 2. Exercises for Language Learning

#### CHAPTER 3

# The Concept of Difficulty

"Divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution."

- Rene Descartes

This chapter introduces the theoretical model of difficulty that forms the basis for the computational prediction approach. The analysis of related work in the previous chapter and the derivation of the difficulty model allow us to determine a more precise formulation of the thesis goals in the first section. The difficulty of exercises is closely coupled with the performance of learners. An exercise is considered to be more difficult if less learners are able to solve it. For a proper analysis of difficulty, we thus need exercise datasets that contain information about the learner performance. For this thesis, five datasets covering three languages and three different exercise types have been collected and are introduced in detail in the second section.

In order to predict and manipulate exercise difficulty, it is necessary to determine a measure for the concept of difficulty. In the third section, the two most common methods are compared to analyze their applicability to the datasets and our task.

# 3.1 Difficulty Model

Difficulty analyses based on observed educational data are common in psychological studies, but the prediction of difficulty for unseen exercises has not received considerable interest prior to our publications, neither in the field of psychology or linguistics nor in the area of natural language processing. This is probably due to the inherent interdisciplinary nature of the task: it requires complex automatic processing techniques, but also a detailed understanding of the underlying cognitive load and the linguistic phenomena.

The difficulty model for this thesis is based on the analysis of related work in section 2.3 and displayed in figure 1.1 on page 4. Difficulty is grouped into content and format factors following the distinction by Sigott (2006). It is then further categorized into the four dimensions: text difficulty, word difficulty, item dependency, and candidate ambiguity. The model incorporates the distinction of micro-level and macro-level processing by Babaii and Ansary (2001). The outer circle including text difficulty and item dependency represents macro-level processing and is visualized using more saturated colors. The inner circle including word difficulty and candidate ambiguity corresponds to micro-level processing.

During micro-level processing, participants deal with *local difficulty*, and macro-level processing captures *global difficulty*. For the content factors, this means that the difficulty of individual words are local difficulties and the combination of words in a text forms the global difficulty. For the format factors, this means that the candidate ambiguity of a single gap contributes to the local difficulty and the item dependency between several gaps to the global difficulty. The macro-level dimensions are thus not clearly separated from the corresponding micro-level dimension, they should rather be interpreted as an aggregation over the local difficulties that also takes the relations between items into account.

### 3.1.1 Content factors

The content of a text-completion exercise consists of a short text from a general domain. The difficulty of this underlying text has an influence on the difficulty of text-completion exercises because more advanced language proficiency skills are required to restore the text.

Text Difficulty The close relationship between readability of the underlying text and the difficulty of the resulting exercise has already been analyzed by Taylor (1953). In a first approach to difficulty prediction, Klein-Braley (1984) experiments with the two readability variables average sentence length and type-token ratio and finds that they are good predictors of the mean difficulty. However, her findings are not confirmed in studies by Dörnyei and Katona (1992) and Skory and Eskenazi (2010). Recently, natural language approaches towards measuring readability have improved tremendously. Chapter 4 introduces the wide range of readability features and contains experiments to verify whether the predictive power of readability features for mean exercise difficulty claimed by Klein-Braley (1984) can be replicated.

Word Difficulty A text is a conglomerate of words. As each text contains easier and harder words, each exercise also contains easier and harder gaps. The mean difficulty of an exercise is thus not informative enough because averaging over all words conceals local difficulties. The individual item difficulty is strongly linked to the difficulty of the solution word. Brown (1989) and Sigott (1995) find that the syntactic class and the frequency of a word correlate with the item difficulty in text-completion exercises. Klein-Braley (1996)

and Kobayashi (2002) additionally analyze word production problems such as inflection. Chapter 5 introduces a wide range of word difficulty features and contributes new measures for cognateness and spelling difficulty.

#### 3.1.2 Format factors

The comparison of different exercise types in related work has shown that the exercise format has a strong influence on the difficulty (see 2.3.4). On the macro-level, the deletion rate has an influence on the dependencies between items. On the micro-level, the gap type and the answer format determine the size of the candidate space and the candidate ambiguity. These dimensions are examined in more detail in chapter 6.

Item Dependency The causes of item dependencies and their effect on item difficulty have not yet been examined. However, Babaii and Ansary (2001) and Sigott (2006) find that C-test items which require macro-level processing tend to be more difficult than items that can be solved using only micro-level cues. Abraham and Chapelle (1992) make a similar observation for cloze tests. As dependent items induce macro-level processing, we assume that items with strong dependencies on other items are more difficult to solve. Section 6.3 analyzes this phenomenon in more detail and evaluates measures for item dependency.

Candidate Ambiguity Text-completion exercises differ with respect to the number of solution candidates. A higher degree of redundancy reduction leads to an increased number of potential solutions. It is assumed that the exercise difficulty is directly related to the degree of redundancy reduction, but this assumption has not yet been closely examined. We determine the *candidate space* computationally in section 6.1. In addition to the size of the candidate space, the *candidate ambiguity* plays an important role. A learner might fail to solve a gap despite knowing the solution because a competing candidate is stronger. In 42 gaps of the English C-test training data, an alternative answer is provided more frequently than the intended solution. Kobayashi (2002) finds that ambiguous cloze gaps are more difficult to solve although multiple solutions are valid. This indicates that higher candidate ambiguity is cognitively more demanding for the learners. Computational approaches to candidate evaluation are introduced in section 6.2.

### 3.2 Thesis Goals

Almost all previous work on the difficulty of text-completion exercises was limited to correlation analyses of very few manually determined features and focused either on content or on format factors. This thesis aims at contributing computational approaches for a wide range of difficulty aspects and motivates the combination of content and format factors.

The main goal is the application of the resulting model for the difficulty prediction of items in different exercise types and for multiple languages.

The results of the difficulty prediction approach have partially been published in Beinborn et al. (2014b) and Beinborn et al. (2015a). To the best of our knowledge, these publications were the first to report difficulty prediction results for C-tests and X-tests. Svetashova (2015) later conducted an additional range of analyses for English C-tests and also obtained good results for difficulty prediction. She particularly analyzed different concepts of measuring difficulty (see section 3.4 and 7.2.6 for details). Being able to reliably measure and predict the difficulty is the first step towards more focused difficulty manipulation.

To summarize, we set the following goals for the thesis:

- (1) Develop measures that can be computed automatically for each of the four difficulty dimensions.
- (2) Build a prediction approach that is based on the measures for the individual dimensions and is able to predict the difficulty of text-completion exercises for several exercise types and languages.
- (3) Apply the difficulty prediction approach for difficulty manipulation.

Chapter 4, 5, and 6 describe the realization of the first goal. The second goal is addressed in chapter 7 and the third goal in chapter 8.

To tackle these tasks, two pre-conditions need to be met. First, datasets containing text-completion exercises and information about learner performance for each exercise are required to train a prediction model. Second, it is important to quantify how the learner performance can be interpreted as exercise difficulty. These two pre-conditions are addressed in the following two sections.

# 3.3 Difficulty Datasets

In order to predict the difficulty of text-completion exercises and evaluate the quality of the prediction, datasets with learner performance information are required. As suitable data was not freely available in digital form, we started collaborations with several testing institutions and also conducted own data collection studies. Table 3.1 provides an overview of all datasets. As most of the exercises form part of a language proficiency test, we use the terms *test* and *exercise* interchangeably. The number of participants varies for each test because the participants were assigned to random combinations of exercises. We therefore specify the average number of participants per gap. The datasets are described in more detail below.

Format	Test type	Tests	Gaps	Participants (per Gap)		
Half-open	C-test en	20/19	400/375	210		
	C-test fr	30/10	600/200	24		
	C-test de	66/16	1320/320	251		
	X-test de	10/ 4	249/ 99	225		
Closed	Cloze en	100/100	100/100	21		

**Table 3.1:** Overview of exercise datasets. The / indicates the split into training and test data.

#### 3.3.1 Cloze Test

We use an existing set of cloze exercises from the Microsoft sentence completion challenge.<sup>26</sup> For this dataset, Zweig and Burges (2012) transformed 1400 sentences from 5 Sherlock Holmes novels (written by Arthur Conan Doyle) into cloze tests. In each selected sentence, they replace a low-frequency content word with a gap and provide the solution along with four distractors. The distractors were generated automatically based on n-gram frequencies and then handpicked by human judges. It should be noted that all distractors form grammatically correct sentences and that the n-gram probabilities for the answer options are comparable. As the original dataset does not contain any difficulty information, we conducted an annotation study. The human upper bound for native speakers on this task was determined by asking an eloquent native speaker of English to answer a subset of 50 cloze questions resulting in 100% accuracy. In order to determine the difficulty for language learners, 20 web surveys were set up each consisting of ten cloze questions (as displayed in figure 1.2). Advanced learners of English were then asked to answer them on a voluntary basis.<sup>27</sup> The participants had heterogeneous language backgrounds and were asked to self-rate their English proficiency on the CEFR scale. A description of the scale was provided, but the results and comments indicate that not all learners were able to identify their proficiency level correctly on the scale. These classifications are therefore not discussed here, but the data is available at http: //www.ukp.tu-darmstadt.de/data/c-tests/difficulty-prediction-for-language-tests.

#### 3.3.2 C-test

The data for English and French C-tests was collected in co-operation with the language center at *Technische Universität Darmstadt*. The institute uses C-tests as placement tests at the beginning of each semester to determine the best course level for each student. They originally followed a manual procedure: the teachers designed a traditional paper and pen-

<sup>&</sup>lt;sup>26</sup>http://research.microsoft.com/en-us/projects/scc, accessed: December 15, 2014

<sup>&</sup>lt;sup>27</sup>Participants were allowed to answer multiple surveys and were instructed not to use any additional resources.

cil test, the students filled the gaps, and the teachers scored each test by hand. As typically more than a thousand students participate in the test each semester, the manual procedure generated a large amount of repetitive work. To benefit from the advantage that the C-test is a half-open format which can easily be scored automatically, we created a web interface for the testing period.<sup>28</sup>

The web interface The web interface contains an administrative view for the test construction by the teachers and a test view for the students. The test construction is a semi-automatic process. The teachers provide a text which is automatically transformed into a C-test. In a second step, the teachers can edit the test if necessary. In particular, they can encode alternative solutions, e.g. color and colour. In the test view, the students first provide their personal data and then read the test instructions and a C-test example. Once they start the test, they have a time limit of 20 minutes to fill in the gaps. The time limit was chosen to be generous, almost all students finished ahead of time. The consultation of additional references was not allowed. All participants are students enrolled at the university. They have heterogeneous backgrounds regarding their language proficiency and mother tongue, but the majority is German.

Each placement test consists of five *C-test paragraphs*. Each paragraph forms a coherent text with three to five sentences and contains 20 gaps. The texts cover different topics to avoid a domain bias. The teachers intend to arrange the paragraphs with ascending order of difficulty. In the remainder of the thesis, a C-test refers to a single C-test paragraph.

The student answers are stored in a *MySQL* database and evaluated automatically. After the testing period, a report that lists the results for each student is provided for the teachers. The participants can score a total of 100 points, one point for each correctly solved gap. The students are then assigned to course levels according to their results. For English, students who score less than 20 points are considered to be absolute beginners, 20–49 points indicate a level of B1, 50–64 points correspond to B2, 65–85 to C1, and a result of more than 85 points corresponds to a course level of C2.<sup>29</sup> In addition, the teachers can inspect the individual results in detail in the administrative view of the interface to analyze borderline cases (see figure 3.1). The web interface was tested in a pilot phase for English with a subset of the participants.

**English** In a first session in April 2013 (S1), 357 participants filled in the same placement test. In a second session in October (S2), three different placement tests were assigned randomly to 463 new participants. As the pilot phase was very successful and the teachers were

<sup>&</sup>lt;sup>28</sup>The development of the interface was conducted in co-operation with the student assistants Jens Haase and Fabian Rokohl.

<sup>&</sup>lt;sup>29</sup>The thresholds can vary slightly for other languages. Personal communication with Martha Gibson, June 12, 2015.

The roots of humanity can be traced back to millions of years ago. The primary evidence comes
from fossils - skulls, skelletons and bone fragments. Scientists have made tools that allow the
em to extract subtle information from ancient bones and their environmental settings. Mod
ern forensic wo in the field and in labora tory can now provide a rich understanding of
how our ancestors lived.

**Figure 3.1:** A screenshot of the administrative view in the web interface. The teachers can inspect the student answers in detail. Correct answers are highlighted in green and wrong answers in gray.

	<b>S</b> 1	S2	S3
Number of C-tests	5	15	25
Number of Gaps	100	300	375
Participants	357	462	1041
Participants per Gap	357	154	210

**Table 3.2:** English C-test data from three sessions

extremely relieved to be freed from the duty of scoring, the automatic procedure was integrated into the course management system of the institute and rolled out to all languages. In the final setup, the placement tests were composed by randomly choosing C-tests from 5 groups of anticipated difficulty. A random combination of 5 C-tests (one from each group) was then assigned to 1,050 new participants in the third session in April 2014 (S3). Due to the random assignment, the number of participants per gap is averaged over all C-tests for S2 and S3 (see table 3.2). During the pilot phase, we additionally requested background information from the participants such as their L1. Unfortunately, this request could not be integrated into S3 due to data protection policies. The linguistic analysis in this thesis is based on the English C-test data from S1 and S2 unless indicated otherwise. The data from S3 is held-out test data for the experiments in chapter 7. Six C-tests of S3 that overlap with tests of S2 and S1 are removed from the test data.

French The French data was collected by means of the same procedure as the English data, but only in the third session. The participants' profile is comparable: students with heterogeneous background who take the placement test in order to participate in a French language course at the language center of *TU Darmstadt*. The teachers had roughly estimated the difficulty of 40 C-tests to categorize them into five groups of difficulty. Each placement test then consisted of a random combination of five C-tests; one from each group. 30 C-tests are used for training and 10 for testing.

German As the language center at *TU Darmstadt* does not conduct placement tests for German language courses, we obtained additional data from the *TestDaF* institute. This institute awards a certificate of German language proficiency which is required for foreign students who apply for courses at German universities. The test results were collected as

part of a calibration study for improved test design. Students enrolled in German courses at the *TestDaF* institute were asked to participate on a voluntary basis. The data is organized in sets of 10 C-tests with assumed ascending order of difficulty. The data is described in detail in Eckes (2011). For the original study, a total of 27 sets have been answered by 5,927 participants from 125 different countries. Only a subset of the data has been released for usage in this thesis and the information about the participants has been removed for anonymization.

The dataset used in the thesis comprises ten sets from 2,220 participants. Eight sets are used for training and two sets are used as held-out test data. In each set, the third and the eighth C-test are *anchor tests* that are similar across sets which results in a total of 82 different C-tests. The results for the anchor C-tests are calculated over all ten sets and form part of the training data. As a consequence, they are of course excluded from the test set. This partition yields 66 C-tests for training and 16 for testing.

#### 3.3.3 X-test

For the X-test format, we obtained German tests from the center of teachers' education at the *University of Duisburg-Essen*. The tests have been conducted as part of a larger evaluation of the language proficiency of prospective teachers which is described in more detail in Scholten-Akoun et al. (2014). The participants are a mix of native German speakers and students with migratory background (24.07%). Their language proficiency is higher than that of the participants in the other tests.

The data consists of 14 X-tests with 25 gaps each. The 787 participants had to complete four tests (i.e. 100 gaps) within 20 minutes. The combination of tests was varied. We use ten X-tests for training and four for testing.

For both German datasets (C-test and X-test), the answers are only available as already encoded correct/false entries. The original answers have not been stored, so that information about the type of errors cannot be derived.

# 3.4 Measuring Difficulty

Measuring the difficulty of a test is an essential task in the field of psychological and educational testing to assure the scalability of diagnostic results. One test usually consists of several items and most approaches distinguish between test difficulty and item difficulty.

This distinction between text difficulty and item difficulty separates items from the context in which they appear. However, items in text-completion exercises are determined by their context. Not only the difficulty, but also the solution of an item can vary in different contexts. The item  $o_{-}$  could easily be solved with of in the context States  $o_{-}$  America,

while it would require further disambiguation in the context write o\_ paper to distinguish between the answer options on, our, and off.

In this thesis, the distinction between the macro-level test difficulty and the micro-level item difficulty is thus captured in the representation of the item. For the evaluation of item difficulty, we aim at a single intuitive measure based on the learner performance for the respective item. Two main theories can be distinguished to determine the difficulty of items: classical test theory and item response theory (IRT).

#### 3.4.1 Classical Test Theory

Classical test theory has a long tradition for the evaluation of human assessments since the 1960s and its measures are still commonly used. The exercises that are examined in this thesis are prototypical test items that are scored with a binary correct/false decision (encoded as 1/0). The teachers evaluated the tests following classical test theory and obtained an overall *test score* for a participant *i* as the proportion of correct answers for the test items *j* over the number of items *k*.

$$TestScore_{i} = \frac{\sum_{j=1}^{k} answer_{ij}}{k}$$

$$TestScore = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} answer_{ij}}{n * k}$$
(3.1)

$$\overline{TestScore} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} answer_{ij}}{n * k}$$
(3.2)

According to classical test theory, this test score (3.1) can be interpreted as an indicator for the true score for each participant biased by some error (Lord et al., 1968). In our setting, the true score is supposed to be the language proficiency of the participant. <sup>30</sup> The mean test score can then be obtained by taking the average over all participants (equation 3.2).

The test score is aggregated over all items in the test. In order to analyze individual items, classical test theory provides the *P-value* which represents the proportion of correct answers of all participants i (i = 1, ..., n) for a specific item j:

$$P\text{-}value_{j} = \frac{\sum_{i=1}^{n} answer_{ij}}{n}$$
(3.3)

 $<sup>^{30}</sup>$ It is assumed that a true score can never be observed because all measurements contain some sort of error. Analyzing the reliability and validity of the text-completion exercises for the purpose of evaluating language proficiency is out of the scope of this thesis. However, there exists a wide range of psychological publications covering these questions. The interested reader is referred to Grotjahn et al. (2002) who provides a summary and reports generally high values for reliability and validity of the C-test. The C-test has been successfully applied to assign students to language levels in many institutes and the teachers experience a very high matching rate.

The mean test score and the P-value are commonly interpreted as test difficulty and item difficulty (e.g. Klein-Braley (1984, 1996)). However, both measures rather represent easiness because higher scores indicate more correct answers, i.e. lower difficulty. In this thesis, the focus is on the difficulty and not on the performance of individuals. Therefore, the more intuitive *error rate* (the proportion of wrong answers) will be used to quantify the difficulty for items (3.4) and tests (3.5). The error rate can be expressed as the inverse of the easiness measures.

$$ErrorRate_{j} = \frac{falseAnswers_{j}}{allAnswers_{j}} = \frac{n - \sum_{i=1}^{n} answer_{ij}}{n} = 1 - P\text{-}value_{j}$$
(3.4)

$$\overline{ErrorRate} = \frac{falseAnswers}{allAnswers} = \frac{n * k - \sum_{i=1}^{n} \sum_{j=1}^{k} answer_{ij}}{n * k} = 1 - \overline{TestScore}$$
(3.5)

There are two reasons for choosing this measure. First, it is a more intuitive measure for difficulty because higher values indicate higher difficulty. Second, it is more common in the area of natural language processing. Many language technology applications report the word error rate as a measure of performance (e.g. for machine translation or speech recognition). In these scenarios, a higher word error rate indicates that the input is more difficult for the system.

## 3.4.2 Item Response Theory

The theoretical foundations of item response theory (IRT) were already discussed in the 1950s and 1960s (Rasch, 1960; Birnbaum, 1968), but they experienced a rise in popularity in the 1980s. Since then, the use of IRT models has become an important measurement tool, especially in the research on construct validity of tests and adaptive testing. The item response theory comprises a set of probabilistic models that emerged from the idea that the test behavior of a person is an indicator for a so-called latent trait which is often also referred to as the ability of the person (Becker, 2004). The difficulty of an item should therefore be expressed with respect to the ability that is required to solve it. As discussed above, the measured ability in our datasets is supposed to be the language proficiency of the participant.

IRT models describe the individual items of a test by item characteristic functions. In the most basic one-parametric logistic model (*1PL* also known as the *Rasch model*), the probability of solving item j depends only on the difference between the participant's ability  $\theta_i$  and the item difficulty  $\sigma_j$  (Rasch, 1960). Eckes (2011) uses such a Rasch model to compare the difficulty of different C-tests and to build a test pool.<sup>31</sup> The Rasch model assumes that all

<sup>31</sup>http://www.ondaf.de, accessed: November 25, 2015

test items discriminate equally well between the participants. Kamimoto (1993) argues that this is not the case for C-tests. He analyzes the discriminative power of different gaps in English C-tests and finds high variance. In order to overcome this weakness, Eckes (2011) condenses the individual items of a test into a single super-item and only reports the overall difficulty of the test.

In this thesis, the focus is on the difficulty of individual items. Thus, the *two-parametric logistic model (2PL)* Birnbaum (1968) which includes a discrimination parameter  $\beta_j$  for each item is more suitable.<sup>32</sup> The item characteristic function for the *2PL* model is defined as follows:

$$P(answer_{ij} = 1) = \frac{e^{\beta_j(\theta_i - \sigma_j)}}{1 + e^{\beta_j(\theta_i - \sigma_j)}}$$
(3.6)

The presented results are calculated using an implementation of the *2PL* model in the *ltm*-package (version 1.0) for *R* with the default settings (Rizopoulos, 2006). The parameters are estimated by maximizing the approximate marginal log-likelihood. The difficulty of an item is then expressed as the ability that is required for a participant to have a 0.5 probability to solve the item correctly. For a better understanding of this difficulty concept, Figure 3.2

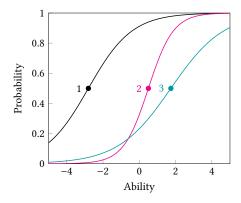


Figure 3.2: Sample item characteristic curves of the 2PL model for three gaps in the English data

shows three item characteristic curves for an easy, a medium and a hard gap as an example. The item difficulty is the point of inflection in the curve, i.e. -2.8, 0.5 and 1.7 for the items 1, 2 and 3 in the figure respectively. It can also be seen, that gap 2 has a higher discrimination parameter (i.e. a steeper slope) than the other two gaps.

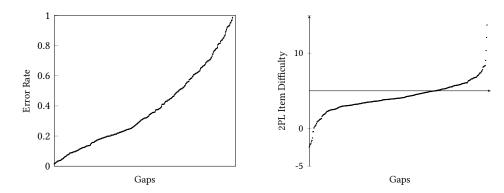
The IRT models are based on a set of very strong assumptions that are hard to meet by many datasets (Hambleton and Jones, 1993). In particular, the assumption of *local independence* between the test items is very likely to be violated in our case; at least for exercises

 $<sup>^{32}</sup>$ In the one-parametric logistic model (Rasch model) the item discrimination parameter  $\beta_j$  is constant for all items. For closed exercises with answer candidates (as in the cloze test), a three-parametric logistic model would be more suitable because it includes an additional "guessing" parameter. For more information on IRT and the different models, the instructional module by Hambleton and Jones (1993) provides a good start.

that contain more than one gap in a single sentence (Schroeders et al., 2014). However, Eckes and Baghaei (2015) show that the effect of local dependence is lower than expected and that a standard *2PL*-Model is sufficient to estimate difficulty. Local dependence will be analyzed in more detail in chapter 6.

### 3.4.3 Comparison of Difficulty Measures

For an informed decision about a suitable difficulty metric, we perform a comparative analysis of the error rate and the *2PL* item difficulty on the training data to examine three aspects: the range of the difficulty, the reliability across samples and the applicability to small sample sizes. In addition, we describe an approach that combines measures of classical test theory and IRT difficulty parameters in so-called performance profiles.



**Figure 3.3:** Sorted difficulty parameters for the gaps in the English training data. Error rates are on the left, *2PL* parameters on the right.

Range of difficulty In our English training data, the item difficulty is almost continuously distributed. In Figure 3.3, the error rates range from 0.01 to 0.99 and the *2PL* item difficulty parameters range from -7.5 to 8.7. In general, the difference between the two difficulty measures is quite small. If the gaps are ranked according to their error rate and their *2PL* item difficulty, the ranks have a Spearman correlation of 0.92. However, the *2PL* model produces two strong outliers with a difficulty parameter of -119.9 and 120.4 that are not displayed in the figure because they are far outside the visible axis range. Interestingly, both outliers have very high error rates (0.93 and 0.96), although they are placed on the opposite ends of the *2PL* difficulty range. We also find items with high error rates that yield "normal" difficulty parameters (around 8), so these two outliers deserve a closer look:

- (1) Early farmers used th\_\_ natural variations selectively breed wild animals, plants and even micro-organisms [...]. (this)
- (2) [...] that the ability of the planet's ecosy\_\_ to sustain and keep up with the demands of future generations cannot be taken for granted. (ecosystems)

	Gaps	1	2	3	4	5	6	7	8	9	10
Mean Error Rate Standard deviation Coefficient of variation		0.83	0.38	0.69	0.27	0.28	0.22	0.80	0.60	0.64	0.07
		0.07	0.06	0.09	0.08	0.11	0.08	0.08	0.09	0.07	0.02
		0.08	0.16	0.13	0.28	0.40	0.34	0.10	0.15	0.11	0.29
2PL Item Difficulty	Mean	2.03	-1.13	0.99	-1.48	-1.25	-1.60	1.21	0.45	0.74	-6.29
	Standard deviation	0.87	0.58	0.50	0.63	0.66	0.47	0.49	0.48	0.52	5.24
	Coefficient of variation	0.43	-0.51	0.51	-0.43	-0.53	-0.30	0.40	1.08	0.71	-0.83

**Table 3.3:** Difficulty statistics for the first ten gaps of a German test averaged over 10 different samples

The first outlier reveals an error in the test design because the sentence is not correct. The preposition to is missing between selectively and breed and either the encoded solution should be these, or variations needs to be in the singular form variation. The participants had of course no chance to solve this item correctly based on their language proficiency, but some (accidentally) provided the intended solution. This leads to a very irregular pattern in the data and as a consequence, a negative value is estimated for the discrimination parameter (i.e. the curve has a negative slope). This indicates, that a participant who is less proficient in English is more likely to solve the item which sounds implausible but might actually be a proper interpretation of a faulty test item. In the second outlier, the high difficulty is due to the ambiguity with the singular form. In this case, the difficulty is estimated to be extremely high because even participants with high ability cannot solve the item.

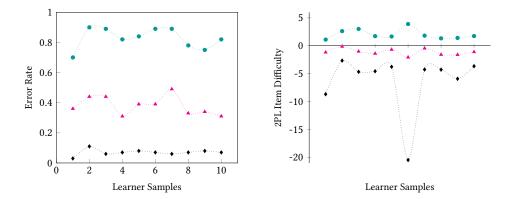
Reliability across samples The IRT models are often advertised with the claim that the person ability and item difficulty parameter estimates are sample-independent (although this is only true for the simple Rasch model) and are therefore more stable than the measures from classical test theory when compared across samples. In our German dataset, two tests have been assigned to ten different samples as anchor tests. This allows us to compare the reliability of our difficulty measures across samples. Table 3.3 provides the aggregated difficulty statistics for the first ten gaps of one of these tests (the results for the second test are comparable). The upper part contains the values for the error rates and the lower part for the item difficulty parameters of the *2PL* model.

It can be seen, that the standard deviations for the *2PL* parameters are considerably higher than for the error rates. This is not easy to interpret because the means are not calculated on the same range. The *coefficient of variation* allows for a better comparison between the two measures because it calculates the ratio of standard deviation and error rate (higher values indicate more variation). It supports the impression that the *2PL* parameters

<sup>&</sup>lt;sup>33</sup>This error had not been noticed during the testing phase.

<sup>&</sup>lt;sup>34</sup>The singular answer *ecosystem* is too short for this gap, but the participants do not seem to take this hint into account.

vary more across samples than the error rates. As these values are averaged over the samples, Figure 3.4 provides a visualization of all samples for three gaps with low, medium and high difficulty. It can be seen, that both measures are quite stable across samples. However,



**Figure 3.4:** Stability of difficulty measures across 10 different samples of learners for Gap 1, 2 and 10. Error rates are on the left, *2PL* parameters on the right.

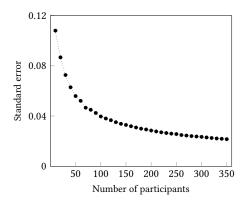
the axis range of the *2PL* parameters is higher and sample 6 produces a very strong outlier for gap 10.

Small samples In the literature, it is usually recommended to use IRT models only for large samples and to rely on classical test theory for small samples, although this is not yet empirically confirmed (Mead and Meade, 2010). Compared to psychological standards, all our datasets are very small. It is thus important to examine whether the number of participants is sufficient to obtain reliable difficulty estimates. Figure 3.5 shows the standard error of the error rates which is calculated for increasing sample sizes.<sup>35</sup> It can be seen that the standard error is reduced to an acceptable level of 0.05 already with 50 participants.

The standard error of the estimated *2PL* item difficulty parameters can also be calculated for increasing sample sizes. Unfortunately, the estimation process often fails for small sets. In an experiment, 20 different runs for each sample size were performed but the results contained so many extreme outliers that the averaged standard errors are too biased to be informative. It can be concluded, that the *2PL* model is not useful for small sample sizes, at least not for the datasets under study.

**Performance profiles** We have seen that the classical test theory is an intuitive measure for assessing the difficulty of an item for a learner group, but the performance of individual learners is only taken into account in the item response theory. Svetashova (2015) shows

<sup>&</sup>lt;sup>35</sup>For each size, the error rate is calculated based on three randomly selected samples of participants from the first English session and the result is averaged. The results for the other three tests are similar.



**Figure 3.5:** Standard error of error rates averaged over all gaps for increasing numbers of participants

that the same exercise item will be difficult for novice learners and easy for advanced learners and proposes to incorporate this performance information in the difficulty concept. She calculates numerous parameters inspired by classical test theory (e.g. percentage of exact matches) and item response theory (e.g. Rasch coefficients) and additionally includes relative measures that represent the difficulty of the item relative to the mean difficulty of the text. All of these measures are calculated with respect to the different CEFR proficiency levels of the learners (A1-C2) resulting in a total of 440 measures. In order to gain a better understanding of the correlations between these measures she applies principal component analysis and clustering. The clustering results show that a reasonable grouping of items by their difficulty can be reached by distinguishing between easy items in easy texts, easy items in difficult texts, difficult items in easy texts and difficult items in difficult texts. Svetashova discusses that this indirectly accounts for the proficiency level of the learner by assuming that difficult texts are generally difficult for novice learners. We have a closer look at this observation in section 4.2.3. In our difficulty model described in section 3.1, the difference between macro-level text difficulty and micro-level item difficulty is captured in the representation of the item and not in the dependent difficulty variable.

Conclusions We have seen that the differences between the classical test theory and the item response theory are rather small for our data; the two difficulty measures are strongly correlated. The analyses show that the parameter estimation process for the *2PL* model is likely to produce extreme values for irregular answer patterns and small sample sizes and in some cases, the estimation process does not even converge. This behavior can provide useful information for item selection but it is not desirable for the task of difficulty prediction which requires stable difficulty estimates. For our datasets, the error rates are very stable across samples and for varying sample sizes. In addition, the concept of error rates is more intuitive and comprehensible for non-experts than the IRT difficulty parameters.

Format	Test type	Mean Error Rate	Participants
	C-test en	$.38 \pm .24$	Students at <i>TU Darmstadt</i> applying for English language courses (≈A1–C1)
Half-	C-test fr	$.52 \pm .26$	Students at <i>TU Darmstadt</i> applying for French language courses (≈A1–C1)
open	C-test de	$.56 \pm .26$	Participants in German language courses as preparation for studying in Germany (≈A2−C2)
	X-test de	$.35 \pm .22$	Native speakers, students at the center of teachers' education at <i>Univ. Duisburg-Essen</i> (C2)
Closed	Cloze en	.28 ± .23	Volunteers, mostly students (≈B1−C2)

**Table 3.4:** Mean error rate and standard deviation for each training dataset

As this can be an important aspect for the acceptance of future applications, the difficulty of an item is expressed by its error rate in the remainder of the thesis. As a consequence, the item difficulty should be interpreted relative to the learner group represented in the training data. For difficulty analyses that focus more on individual differences, item response theory or a combinatorial measure like the performance profiles might be a better choice. Chapter 4 aims at replicating previous work for predicting the mean error rate based on readability features, and in chapter 7 the individual error rate of each gap is predicted based on the new difficulty model. Table 3.4 provides an overview of the mean error rates for the different training datasets. The error rates should not be directly compared across test types because the participants had different backgrounds.

## 3.4.4 Measuring Prediction Quality

Throughout the thesis, we perform several prediction experiments. Classification experiments with multiple classes are evaluated using the conventional evaluation metrics precision(P), recall(R), and  $F_1$ . If only two classes are distinguished and the classes are balanced, the results are evaluated based on the accuracy(A). In the equations below, TP refers to true positives, FP to false positives, TN to true negatives and FN to false negatives. These counts can only be determined with respect to one specific class. To evaluate the overall performance of the classification, the metrics are reported as weighted average over all classes.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{3.7}$$

$$F_1 = \frac{2 * P * R}{P + R}$$
  $A = \frac{TP + TN}{TP + FP + TN + FN}$  (3.8)

Regression experiments are evaluated using the conventional evaluation metrics *Pearson* correlation (*Pearson's r*) and root mean square error (*RMSE*) for linear regression. In the equations 3.9 and 3.10, n refers to the number of gaps and  $\sigma$  to the standard deviation.

$$Pearson's r = \frac{covariance(errorRates, predictions)}{\sigma(errorRates) * \sigma(predictions)}$$
(3.9)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (errorRate_i - prediction_i)^2}{n}}$$
 (3.10)

The *significance* of Pearson's r is calculated on the basis of the *paired.r* function in the *psych*-package (version 1.4.5) for R (Revelle, 2014). In some cases, additional metrics are required. These are introduced in the corresponding sections.

For most experiments, we perform *leave-one-out cross-validation*. There exist several methods to evaluate cross-validation results, but many of them can yield misleading results. We follow the recommendation by Forman and Scholz (2010) and calculate the results only once over the collected predictions from each fold.

# 3.5 Chapter Summary

This chapter has introduced the difficulty model that forms the theoretical basis for this thesis. Exercise difficulty is interpreted as a combination of the four dimensions text difficulty, word difficulty, candidate ambiguity, and item dependency. For the analytical basis of the thesis, five difficulty datasets have been introduced. Two datasets have been created in cooperation with the language centre at *TU Darmstadt*. Two other datasets have been obtained from educational researchers at the *TestDaF* institute and the *University of Duisburg-Essen*. The last dataset is a set of existing cloze questions that have been enriched with error rates from a learner study conducted for this thesis. Based on these datasets, the main approaches for measuring difficulty could be compared. The results of this comparison led to the decision for the error rate as measure of difficulty.

The analysis of related work in the previous chapter and the development of the difficulty model enabled us to provide a more precise determination of the thesis goals. In the following chapters, each difficulty dimension is analyzed in detail and operationalized into objective, measurable features that can be determined automatically. In chapter 7, all individual measures are combined into an integrative model to tackle the task of gap difficulty prediction. An elaborate prediction of difficulty facilitates target-oriented difficulty manipulation. Existing strategies for text selection, text manipulation and distractor selection mainly rely on implicit difficulty assumptions by experts that are not directly measurable and often do not account for language learners. In chapter 8, we analyze how the integrated

#### Chapter 3. The Concept of Difficulty

difficulty prediction approach can contribute to improved text selection. The measures for word difficulty and candidate ambiguity can guide manipulation strategies for lexical simplification and distractor manipulation respectively.

#### CHAPTER 4

# **Text Difficulty Prediction**

"Reading is to the mind what exercise is to the body."

Joseph Addison

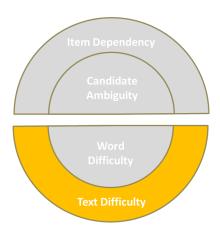


Figure 4.1: Text difficulty: macro-level dimension of the exercise content

For all text-based exercises, the difficulty of the underlying text determines the available context for the participant. A more challenging text increases the difficulty of the exercise as the participant's orientation in the text becomes more complicated (compare Brown (1989)). Research on text difficulty is commonly approximated under its inverse concept of readability as in the following quote by DuBay (2004, p. 3):

Readability is what makes some texts easier to read than others.

A wide range of approaches have been proposed to measure readability based on the linguistic complexity of the text. Collins-Thompson (2014, p. 98) highlights the importance of focusing not only on the text, but also on the reader:

In addition to text characteristics, a text's readability is also a function of the readers themselves: their educational and social background, interests and expertise, and motivation to learn, as well as other factors, can play a critical role in how readable a text is for an individual or population.

Nevertheless, he treats the terms *text difficulty* and *readability* as synonyms to avoid confusion. We follow this convention in this chapter. The first section provides an overview of existing readability measures and features from all linguistic levels. In previous work, a strong relation between readability and the difficulty of text-completion exercises has been claimed (Klein-Braley, 1984). The second section contains experiments for analyzing the reproducibility of these results for the difficulty datasets used in this thesis. As most related work has focused on readability of texts in the readers' L1, we discuss the challenges for L2 readability in the third section and analyze the influence of two important learner variables: the L1 and the reading experience. Most arguments in the first and the third section have already been discussed in Beinborn et al. (2012) and Beinborn et al. (2014b).

# 4.1 Readability Measures

Readability measures have a long history, especially in the American education research. The need for these measures is rooted in a very practical task: teachers search for texts that best fit the knowledge level of their students. According to Vygotsky's *zone of proximal development* (Vygotsky, 1978) the range of suitable texts that a learner can manage without help is very small. Texts which do not challenge the student easily lead to boredom, whereas texts with overly complex language might lead to frustration when no tutoring is available.

From a diachronous view, readability measures have continuously taken more and more features into account. Early measures in the 1960s worked only with surface-based features and manually determined constant coefficients. Later approaches successively added features from the lexical, syntactic, semantic, and discourse dimensions as the respective technologies became available. As the number of features was steadily growing, the need for machine learning methods emerged. Supervised learning methods use training data from corpora to determine the significant features for each readability level. The most common readability corpora are based on the *WeeklyReader corpus* or the *Common Core State Standard*. The learned feature weights for the training data can then be applied for the prediction of readability levels of unseen texts. Recently, progress in the field of text classification has led to a new perspective on readability measures (Collins-Thompson (2014) provides a good overview). Supervised learning algorithms automatically combine several

<sup>&</sup>lt;sup>36</sup>The traditional measures expressed readability in terms of the twelve US school grades. More recent approaches only distinguish two to four different readability levels.

<sup>37</sup>http://www.weeklyreader.com/

text properties extracted from training data and learn to associate them with the corresponding readability class. Naïve Bayes (Collins-Thompson and Callan, 2005), k-nearest neighbors (Heilman et al., 2007), and support vector machines (Schwarm and Ostendorf, 2005) have all been tested in previous approaches. This automation has led to an explosion of the text features that are taken into account. The features can be grouped into various dimensions focusing on different aspects of the text. In the following, we first introduce the traditional measures. Then we discuss the lexical, morphological, syntactic, semantic, discourse, conceptual, and pragmatic aspects of readability and give an overview of the features that operationalize them. The features range from simple POS-tag counts over syntactic structures to complex language models and discourse relations. Detailed analyses and results of the individual features can be found in Vajjala Balakrishna (2015), Feng and Huenerfauth (2010), Pitler and Nenkova (2008), and Graesser and McNamara (2004).

### 4.1.1 Traditional Readability Measures

Traditional approaches to readability such as the Flesh–Kincaid Grade Level (Kincaid et al., 1975), the Automatic Readability Index (Smith and Senter, 1967) and the Coleman–Liau Index (Coleman and Liau, 1975) date back even until the 1960s and operate only on two main features, word length and sentence length. In addition to the length features, the SMOG grade (McLaughlin, 1969) and the Gunning-Fog Index (Gunning, 1969) also consider the number of complex words defined as words with three or more syllables. DuBay (2004) gives a comprehensive overview of traditional readability measures and Benjamin (2011) evaluates their usability for teachers. These measures have been incorporated into guidelines for writing standards and are also implemented in standard word processors. However, they have also been subject to criticism as they only capture structural characteristics of the text and can easily be misleading. Consider the following two example texts:

- (1) There were ten green bottles hanging on the wall, ten green bottles hanging on the wall. And if one green bottle should accidentally fall, there'd be nine green bottles a-hanging on the wall.
- (2) This above all: to thine own self be true, And it must follow, as the night the day, Thou canst not then be false to any man. Farewell, my blessing season this in thee!

Text 1 is a popular children's song and text 2 is a famous quote from Shakespeare's play *Hamlet*. Both texts consist of two sentences and contain 34 words. Most readers would probably agree that the first text is significantly easier to read. However, the traditional

Readability measure	Text 1	Text 2
Flesh-Kincaid grade level	6.3	5.6
Automatic readability index	8.0	4.8
Coleman-Liau index	10.3	6.4
SMOG index	4.4	4.4
Gunning-Fog score	8.0	8.0

**Table 4.1:** Readability scores for text 1 and text 2. The measures return the US grade level that is typically required to comprehend the text. Higher scores thus express lower readability of the text.

readability measures (see table 4.1) actually indicate the opposite. They return lower readability for text 1 (expressed as a higher grade level) because it contains longer words.<sup>38</sup>

In order to capture the perceived readability disparity in the two example texts above, more advanced linguistic features that operate on the actual content are required.

#### 4.1.2 Lexical-Semantic Features

In order to understand a text, a reader first needs to map words to concepts. In a second step, the word meanings are combined to construct the meaning of complex phrases and sentences. The difficulty of these two steps can be modeled by features capturing lexical and compositional semantics respectively. On a more global level, lexical diversity measures capture the range of vocabulary in a text.

Lexical semantics Many psycholinguistic studies have shown that *frequency* effects play an important role in language comprehension (Brysbaert and New, 2009). High-frequency words enable faster lexical access and should therefore contribute to higher readability. For English, the word length used in the traditional measures is in principle a good approximation of word frequency because frequently used words tend to be rather short compared to more specific terms (Sigurd et al., 2004). However, we can find many exceptions to this approximation.<sup>39</sup> As an alternative, Dale and Chall (1948) propose the use of word lists that are based on the frequency of words. If many words of a text do not occur in the list, they assume higher text complexity. For readability estimates, the number of high and low frequency words is then aggregated over the whole text. In more recent work, manually constructed word lists have been replaced with automatically trained *language models* (see below).

<sup>&</sup>lt;sup>38</sup>It should be noted that the minimum text size recommended for readability measures is 100 words. If we add the subsequent two verses of each text to reach this number, the tendency of the readability measures remains the same. We omitted these verses here for the sake of brevity.

<sup>&</sup>lt;sup>39</sup>Compare, for example, *together* (length: 8, frequency in the American National Corpus: 4,004) and *sag* (length: 3, ANC frequency: 27). Frequency information is available at <a href="http://www.anc.org/data/anc-second-release/frequency-data">http://www.anc.org/data/anc-second-release/frequency-data</a>, accessed: June 19, 2015

Compositional semantics Compositional semantics has been addressed by several approaches, but they have rarely been applied to readability. An exception is the work on readability of German texts by Vor der Brück et al. (2008) which is based on the semantic *Wocadi*-parser (Hartrumpf, 2003). In this approach, the amount of conceptual nodes and relations in the semantic representation is considered as indicator of semantic complexity. Their findings indicate that these features correlate well with human judgments of readability. Unfortunately, the parser often fails to build a semantic representation which limits the robustness of their approach.

Language models Language models are a more robust way to combine lexical and compositional semantics. Instead of absolute frequencies as in word lists, language model approaches are based on word probabilities. The use of language models is a common technique in speech recognition and machine translation in order to determine the probability of a term in a given context. Collins-Thompson and Callan (2005) have shown that this notion of the probability of a term can easily be transferred to readability, since it is generally understood that a sentence is more readable if it uses very common terms and term sequences. Higher n-gram models as used by Schwarm and Ostendorf (2005) can even account for collocation frequencies indicating different usages of content words (e.g. hit the ball vs. hit rock bottom). Language models can easily be re-trained for new domains and new languages and could thus be dynamically adapted to specific learner groups. In most recent works in natural language processing, written word frequencies are extracted from the Web1T corpus based on web texts (Brants and Franz, 2006) and spoken word frequencies from the SUBTLEX corpus based on movie subtitles (Brysbaert and New, 2009).

Lexical diversity The *lexical diversity* determines the vocabulary range of a text. If the same concept is expressed with different words, the reader has to recognize the similarity relation of the words in order to understand the shared reference. Lexical diversity is usually measured by the *type-token ratio*, where type is a word and token refers to the different usages of the word in the text. Graesser and McNamara (2004) have implemented the type-token ratio for their readability framework *Coh-Metrix* and Vajjala and Meurers (2012) experiment with different variants of it. A low ratio indicates that words are frequently repeated in the text. This characteristic might decrease the stylistic elegance of the text, but it facilitates text comprehension. Recall the two example texts in the previous section. The type-token ratio of the rather repetitive children's song is 0.5, whereas the more complex Shakespeare excerpt has a type-token ratio of 0.9. The type-token ratio can thus be a strong indicator for readability which has an influence on exercise difficulty. Klein-Braley and Raatz (1984) and Karimi (2011) show that lower lexical diversity leads to lower difficulty of C-tests.

### 4.1.3 Syntactic Features

Syntactic features measure the grammatical difficulty of a text. Especially for language learners, complex syntactic structures are major text comprehension obstacles. The surface-based measures estimate the syntactic difficulty by considering the average sentence length. A longer sentence might indeed indicate a more complex structure, however, it could also simply contain an enumeration of concepts. In recent approaches, the grammatical structure is represented by part-of-speech (POS) patterns and parse trees, as described below.

POS-tagging In readability measures, POS-tagging is mainly used for the distinction between *content* and *function words*. Content words carry lexical meaning, whereas function words indicate syntactic relations (e.g. articles or conjunctions). A high number of content words indicates high lexical density (Vajjala and Meurers, 2012). Feng and Huenerfauth (2010) additionally determine the absolute and relative occurrences of the different POS-tags in the sentence and find that high usage of nouns and prepositions is an indicator for increased text complexity. Heilman et al. (2007) highlight the occurrence of different verb tenses as indicators for text complexity, especially for second language learners. Grammatical constructions are usually acquired step by step and complex structures such as the use of the passive voice occur in later stages. Infrequent verb tenses might thus strongly inhibit a learner's comprehension of the text. Lu (2012) approximates the syntactic complexity by calculating the *verb variation*. The verb variation is similar to the more general *type-token ratio*, but only includes verbs.

Parsing Another method to assess syntactic complexity are parsing features. Syntactic parsers analyze the grammatical structure of a sentence and return a formal syntax representation. For readability measures, the number and type of noun and verb phrases are considered relevant (Schwarm and Ostendorf, 2005; Heilman et al., 2007). In addition, Schwarm and Ostendorf (2005) include the depth of the parse tree and the number of subordinated sentences in order to model the sentence complexity. Similarly, Vajjala and Meurers (2012) consider the number of clauses per sentence and the number of subordinations and coordinations. According to Tonelli et al. (2012), a text is easier to read if it exhibits low syntactic variability, i.e. if the syntactic structure of neighboring sentences exhibits high similarity. This can be computed by detecting the largest common parse subtree of two sentences. Babaii and Moghaddam (2006) examine the effect of syntactic complexity on C-test difficulty, but the results are not very conclusive.

### 4.1.4 Discourse Features

Discourse features model the structure of the text and particularly analyze *cohesion* and *coherence*. In the readability literature, all inter-sentential relations are qualified as related to discourse.

Cohesion A text is considered to be *cohesive* if the individual sentences are reasonably tied together on the linguistic level. This can be obtained by using cohesive markers such as connectives. Pitler and Nenkova (2008) build a discourse language model based on the annotations from the *Penn Discourse Bank*. This model determines how likely it is for each grade level that the text contains implicit or explicit discourse relations. Tonelli et al. (2012) manually create a list of additive, causal, logical, and temporal connectives for Italian. In addition, they calculate the ratio between causal or intentional particles and causal or intentional verbs. Causal and intentional verbs are identified manually by exploiting category and gloss information from *WordNet*. Flor et al. (2013) measure the degree to which a text tends to use words that are highly inter-associated in the language as an indicator for lexical cohesion. They introduce the measure lexical tightness that is based on pointwise mutual information of words and show that it correlates with the grade level for a text.

Coherence A text is considered to be *coherent* if the sentences are reasonably linked on the semantic level. In the readability literature, this is often approximated by the use of co-references. Graesser and McNamara (2004) analyze co-references in more detail and determine the relation between two consecutive sentences. They interpret entities that occur in both sentences (noun overlap) as indicators for coherence. Pitler and Nenkova (2008) generate entity grids that capture how the center of attention shifts from one entity in the text to another as postulated in the centering theory (Grosz et al., 1995). Feng and Huenerfauth (2010) keep track of the number of entity mentions and the pronoun density. In addition, they assume that a higher number of active entities poses a higher working memory load on the reader. In order to determine the active entities, they identify *lexical chains*. A lexical chain is formed by entities that are linked through semantic relations such as synonymy or hyponymy. The length and the sentence span of the chain are interpreted as indicators for text complexity.

### 4.1.5 Readability for Other Languages

Several researchers also implemented readability calculations for other languages, e.g. for French (François and Fairon, 2012), German (Hancke et al., 2012; Vor der Brück et al., 2008), Portuguese (Aluisio et al., 2010), Italian (Dell'Orletta et al., 2014; Tonelli et al., 2012), Swedish (Pilán et al., 2014; Larsson, 2006), Japanese (Sato et al., 2008), and Arabic (Al-Khalifa and Al-Ajlan, 2010). In order to do that, it is not enough to simply adapt the features established

for English. The particular characteristics of a given language also need to be considered in the feature selection. Morphological aspects, for example, are less important for English, but crucial for agglutinative languages. Some of the features described above rely on the use of lexical resources such as *WordNet* and thus limit the applicability to other languages. In this thesis, we focus on readability measures that are available for English, German, and French, as these are the languages of our datasets. An overview of the implemented readability features can be found in the appendix in section A.

# 4.2 Readability and Reduced Redundancy Exercises

The relationship between readability and text-completion exercises has already been discussed in chapter 3.1. In chapter 3.4, we have seen that the difficulty of a text-completion exercise can be measured by the mean error rate which is averaged over all gaps and all participants. As the exercises in our datasets are used as placement tests for comparing students, the difficulty of different tests should be balanced. Readability measures can thus serve as a predictor to determine tests with comparable difficulty.

### 4.2.1 Correlation of Readability with Difficulty

The first work on the difficulty of C-tests was a correlation analysis of readability features by Klein-Braley (1984). She correlated the mean difficulty for English C-tests with the readability of the text (Klein-Braley, 1984). She focuses on two features: type-token ratio and average sentence length in words (abbreviated as *TTR* and *XSL*) and tests them on C-tests from two groups: third graders and fifth graders. She reports a Pearson correlation between difficulty and readability of 0.76 for third graders and of 0.69 for fifth graders. It should be noted that this correlation was calculated for the full set of each dataset and not tested on unseen data. In the following experiment, the procedure is repeated for our English, French and German C-test training datasets. In addition, the results are compared to those obtained using the wider range of implemented readability features (see section A). In order to be able to compare with previous work, we also use a linear regression approach here.

Table 4.2 shows the results. In the upper part of the table, only the type-token ratio and the average sentence length are used as features. It can be seen, that the correlation of these features with the mean test difficulty of our datasets is considerably lower than the correlation reported by Klein-Braley (1984). The best results are obtained for the French dataset. The column *Included Features* contains the features that have been included into the regression equation. For French, the average sentence length is not even considered for the regression equation. For English, the linear regression algorithm falls back to the

<sup>&</sup>lt;sup>40</sup>She measured difficulty as the P-value but as it is simply the inverse of the error rate this difference can be neglected.

Features	Dataset	Included Features	Pearson's r
	C-test en	Constant prediction: .37	.00
	C-test de	TTR, XSL	.39
TTR, XSL	C-test fr	TTR	.47
	Klein-Braley (1984) (3rd grade)	TTR, XSL	.76
	Klein-Braley (1984) (5th grade)	TTR, XSL	.69
	C-test en	AvgWordLengthInSyllables	.55
17 (see appendix A)	C-test de	FeaturesDe	.55
- <b>-</b>	C-test fr	FeaturesFr	.59

**Table 4.2:** Correlation of readability features with mean difficulty on the training set. *FeaturesDe*: AvgWordLengthInCharacters, ChunksPerSentence, NounChunksPerSentence, TTR, VerbsPerSentence *FeaturesFr*: AdverbsPerSentence, AvgSentenceLength, AvgWordLengthInSyllables, ChunksPerSentence, NounsPerSentence

baseline of using the average difficulty because both features do not correlate sufficiently with the difficulty. In the lower part of the table, the set of available readability features is extended. This leads to a strong improvement, but the results reported by Klein-Braley (1984) still remain far out of reach.

It should be noted that only few features from the available set of 17 features are included in the regression equation (one feature for English and five features for French and German). The average word length is relevant for all three languages and it is the only relevant feature for English. The number of chunks per sentence is predictive for French and German. It approximates the average sentence length but provides more information about the complexity of the sentence.

## 4.2.2 Prediction of Mean Difficulty

As noted above, the correlations are calculated on the full training set. This setup has the disadvantage that it does not provide insights into the predictive power of the features. In order to overcome this weakness, we perform leave-one-out cross-validation on the dataset. This means that we train a linear regression model on all C-tests but one, and predict the mean difficulty of the excluded test with the trained model. We repeat this procedure for each test in the dataset. The Pearson correlation is then calculated over all individual predictions. The set of features that are considered by the classifier can be different for each fold because the regression equation depends on the training set.

The results in the upper part of table 4.3 show that the readability features by Klein-Braley cannot reliably predict the mean difficulty. Only for the French dataset, the mean difficulty is predicted relatively well. For English, we even obtain a negative correlation. We have seen that the best prediction for the training set was a constant difficulty of 0.37.

Features	Dataset	Pearson's r
TTR, XSL	C-test en C-test de C-test fr	66 .26 .39
17 (see appendix A)	C-test en C-test de C-test fr	.04 .39 14

**Table 4.3:** Pearson correlation for mean difficulty prediction based on readability features using leave-one-out cross-validation on the training set. The feature combination that is used by the classifier can be slightly different for each fold because it depends on the training set.

In this case, the difficulty prediction is not very informative because it would be constant for any new test.

Extending the feature set to the full set of readability features leads to improvements for German. For French, the prediction quality gets worse leading to a negative correlation value and for English, the correlation between actual and predicted difficulty is close to zero.

In general, these results show that predicting the mean test difficulty based on standard readability features is not very promising. The results in Klein-Braley (1984) have never been verified on unseen data, so they might have been overly optimistic. Another explanation could be that the mean difficulty is simply not a very useful measure for text-completion exercises. This aspect is discussed in more detail in the following section.

# 4.2.3 The Concept of Mean Difficulty

The mean error rate, which has been used as dependent variable in the readability experiments above, is averaged over all gaps in a text. Figure 4.2 visualizes the measure for the 20 texts in the English training dataset. The texts are grouped according to their order of presentation to the students. One test session consisted of five texts and the teachers intended to arrange them with ascending difficulty.

The results show that the difficulty estimate of the teachers was not accurate. The first test of a sequence is indeed easier than the last one, but a true ordering has not been achieved for any test session. This underlines the need for objective difficulty estimates to assure proper test conditions. However, the high standard deviations indicate that the mean error rate is not a very informative measure because each test contains very easy and very difficult gaps. Imagine an extreme case, in which half of the gaps can be solved by all learners and the other half by almost no one. The test is then assigned a medium difficulty, but the results are not useful for discrimination between learners.

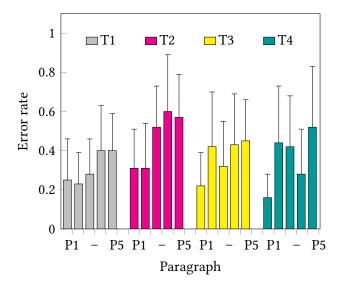


Figure 4.2: Mean error rate and standard deviation for the paragraphs 1-5 of the four tests

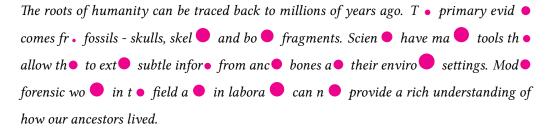


Figure 4.3: Visualisation of error rates for each gap

In the training data, the error rates of the individual gaps range from 0.01 to 0.99 and are almost continuously distributed. Figure 4.3 shows an example for the high variance of the gap difficulty within a single paragraph. The error rates in the example are indicated by the size of the circles: bigger circles visualize higher difficulty. In order to account for the high variance within tests, it is reasonable to focus on the difficulty of individual gaps.

Readability measures can only contribute as a constant factor to the difficulty of all gaps. As the difficulty of each individual gap deviates strongly from the mean difficulty, traditional readability scores do not provide sufficient information. This indicates that the concept of an average readability of a text that has been established for native speakers might not be suitable for foreign language learners. The following section provides an analysis of the reading processes for language learners and the consequences on L2 readability measures.

# 4.3 L2 Readability

Almost all of the approaches to readability discussed in section 4.1 are directed at native speakers reading in their mother tongue (L1). A naïve approach to readability for language learners could simply apply the existing L1 models on texts in the foreign language (L2, L3,...) to calculate the readability. However, the analyses on text-completion exercises above indicate that an averaged mean is not informative enough for the reading processes of language learners. Greenfield (2004) argues that the traditional formulae already give a reasonable approximation for L2 readability and that their ease of application can make up for other shortcomings. On the contrary, Carrell (1987) and Brown (1998) raise the concern that the traditional work on readability cannot capture the different requirements for L2 readability. They particularly mention the need to consider the reader's background knowledge. François and Fairon (2012) stress the importance of multi-word expressions in L2 readability, but their experiments do not validate this assumption. Heilman et al. (2007) and Vajjala and Meurers (2012) include measures of syntactic complexity to assess readability for language learners. However, they find that lexical features outperform the syntactic features in their experiments.

In order to properly analyze the requirements for L2 readability, we first need to understand how L2 learning differs from L1 acquisition. We then discuss the resulting differences of L1 and L2 readability in more detail.

## 4.3.1 L1 Acquisition and L2 Learning

The acquisition of the native language and the process of learning a second language progress in different ways. The process of language learning is a strongly debated topic in psycholinguistics. Many terminological distinctions have been attempted and discussed without reaching common ground (see DeKeyser (2009) for an overview). In the following, we focus on *L2 learning* that starts after the L1 has already been acquired or as Jiang (2000) defines it: "second language learning with insufficient natural exposure and/or with established L1 linguistic and semantic systems". We do not account for L2 acquisition by bilingual children but focus on more conscious processes as in standard L2 learning settings. In order to emphasize the aspect of consciousness in second language learning (Ellis, 1994), we will distinguish between *L1 acquisition* and *L2 learning*. In this section, we elaborate on several differences with a focus on the lexical comprehension of texts.

**L2 input** The differences between L1 acquisition and L2 learning already emerge from the general setting; the basic L1 knowledge is learned from the unstructured input children receive from their environment. Therefore, their language knowledge is based on experience with direct day-to-day communication and occurs in very concrete situations. The L2, on the other hand, is usually learned gradually by instruction (Cook et al., 1979) following

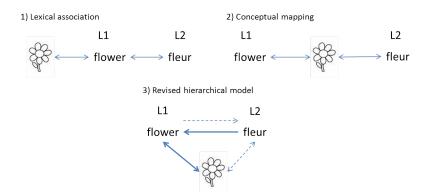
a more conscious process that also requires more structured input (Schmidt, 1995). The poverty of the stimulus is a term coined by Chomsky (1965) initially pointing to the quality of the language input that children receive from their environment during L1 acquisition. It describes the fact that humans often use incomplete or even ungrammatical sentences in spoken language. In language learning settings, it is generally assured that learners are presented with a more controlled use of language and also receive direct error feedback through corrections. However, Jiang (2000) argues that L2 learners also suffer from the "poverty of the input". In his argument, Jiang does not refer to the quality of the input but to the frequency. Native speakers are confronted with the language they are about to learn all the time and cannot verbally communicate using another language. L2 learners usually only receive input in the foreign language during a few hours per week (Nation, 2003) and are less dependent on it.

It can be concluded that L2 learners have to learn from less frequent, but more controlled language input than native speakers. Thus, the learning process needs to follow more conscious patterns in order to extract more information from less input. For this aim, L2 learners have access to a wider range of learning strategies than children that acquire their L1 (Meara, 1988) and can thus modulate and adapt their learning process. Approaches to L2 readability should consider these more conscious strategies of dealing with language input and account for linguistic inferences.

L2 mapping processes The native language is usually acquired in the first years of child-hood, whereas an L2 is generally learned after the L1. This means that a certain level of proficiency in the L1 already exists. Psycholinguistic researchers are especially interested in the differences of language representation in the L1 and the L2. Many experiments with bilinguals are conducted in order to shed light on the question how the different languages are managed in the brain.

Two general mapping models for L2 acquisition can be distinguished: *lexical association* and *conceptual mapping* (see figure 4.4 model 1 and 2). Lexical association is based on the form of words; learners are assumed to map the L2 form on an existing lexical entry in the L1. The model of conceptual mapping, on the contrary, is based on the idea that the L1 and the L2 word both independently point to the same conceptual representation. Thus, in the first model, learners have to map L2 words on L1 words, whereas the second model assumes that learners map L2 words on concepts already associated with L1 words. Kroll and Stewart (1994) propose the *revised hierarchical model* that combines the two ideas. They argue that lexical associations are predominant in early stages of L2 learning and that with increasing L2 competence more direct conceptual links are established (see Figure 4.4 (3)). Kroll and Stewart (1994, p. 168) note that this revised hierarchical model can, for example, explain the finding that translation from L2 to L1 is usually faster than translation from L1 to L2:

### Chapter 4. Text Difficulty Prediction



**Figure 4.4:** Language representation in the minds of language learners. Lexical association (1) and conceptual mapping (2) were both introduced by Potter et al. (1984). Kroll and Stewart (1994) combine the two mapping processes into the revised hierarchical model (3) and argue that the different connections have varying strength (the bold arrows signal strong links, dashed arrows stand for weaker links).

[...] translation from the first language to the second is conceptually mediated, whereas translation from the second language to the first is lexically mediated.

When reading a text in a foreign language, the learner works on L2 words and figures out their meaning in L1. As this is the lexically mediated direction, lexical aspects are particularly important for L2 readability. In a refined version of the revised hierarchical model, Jiang (2000) makes the temporal aspect of lexical development more concrete. He defines three successive stages ranging from pure lexical association at the beginning to conceptual mapping at later stages. Lexical association is thus particularly dominant for less proficient learners. For intermediate to advanced learners both mapping routes – lexical and conceptual – are activated and are used depending on the given cues. Both mapping processes should thus be considered for L2 readability.

Cross-lingual influence The mapping models introduced above indicate that L2 learning is influenced by the background knowledge of the learner. As the L1 is already present, basic concepts of languages such as the different behavior of word classes or the syntactic coordination of arguments are already known. In addition, the specific properties of the L1 influence the acquisition of the L2. These phenomena of *cross-lingual transfer* have been thoroughly researched on several dimensions. Odlin (1989) discusses cross-lingual effects on grammar, vocabulary and pronunciation acquisition. Jiang (2000) focuses on the negative effects of L1 interference that can be deduced from word choice and usage errors during production.

Ringbom and Jarvis (2009) distinguish between *system transfer* and *item transfer*. In system transfer, principles for organizing the elements of the L1 are transferred to the L2. This applies particularly to the morphological behavior of words and to the syntactic com-

bination of words into phrases. Syntactic constructions in the L2 that are parallel to constructions in the mother tongue are more likely to be used by learners (Zobl, 1980). On the other hand, non-familiar constructions are more likely to cause problems. Kobayashi (2002) analyzed that Japanese learners struggle particularly with the usage of determiners in English cloze tests because the concept of determiners does not exist in Japanese. These transfer patterns are exploited in automatic approaches for L1 identification. For this task, the structure of texts by language learners is analyzed to predict the L1 of the learner. The best systems reach an accuracy of 85% and perform considerably better than human experts (Malmasi et al., 2015). The transfer effect is also reflected in the observation that an L3 (or L4, L5, ...) is usually acquired more easily than an L2 (Cenoz, 2003) because multilingual learners can already rely on more linguistic knowledge.

Item transfer refers to the direct mapping of L2 and L1 items. It is mainly a phenomenon in the lexical dimension and particularly important for language comprehension. Ringbom and Jarvis (2009, p. 11) note:

The cross-linguistic similarities that underlie item transfer are a concretely perceived similarity of form usually combined with an assumed similarity of function or meaning. [...] item learning has a predominantly positive effect on learning, notably on learning for comprehension.

Cross-lingual item transfer is facilitated for language pairs that contain many cognates (Ellis, 1994). Cognates are words that share a related form with their translation in another language, e.g., elegance and elegancia in Spanish (see section 5.2 for a more detailed explanation). De Groot and Keijzer (2000) show that cognates are acquired more easily and also persist longer in memory independent of their frequency. They argue that the form overlap of the two words is a strong cue for their successful association.

Cross-lingual item transfer is particularly relevant for recognition tasks where the lexical form is overtly available so that lexical association is strongly facilitated. Reading comprehension is a typical example for a recognition task. Other recognition tasks comprise, for example, the lexical decision whether a word belongs to the language, and L2 to L1 translation. However, the cognate facilitation effect persists even across different alphabets, e.g. for Japanese learners of English (Hoshino and Kroll, 2008), and can also be found in the productive picture naming task (Kroll and Stewart, 1994; Lotto and de Groot, 1998). It is assumed that not only the lexical, but also the phonetic similarity of cognates contribute to these facilitation effects. The numerous findings on cross-lingual influence in language learning – and in particular on the cognate facilitation effect – highlight the importance of taking the native language of the learner and also other previously acquired languages into account for the preparation of suitable reading material.

**Reading experience** As L2 learners are older than children acquiring the L1, they also tend to have a more advanced educational background and have already developed higher

intellectual abilities (Cook et al., 1979). L2 learners often start reading in the foreign language very early in the learning process whereas native speakers already have completed a period of several years limited to verbal communication before they face texts. Thus, L1 readers still have to learn standard text processing, principles of cohesion and coherence and general linguistic competence, whereas L2 learners can already build on these skills from their L1 experience. Meara (1988) argues that most studies on lexical development in children do not manage to disentangle their findings on vocabulary acquisition from cognitive development. For older L2 learners, these processes can be more clearly separated. L2 readability can thus abstract from factors concerned with matching texts to children's intellectual level.

In spite of the intellectual advantages of L2 learners, the vocabulary growth progresses significantly slower than in the L1 (Webb and Chang, 2012). The acquisition of the standard syntactic structures in a foreign language can be successfully completed after a certain period, whereas the vocabulary acquisition is a continuous process. Lervåg and Aukrust (2010) show that the vocabulary size is a critical predictor for reading comprehension by L2 learners. These findings indicate that texts for L2 learners can be conceptually more difficult than for younger L1 readers, but more attention should be given to lexical aspects.

### 4.3.2 Consequences for L2 Readability

From the discussion above, we conclude that L1 readability approaches cannot be directly transferred to L2 readability because L1 acquisition and L2 learning are two different processes. L1 readers already master the basics of the language before they attempt to read texts. L2 readers, on the other hand, often fail to build a representation of a text because they focus on unknown words or constructions.

L1 readability is usually determined for full texts with a minimum length of 100 words. For the readability calculation, many text features are extracted and usually normalized by the length of the text. The normalization conceals the local difficulties and aggregates them into a mean readability measure. For L1 readability, this is a reasonable approach, as readers can consolidate the information in order to facilitate the comprehension of a complex sentence and to dissolve misconceptions and ambiguities. For L2 learners, local difficulties can inhibit the comprehension of the full text. If the mapping processes described above fail to map unknown words on known words or concepts, the learner cannot build a representation of the text content because lexical knowledge is crucial for text comprehension (Laufer and Ravenhorst-Kalovski, 2010). Consider, for example, the following sentence:

He swam up past the boulders made of spongy gishy-gosh and flew right by the herd of floating feeding fipple-fosh.<sup>41</sup>

<sup>&</sup>lt;sup>41</sup>Retrieved from http://www.magickeys.com/books/noblegnarble/index.html, page 9-10, accessed: June 20, 2013

This sentence is taken from the book "The Journey of the Noble Gnarble" by Daniel Errico which was written for young children. It is thus considered to be adequate for inexperienced L1 readers. However, this sentence could be problematic for L2 readers, because it contains two irregular past tense words (swam and flew). The words boulder, spongy and herd might also be unknown as they are domain-specific. In addition, the compounds gishy-gosh and fipple-fosh are made up by the author. They consist of regular English sound patterns, but do not actually exist. L1 readers can easily accept that these new words denote new concepts specific to the world described in the book. L2 learners, on the contrary, will try to map the unknown words on concepts they already know. As this process is bound to fail, they will likely get confused about the sentence and blame their insufficient L2 proficiency instead of simply accepting the new words. In order to avoid that these mapping failures lead to learner frustration, it is important to account for the local word difficulties. It is thus reasonable to compute the readability on the sentence level (Volodina and Pijetlovic, 2013) or give feedback on even smaller units.

As we explained above, the learner's native language background is very important for determining the difficulty of L2 concepts. Learners with the same mother tongue are more likely to experience the same comprehension obstacles because they follow similar analogies. In order to capture the cross-lingual influence, it is necessary to retrieve L2 texts dedicated to learners with a specific mother tongue. In addition, knowledge from previously acquired languages should also be taken into account. Consider the following example:

**Definition** of **moral**: **Private** <u>conduct</u> **based** on **strict** <u>adherence</u> to a **sanctioned** or **accepted code** or **dogma** of what is right or wrong, <u>particularly</u> as <u>proclaimed</u> in a sacred **book**, or by a **non-secular group** or **sect**. 42

This definition of moral describes a very abstract concept using mainly academic words and is thus hard (or probably impossible) to understand for beginning L1 readers. For L2 learners, it would also be a challenging sentence but they can rely on a bigger set of comprehension strategies here. First of all, they are probably familiar with the concept of moral and already have an approximate definition in mind. This knowledge enables them to map the words more easily to the known concepts. A German native speaker can easily comprehend the words marked as bold because they are cognates and have a similar or even identical spelling in German. In addition, the underlined words have Latin roots and are shared among many Western European languages. The learner might have come across them in a previously learned language or in scientific texts. If all these words are successfully resolved due to their lexical similarity to other languages, only very basic vocabulary remains. If L2 learners combine the cognateness cues with their conceptual understanding, they should be able to comprehend the sentence much better than inexperienced L1 readers.

<sup>&</sup>lt;sup>42</sup>Retrieved from http://www.businessdictionary.com/definition/moral.html,19.6.2015

From the discussion above, we can conclude that readability estimates for language learning scenarios need to focus more on the individual words than on the aggregated mean text complexity. Ideally, the background knowledge of the learners (in particular their L1) should also be taken into account.

# 4.4 Chapter Summary

This chapter has first introduced the concept of readability and provided a detailed overview of existing readability features for multiple linguistic levels. In the second section, the relationship between readability and mean difficulty has been evaluated in the context of text-completion exercises. The results show that the claims in previous work about the predictive power of readability features for the difficulty of text-completion exercises are not supported by our data. In the subsequent analysis, we took a closer look at the high standard deviations of the error rate. We concluded that the lack of correlation between readability and difficulty is probably rooted in the phenomenon that the mean error rate conceals local difficulties and is therefore not informative enough. This aspect was further elaborated in the third section. In the traditional view on readability, the text difficulty is an aggregated mean over the elements in the text. This view fits with the holistic comprehension and interpretation approaches by native speakers but does not necessarily match the processes of L2 readability. In the third section, the differences between L1 and L2 readability have been discussed from a psycholinguistic point of view. For native speakers, vocabulary acquisition is already progressed very far when they first attempt reading. The main focus of L1 readability is thus on matters of text organization to account for the learners' lack of reading experience and on the age suitability of the text. In contrast, L2 learners can rely on reading experience from their L1, but struggle with lexical difficulties. We conclude that L2 learners rather attempt to understand a text step-by-step and consciously use comprehension strategies such as cross-lingual item transfer and contextualized inferences. They profit from their existing background knowledge, especially regarding the mother tongue and previously learned languages. As a consequence, L2 readability should focus more on local word difficulty instead of aggregated means. In previous work on readability, some aspects of word difficulty have also been considered (e.g. in language model approaches), but they have been averaged over the whole text. In the following chapter, we aim for a more explicit operationalization of word difficulty.

### CHAPTER 5

# Word Difficulty Prediction

"Words are, of course, the most powerful drug used by mankind."

- Rudyard Kipling

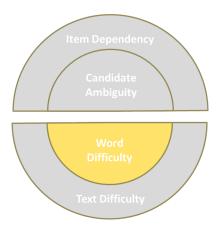


Figure 5.1: Word difficulty: micro-level dimension of the exercise content

For language learners, lexical knowledge is considered to be the major factor for text comprehension (Laufer and Ravenhorst-Kalovski, 2010). Lervåg and Aukrust (2010) show that the vocabulary knowledge can predict the growth of reading comprehension skills for learners and argue that additional vocabulary instruction should be given a high priority in L2 learning. The vocabulary growth in L2 learning proceeds significantly slower than in L1 acquisition and might even reach a dead end, called *lexical fossilization* (Jiang, 2000).

In the previous chapter, text difficulty was introduced as an aggregated mean over the elements in a text. For L2 learners, it is important to put the focus on the difficulty of individual words. The process of learning the basic syntactic structures can be considered

to be more or less completed at a certain point, but vocabulary acquisition is a continuous process that remains important even for advanced learners.

However, assessing whether a learner knows a word is not trivial. In the following quote, Brown et al. (2005, p. 819) summarize the wide range of word knowledge that is described in a detailed analysis on teaching and learning vocabulary by Nation (1990):

Word knowledge is not all or one. Rather, there are different aspects, such as knowledge of the spoken form, the written form, grammatical behavior, collocation behavior, word frequency, stylistic register constraints, conceptual meaning and the associations a word has with other related words.

If there are many factors involved in knowing a word, *not* knowing a word can have many reasons. The training data of our text-completion exercises contains a surprisingly high variety of wrong answers. For the English C-tests, the participants provide on average 19 different answers per gap, ranging from close variants of the solution (e.g. typos) to completely different words. This supports the assumption that knowing a word is not a binary feature, but a rather diffuse conglomerate of many aspects.

In the following, we discuss the aspects of word difficulty in more detail and focus on their operationalization in computational features. We first discuss classical word features that have already been used quite often and illustrate them with examples from the training data. We then focus on two phenomena specific to language learning, namely cognateness and spelling difficulty. As previous work addresses these phenomena only partially, we contribute own approaches to measure them. The new approaches are evaluated on additional corpora to analyze their predictive power.

### 5.1 Classical Word Features

This section provides a detailed analysis of factors that contribute to word difficulty and discusses implementations for computational approaches.

## 5.1.1 Word Familiarity

In natural language processing, it is very common to approximate the probability that a reader knows a word by the word frequency. More frequent words are considered to be more likely known. If we compare the solutions of the easiest (1) and the most difficult gap (2) in our training data, it is obvious that *you* is easier because it is more frequent than *plentiful* and therefore more likely to be known.

- (1) If y are looking for new experiences, [...] (you)
- (2) [...], people may try self-employment because the opportunities seem plen\_\_\_ and financing is easy to get. (plentiful)

This effect has also been identified by Kobayashi (2002) in his study on cloze tests. In chapter 4, we have seen that word frequency was first approximated by the word length and in more recent approaches by word counts in large corpora. The direct use of word counts has the advantage that it can be adapted to different domains and can thus also predict expert vocabulary. However, the relevance of the frequency effect has mainly been studied for L1 acquisition and might be less relevant for L2 learning.

As elaborated in section 4.3.1, the L1 is acquired by direct communication and imitation. More frequent words are heard and used more often and are thus more familiar to the learner. For L2 learners, it also holds that the probability of previously having encountered a word is increased if the word is more frequent, but this is not the only factor. L2 learning is a more conscious process that is usually moderated by the learning material. De Groot and Keijzer (2000) even argue that the success of learning completely new words is independent of the frequency of the concept. In their experiments, they teach artificially created words to learners and find that the frequency of the source word has no effect on the learning success nor on the retention of the word. However, the use of artificial words and the strict experimental conditions do not allow a generalization of their findings. The participants had no chance to actually encounter or apply the new words outside the laboratory, which of course rules out frequency effects that are rooted in word usage.

We conclude that frequency can be a strong indicator for word familiarity, but that other factors also need to be taken into account especially for language learners with insufficient exposure to the target language.

Polysemy and Concreteness Two other aspects of word familiarity are *polysemy* and *concreteness*. Polysemous words complicate the interpretation of a sentence because they have to be disambiguated first. For L2 learners, a polysemous word can be particularly confusing as they might know one sense of the word, but not be aware of the fact that it can also denote something else. In addition, it has been found that L2 learners cannot successfully apply guessing strategies for the meaning of an unknown word if the word is polysemous (Verspoor and Lowie, 2003). Polysemy of a word can also override frequency information as the different senses are not equally frequent. For example, *well* is a highly frequent word, but the occurrence as a noun in the sense of fountain is relatively rare. High polysemy of words is an indicator of low readability for both L1 and L2 readers, but the effect on L2 readers is even stronger.

In texts for beginning readers, we typically find a high number of concrete words because they can more easily be mapped to specific concepts. Concrete words also activate a richer network of semantic information (e.g. mental imagery or wider contextual information) than abstract words (Kaushanskaya and Rechtzigel, 2012). If the word mapping models described in section 4.3.1 apply, word concreteness should be less important for L2 learners because they are already familiar with the source concepts (recall the example

with the definition of *moral* in section 4.3). The L2 word then only needs to be associated with the L1 word. However, experiments by De Groot and Keijzer (2000) do not support this assumption. They show that concrete words are learned faster and retained longer than abstract words. However, if the abstract words are cognates, the learning success is as great as for concrete non-cognates. A possible explanation for this finding could be that the two mapping routes influence each other. If the lexical association is facilitated, the conceptual influence of the word concreteness is reduced.

In the *CohMetrix* readability framework (Graesser and McNamara, 2004), polysemy and concreteness are determined on the basis of *WordNet* relations (Fellbaum, 1998). Polysemy is measured by the number of senses of a word (in *WordNet* senses are provided as synsets). For concreteness, they assume that words which correspond to higher nodes in the ontology hierarchy are less concrete. They thus count the number of hypernym relations that lead from the current word to the root of the hierarchy as indicator for concreteness.<sup>43</sup> Aspects of concreteness have mainly been studied in the area of metaphor interpretation (Turney et al., 2011). For English, human concreteness ratings can be obtained from the *MRC* database (Gilhooly and Logie, 1980) and from a more recent collection by Brysbaert et al. (2014).

### 5.1.2 Morphology and Compounds

English has a rather shallow morphology compared to many other languages. However, many errors in the training data are caused by wrong inflection or derivation. In both examples below, the lemma forms of the words (*death*, *environment*) are provided much more often than the correct solutions.

- (3) There has been a reduction in dea\_ of children. (deaths)
- (4) [...] information from ancient bones and their enviro\_\_\_ settings. (environmental)

Compound words that consist of several parts e.g. *greenhouse* are also considered to be more complex than simple root words because the learner needs to combine the meaning from the parts. This is particularly relevant for German because it contains a high number of compounds.

Word splits in text-completion exercises In most text-completion exercises, a part of the solution word is already provided. This is a helpful hint as it reduces the number of answer options (see chapter 6), but it can also be misleading if the given part splits the word in an unfortunate way, as in the following example.

(5) It is not easy to design and build a mac\_\_ that is both, efficient and durable. (machine)

<sup>&</sup>lt;sup>43</sup>We assume that words that are too concrete, i.e. extremely low in the sense hierarchy, should also be more difficult (as *macaque* compared to *monkey*), but we have not yet found any studies supporting this hypothesis.

In this case, the prefix splits the word in the middle of a syllable which results in a phonetic change and provokes answers with the pronunciation /mæc/ such as \*macanics, \*macanism, \*macanical, macbook, macphone, and macro instead of the original pronunciation /mæʃ/. Hyphenation dictionaries or rule-based hyphenation approaches can provide information about syllable boundaries to account for this phenomenon. If the prefix splits a word at a compound boundary (such as greenhouse), the learner does not receive information about the second stem of the word which might also lead to increased difficulty. This can be checked using decompounding algorithms (Erbs et al., 2015).

### 5.1.3 Syntactic Behavior and Context Specificity

As the introductory quote by Brown et al. (2005) indicates, word knowledge cannot be obtained by looking at isolated words. Words always occur in combination with other words and word difficulty can differ depending on the context and the syntactic behavior of the word. Heilman et al. (2007) highlight the importance of considering the *syntactic complexity* as a difficulty factor for language learners.

Word class The word class has been studied as a difficulty indicator for text-completion exercises by several researchers, but with mixed results. Several researchers find that function words are easier to solve than content words (Brown (1989) and Kobayashi (2002) for cloze tests and Dörnyei and Katona (1992) and Kobayashi (2002) for C-tests), but Klein-Braley (1996) claims that prepositions are often harder for learners. Sigott (1995), on the contrary, could not confirm any effect of the word class on C-test difficulty.

Technically, the word class can be determined by applying a part-of-speech (POS) tagger on the input. Commonly, all articles, prepositions, conjunctions and pronouns are considered to be function words. Another interesting aspect is the probability of a POS sequence. Phrases with an adjective between a determiner and a noun (*the red house*) are more frequent than phrases in which the adjective is placed after the noun (*the house red*).<sup>44</sup> This can be determined by counting the occurrences of POS sequences in manually tagged corpora.

Context specificity When reading a foreign text, learners often guess the meaning of words. These informed guesses are called *lexical inferencing* (Haastrup, 1991). A learner's ability to perform lexical inferencing strongly contributes to her reading comprehension (Zhang and Koda, 2011). According to the *distributional hypothesis*, the inferred meaning of a word largely depends on the context in which the word has been presented (Firth, 1957). This hypothesis has received a lot of attention in the field of meaning acquisition in the L1 and has caused a theoretical shift towards distributional methods. It also initiated

<sup>&</sup>lt;sup>44</sup>Note that the second example is rare, but possible as in We paint the house red.

many experimental studies as the one by McDonald and Ramscar (2001) who show that the similarity of two words is judged differently depending on the context in which they are presented. Distributional properties can provide an explanation how children overcome the poverty of the stimulus and manage the rapid acquisition of new words. The sparse vocabulary knowledge of language learners is a hard obstacle for reading comprehension (Lervåg and Aukrust, 2010) and they have to rely even more on contextual information for lexical inferencing. Very specific contexts can narrow down the possible meaning of a word so that it can easily be guessed by learners. Consider the following example:

### (7) The doctor diagnosed him with Ewing's sarcoma and proposed chemotherapy.

The phrase diagnosed him with is almost always followed by the name of a disease. In combination with the clue chemotherapy, the reader can guess that Ewing's sarcoma is probably a cancer type. This phenomenon can also explain why language learners feel more comfortable with texts from their expert domain. Even though the text might be more complex and the vocabulary very infrequent, their high domain knowledge facilitates the conceptual understanding and thus enables an easier mapping.

The probability of a word in a given context can be determined by higher-order language models as applied in Schwarm and Ostendorf (2005). Additionally, the *context specificity* can be operationalized by using a distributional thesaurus (Lin, 1998), collocation frequencies (Anagnostou and Weir, 2006) and semantic similarity methods (Zesch and Gurevych, 2010). These approaches provide information about the words that are likely to occur in a given context. If all the candidate words are semantically close, the context is very specific and the meaning of an unknown word can probably be approximated more easily. A more detailed description of approaches for measuring candidate fitness can be found in section 6.2. For the context specificity of the solution, the bigram and trigram frequency of the word and the left and right neighbors are calculated.

Lexical inferencing on the basis of context specificity is performed by L1 and L2 learners and is conceptually mediated. In addition, L2 learners can also rely on the orthographic similarity of the unknown word to a word in their mother tongue in order to infer the meaning. The cross-lingual orthographic similarity of two words is called cognateness and we will elaborate on this feature in the next section. Both guessing strategies – lexical and conceptual – might of course also lead to wrong mappings and thus to severe misunderstanding. In order to provide useful learning materials, it is important to be aware of these processes and be able to predict comprehension facilitation by cognates or context specificity as well as probable misconceptions caused by false friends.

## 5.2 Cognates

Cognates are words that have a similar lexical form in the L1 and the L2. A strict definition only considers two words as cognates if they have the same etymological origin, i.e. if they are genetic cognates (Crystal, 2011). Language learners usually lack the linguistic background to make this distinction and will use all similar words to facilitate comprehension regardless of the linguistic derivation. For example, the English word strange has the Italian correspondent strano. The two words have different roots and are therefore genetically unrelated. However, for language learners the similarity is more evident than, for example, the English–Italian genetic cognate father–padre. We thus apply a more operational cognate definition for the language learning setting. We consider as cognates all word pairs that are sufficiently similar to be recognized as related by language learners. This similarity facilitates the word association independent of other features such as frequency or concreteness of the word (De Groot and Keijzer, 2000; Lemhöfer and Dijkstra, 2008). In this section, the importance of cognates for language learning and approaches for measuring cognateness are discussed. The main contribution is the introduction of a new approach for cognate production that works even across different alphabets. The approach is described and evaluated in depth. This contribution has been published in Beinborn et al. (2013) and parts of the general discussion can be found in Beinborn et al. (2014a).

## 5.2.1 Cross-lingual Transfer

For L1 readability calculations, words containing Latin roots will often be considered more difficult because they are usually longer. In addition, they are mainly used in academic contexts and thus have a low frequency in general domain corpora. In L2 settings, on the contrary, these words can facilitate the comprehension for language learners with a romance language background due to their *cognateness*. Tharp (1939) already criticizes that pure frequency counts "ignore the lack of burden of 'gift' words" such as cognates. Ferreira de Souza (2003) studies how cognates contribute to reading comprehension and finds a positive effect for the tasks of skimming and free written recall. In Uitdenbogerd (2005), the readability of French books was judged by English native speakers based on the first 100 words. The results show a correlation between the number of cognates in the text and its assumed readability. This finding was not considered by recent work on readability because most approaches targeted native speakers. The cognateness cues contribute to the advantage of L3 (L4, L5, ...) learners over L2 learners mentioned earlier. Multilingual learners can rely on a wider range of linguistic knowledge and can access a bigger set of potential cognates in order to facilitate lexical association.

The cognateness of words is thus an important feature for word difficulty. However, not all cognates actually have the same meaning. So-called *false friends* look lexically similar but mean different things. Famous examples for false friends are the German word

Czech trigger	Translation	Other associations
nudle	Nudel(n)(15)	
$\check{s}vagr$	Schwager(13)	
$br\acute{y}le$	Brille(12)	$br\"{u}llen(4)$
$\check{s}lak$	Schlag(12)	Schlagsahne(3), schlagen(2)
kned lik	$Kn\ddot{o}del(10)$	$gn\ddot{a}dig(3),niedlich(2)$
$\check{z}old$	Sold(9)	Zoll(5), Gold(2), verkauft(2), Schuld(2)
cil	Ziel(9)	Himmel(2)
$s\'al$	Saal(8)	Salz(13)
$ta\check{s}ka$	Tasche(8)	Aufgabe(4), Tasse(4), Taste(2)
$sk\check{r}i\check{n}$	Schrein/Schrank(5)	Bildschirm/Screen(3), schreien(2)
$farcute{lpha}\check{r}$	Pfarrer(4)	Fahrer(7), fahren(5), Fahrrad(4)
$flcute{e}tna$	$Fl\ddot{o}te(4)$	Flotte(4), fliehen(2)
$val\check{c}\acute{\imath}k$	Walzer(3)	Walze(4), falsch(2)
muset	$m\ddot{u}ssen(3)$	Museum(11), Musik(3), Mus(2), Muse(2)
$tali\check{r}$	Teller(2)	Taler(5), $zahlen(3)$ , $teilen(2)$
knoflik	Knopf(1)	Knoblauch(11), $knifflig$ (4)
šunka	Schinken(1)	Sonne(2), schunkeln(2)

**Table 5.1:** Results of the cognate study: Czech cognates with their correct translations and other associations named by the participants. The number of mentions are provided in brackets and the words are ordered by the number of mentions of the correct translation.

Gift (meaning poison and not present) and the Spanish word embarazada (meaning pregnant and not embarrassed). False friends have the opposite effect on readability; they lead the reader towards a wrong interpretation of the sentence. Both, true and false cognates need to be detected and incorporated into readability measures.

### 5.2.2 Cognate Study

In a small-scale study, we examined whether cognates can facilitate the comprehension of unknown words even in the absence of context and background knowledge of the target language. We selected 17 Czech-German cognates and presented to 15 native German speakers who did not have experience with Eastern-European languages. The participants were asked to name up to 3 guesses for the German translation of the Czech source word. They were not allowed to use a dictionary or any other linguistic material. Table 5.1 gives an overview of the Czech source words together with the correct German translations and other German associations named by more than one participant. The results in table 5.1 show that the participants succeed in guessing the right meaning in most of the cases. Some of the Czech words are strongly associated with their correct German trans-

<sup>&</sup>lt;sup>45</sup>The words were selected from http://de.wikipedia.org/wiki/Tschechische\_Sprache#Deutsche\_Lehnw.C3.B6rter im Tschechischen, accessed: June 1, 2013

lations (e.g. nudle-Nudel, švagr-Schwager) and do not provoke other associations. For other words, the cognateness relation is not strong enough (šunka-Schinken) or the correct cognate is suppressed by stronger false friend associations (e.g. knoflik-Knoblauch, muset-Museum). The results show that the participants were able to guess the correct meaning of a foreign word without any context information. If they had seen the words within an understandable context, they would probably have an even better intuition for the meaning. Consider, for example, the test word muset which is the only verb in the list. Some participants mentioned after the test that the other cognates had primed them for noun meaning, which explains the variety of noun associations named here. In a context that clearly signals the verb property of muset, the noun associations would probably have been suppressed.

Another interesting aspect is the influence of languages besides the L1. As we noted before, people rely on all previously learned languages when trying to detect cognates. Our participants all have very good knowledge of English and almost all of them also know French. In addition, each of them (except for one participant) has studied at least one of Spanish, Italian or Latin. This is also reflected in the associations. For example, the German association Himmel for the Czech word cil is very likely rooted in the similarity to the French word ciel or the Spanish word cielo. Both mean heaven in English and thus trigger the German translation Himmel. The same process might apply when proposing Salz (Spanish: sal, French: sel, English: salt) for sal. English translations might have caused the associations for Aufgabe (=task) for taška, verkauft (=sold) for zold and Bildschirm (=screen) for skrin. It should be noted that all the cross-lingual associations lead to wrong mappings in our examples. These false friend pairs might cause confusion for the learners and need to be handled separately. Czech is not very close to Western European languages which are the background languages of our participants. If they had more experience with Eastern European languages, the cross-lingual mappings could be more successful.

The results of the cognate study show that learners are able to guess the meaning of words that are related to their own language (or to other languages they already know) without having received any education in the new language. A more detailed study by Berthele et al. (2011) comes to the same result. The results confirm the intuition of many language learners and teachers that cognateness of words facilitates their mapping to L1 words and concepts. The ability to guess unknown words also has an impact on the text readability (Zhang and Koda, 2011). A text containing many infrequent words can still be easy to understand for a learner, if the infrequent words exhibit high cognateness.

### 5.2.3 Measuring Cognateness

Determining the cognateness of a word for a learner requires that it be checked against a list of known cognates from all the languages that the learner knows. In previous work on cognates and in our pilot study, the words are manually selected by human experts. How-

ever, there exist several approaches that can be applied for automatic cognate retrieval. As each of the following measures for string similarity and phonetic similarity can only capture a fraction of cognates, they are often applied in combination as features for supervised machine learning classifiers (Sepúlveda Torres and Aluisio, 2011).

String similarity Many different similarity measures have been applied on candidate pairs to distinguish between cognates and unrelated words. The majority are *string similarity* measures that assess the orthographic similarity of two words. The *longest common subsequence ratio* (*LCSR*), for example, calculates the ratio of the length of the longest (not necessarily contiguous) common subsequence and the length of the longer word (Melamed, 1999). Kondrak and Dorr (2004) apply generalizations of *LCSR* to bigrams and trigrams for the detection of confusable drug names. Inkpen et al. (2005) and Montalvo et al. (2012) show that different variations of the *Dice* measure (introduced by Adamson and Boreham (1974)) return the best results. In this context, Dice is defined as "twice the number of shared letter bigrams by the total number of bigrams in both words".

$$Dice(X, Y) = \frac{2 * |bigrams(x) \cap bigrams(y)|}{|bigrams(x)| + |bigrams(y)|}$$
(5.1)

(Brew and McKelvie, 1996) proposed the related measure *XDice* which uses extended bigrams, i.e. trigrams without the middle letter. They use the union of bigrams and extended bigrams, but the description in Inkpen et al. (2005) leads to the conclusion that they apply the formula only on the extended bigrams without the standard bigrams. In this thesis and in our publication (Beinborn et al., 2013), we follow the second interpretation to allow comparison to related work. Other measures for cognate identification focus on the prefix similarity (Simard et al., 1992) and the consonant similarity (Danielsson and Mühlenbock, 2000).

Phonetic similarity The string similarity measures do not capture similarity in the pronunciation of words very well. Different languages often encode the same phoneme with different spelling (e.g. the English word *rice* is pronounced similarly as the German cognate *Reis*, but their *Dice* and *XDice* values are 0). The phonetic similarity measures *Soundex* (Russell and Odell, 1918), *Editex* (Zobel and Dart, 1996) and *Aline* (Kondrak, 2000) calculate the similarity of two words by approximating phonetic characteristics. These approaches have been developed for English and do not scale well for languages with different grapheme to phoneme alignment. The *LexStat* measure (List, 2012) aligns the phonetic transcription of two words to determine their cognateness. The approach relies on phonetic transcription resources that are only available for a few languages.

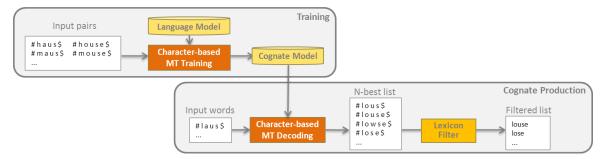


Figure 5.2: Cognate production workflow

Pattern-based approaches Most cognate pairs follow regular production processes as in vision-visión and tradition-tradición. Not all of these processes can be captured by similarity measures; for example, the English-Spanish pair accustomed-acostumbrado has very low similarity values but can easily be identified as related by language learners. This phenomenon can be captured by approaches which focus on the production rules that transform a word into its cognate in a foreign language. The approach by Mulloni (2007) learns the edit operation associations from a list of cognates and generalizes them to candidate rules. Statistical scores are then assigned to each rule to measure the association between the left-hand side and the right-hand side of the rule. Similarly, Gomes and Pereira Lopes (2011) use standard character alignment for the extraction and generalization of substitution patterns and introduce the new measure SpSim. Their results indicate that pattern-based approaches outperform standard string similarity measures in the task of cognate identification.

The pattern-based approaches described above all rely on string alignment that is based on exact matching of characters and can thus not be applied for language pairs with different alphabets. In order to overcome this weakness, we developed a new algorithm for *cognate production (COP)* that can be applied to any language pair sharing cognates (Beinborn et al., 2013). We describe this approach in more detail in the following subsection.

## 5.2.4 Cognate Production Approach

In the *COP* approach, we model the regular production patterns of cognates using *character-based machine translation*.

Figure 5.2 gives an overview of the *COP* architecture. We use the existing statistical machine translation engine *Moses* (Koehn et al., 2007). The main difference between character-based machine translation and standard machine translation is the size of the lexicon. The tokens in *COP* are character n-grams instead of words; therefore less training data is required. Additionally, distortion effects can be neglected as reordering of n-grams is not a

common regular morphological process for cognates.<sup>46</sup> Thus, we deal with less variation than standard machine translation.

Training As training data, we use existing lists of manually collected cognates (see sections 5.2.5 and 5.2.6 for descriptions of the data) and perform the following pre-processing steps. All duplicates, multi-words, and all word pairs that are identical in source and target are removed because they are trivial to find. The remaining words are lowercased and # is added as start symbol and \$ as end symbol of a word. Then all characters are divided by blanks. *Moses* additionally requires a language model. We build an *SRILM* language model (Stolcke, 2002) from a list of words in the target language converted into the format described above. On the basis of the input data, the *Moses* training process builds up a phrase table consisting of character sequences in our case. As a result of the training process, we get a cognate model that can be used to produce cognates in the target language from a list of input test words.

**Cognate production** Using the learned cognate model, *Moses* returns a ranked n-best list containing the *n* most probable transformations of each input word. In order to eliminate non-words, the n-best list is checked against a lexicon list of the target language. The filtered list then represents the set of produced cognates. Note that the list will contain both, true and false cognates, and needs to be further processed to identify false friends.

**Evaluation metrics** In order to estimate the cognate production quality without having to rely on repeated human judgment, we evaluate *COP* against a list of known cognates. The *coverage* indicates the ratio of source words for which our approach was able to produce the correct target cognate.

Existing cognate lists only contain word pairs as cognates, but a word might have several true cognates. For example, the Spanish word m'usica has at least three English cognates: music, musical, and musician. Therefore, even a perfect cognate production process would not always be able to rank the designated gold cognate at the top position. In order to account for this issue, we evaluate the *coverage* using a relaxed metric that counts a positive match if the gold standard cognate is found in the n-best list of cognate productions. We determined n = 5 to provide a reasonable approximation of the overall coverage for the training data and refer to it as Cov@5.

We additionally calculate the mean reciprocal rank (MRR) as

$$MRR = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{1}{rank_i}$$
 (5.2)

<sup>&</sup>lt;sup>46</sup>We set the distortion limit to zero and use these standard parameters: -weight-l 1 -weight-d 0 -weight-w -1 -dl 0 -weight-t 0.2 0.2 0.2 0.2 0.2

where C is the set of input words and  $rank_i$  is the rank of the correct cognate production. The MRR is a metric often used in information retrieval for evaluating any process that produces a ranked list of possible responses. The reciprocal rank is the inverse of the correct rank. The mean reciprocal rank is the average of the reciprocal ranks of the test list. For example, if the target cognate is always ranked second-best, then the MRR would be 0.5.

Note that in the language learning scenario, words that might be associated with the foreign word by learners, but are actually not true cognates also need to be considered (e.g. the English word mystic might also be mistakenly associated with m'usica by language learners). Unfortunately, an evaluation of the false friends produced by COP is not covered by those metrics and thus left to an additional analysis performed in section 5.2.7.

### 5.2.5 Experiments

We conduct a set of experiments that cover different aspects of the cognate production process. First, we test whether the approach is able to learn simple production rules in a minimal model. In a second step, we select optimal parameters and test the influence of the size and quality of the available training data. We then compare the best model to previous work. These experiments are conducted for the language pair English–Spanish. Results for other language pairs are discussed in section 5.2.6. All cognate lists are available at <a href="http://www.ukp.tu-darmstadt.de/data/cognate-production">http://www.ukp.tu-darmstadt.de/data/cognate-production</a>.

As this is the first approach to cognate production, we cannot directly compare the results to a state-of-the-art baseline. A possible baseline could be a random permutation of characters, but it would lead to useless results. In the last experiment, we compare our results to the performance of existing algorithms developed for cognate identification. The experiments are evaluated using coverage and *MRR*. The upper bound for *Cov@5* would presumably be close to 100% for eloquent bilingual speakers. The upper bound for the *MRR* cannot be exactly determined as discussed in the previous section.

**Minimal model** For a first proof of concept, we test whether COP is in principle able to capture the production processes from one language to another by using a very controlled setting. We take a list of ten English–Spanish cognate pairs all following the production process where  $\sim tion$  becomes  $\sim ción$ ; see table 5.2.

We train a model with these word pairs and as a result can successfully produce cognates following the same process, e.g. invitation-invitación. The model is also able to generalize to unseen character sequences. Even though all examples in the training list either terminate in the sequence  $\sim ition$  or  $\sim ation$ , the input fiction correctly produces ficción. We can conclude that COP succeeds in learning the necessary patterns for cognate production.

<sup>&</sup>lt;sup>47</sup>The common evaluation metric for machine translation is the *BLEU* score (Papineni et al., 2002), but this would be misleading in our setting. *BLEU* assigns considerably high scores for partial matches which are not useful for cognate production.

	English	Spanish
Training	tradition application organization administration implication classification nation conjugation information amputation	tradición aplicación organización administración implicación clasificación nación conjugación información
Test examples	$invitation \ fiction$	invitación ficción

**Table 5.2:** Training lists for minimal model

In the following, we investigate whether *COP* can also be applied to noisy training data containing a mixture of many different production processes.

Data We collect and proofread cognates from freely available resources and merge them with the list used by Montalvo et al. (2012) resulting in a list of 3,403 English-Spanish cognates. We split it into a training set (2,403), development set (673), and test set (327). We train a cognate model on the training set and evaluate the cognate production on the test set. We compare the results for different parameters: the character n-gram size used for tokenization, the order of the language model, and the lexicon used for filtering. In addition, the effect of tuning the model on the development set is evaluated. Table 5.3 shows the coverage in the 5 best productions and the *MRR* for each parameter. The results are discussed below.

**N-gram size** We start with the n-gram size parameter that determines the tokenization of the input; the respective format for unigrams, bigrams, and trigrams for the word banc looks as follows:

unigrams: # b a n c \$ bigrams: #b ba an nc c\$ trigrams: #ba ban anc nc\$ Higher order n-grams in general increase the range of the patterns and thus lead to better alignment. However, they also require a larger amount of training data; otherwise the number of unseen instances is too high. In our experiments, bigrams produce slightly better results than unigrams and trigrams; this is in line with findings by Nakov and Tiedemann

<sup>&</sup>lt;sup>48</sup>We used the resources at http://www.cognates.com, http://www.intro2spanish/com/vocabulary/00-lists/\*, and http://www.cognates.org/research/mfcogn.html, all accessed: February 11, 2013.

		Cov@5	MRR
	Unigram	.63	.43
1)	Bigram	.65	.49
	Trigram	.51	.40
2)	LM-order 5	.68	.48
2)	LM-order 10	.65	.49
2)	Web1T-Filter	.68	.52
3)	Wordlist-Filter	.65	.54
4)	Moses Tuning	.66	.54

**Table 5.3:** Parameter selection for *COP*. The settings in bold are kept constant for the other experiments.

(2012). Thus, character bigrams are used in the following experiments. The optimal setting for this parameter might vary with respect to different languages and datasets.

Language model The next parameter is the language model which determines the probability of a sequence in the target language, e.g. a model of order 5 considers sequences of character n-grams up to a maximum length of 5. Order 5 seems to be already sufficient for capturing the regular character sequences in a language. However, the ranks for the order-10 model are slightly better and as our vocabulary is very limited, training the bigger model can still easily be performed. The language model of order 10 is used for the remaining experiments.

**Lexicon filter** For filtering the n-best cognate productions, two different lexicon filter lists are tested. A relatively broad one extracted from the Spanish *Web1T* (Brants and Franz, 2006) word counts, and a more restrictive corpus-based list.<sup>49</sup>

The more restrictive filter decreases the coverage as it also eliminates some correct solutions, but it improves the *MRR* as non-words are deleted from the n-best list and the ranking is adjusted accordingly. The choice of the filter constitutes a trade-off between cognate coverage and the quality of the n-best list. For our language learning scenario, the more restrictive filter is chosen in order to assure high quality results in expense of coverage.

**Machine translation parameters** In the initial experiments, we have set the parameter weights according to introspection. However, it is common practice to tune the machine

<sup>&</sup>lt;sup>49</sup>The stricter word list is available at: http://www.umich.edu/~archive/linguistics/texts/lexica/span-lex.zip accessed: December 12, 2015.

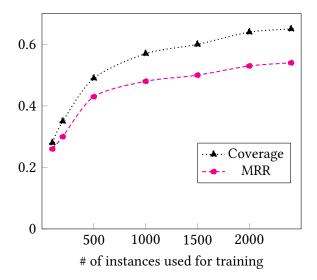


Figure 5.3: COP learning curve

translation parameter weights by applying minimum error rate training (Och and Ney, 2003) using the development set. For comparison, we re-ran the experiments on a tuned model but the difference is minimal. Tuning optimizes the model with respect to the *BLEU* score. For our data, the *BLEU* score is quite high for all produced cognate candidates, but it is not indicative of the usefulness of the transformation. A word containing one wrong character is not necessarily better than a word containing two wrong characters. This explains why tuning has little effect.

Generally, *COP* reaches a coverage of about 65%. If an n-best list with the 100 best translations is considered (instead of only 5), the coverage increases only by less than 1% on average, i.e. the majority of the correct cognates can be found in the top 5. This is also reflected by the high *MRR*. In the following experiments, the optimal parameter setting (highlighted in table 5.3) is used.

Training data size & quality As we have seen in the experiments in section 5.2.5, *COP* is able to learn a production rule from only few training instances. However, the test dataset contains a variety of cognates following many different production processes. Thus, we evaluate the effect of the size of the training data on *COP*. The learning curve in Figure 5.3 shows the results. As expected, both coverage and *MRR* improve with increasing size of the training data, but we do not see much improvement after about 1,000 training instances. Thus, *COP* is able to learn stable patterns from relatively few training instances.

However, even a list of 1,000 cognates is a hard constraint for some language pairs. Thus, the following experiments test whether satisfactory results can also be produced with lower quality sets of training pairs that might be easier to obtain than a list of cognates.

		Size	Cov@5	MRR		Size	Cov@5	MRR
Cognates		1,000	.57	.48		2,403	.65	.54
	UBY	1,000	.53	.47	-	6,048	.69	.56
Translations	UWN	1,000	.50	.43		10,531	.69	.54
	Bing	1,000	.51	.44		5,567	.64	.54
Knowledge-fr	:ee	1,000	.21	.18	-	34,019	.47	.33

Table 5.4: Influence of data quality and size

We extract word pairs from the freely available multilingual resources *UBY* (Gurevych et al., 2012) and *Universal WordNet* (*UWN*, de Melo and Weikum (2009)). *UBY* combines several lexical-semantic resources, but all translations are extracted from *Wiktionary*. <sup>50</sup> *UWN* is based on *WordNet* (Fellbaum, 1998) and *Wikipedia* and provides automatically extracted translations for over 200 languages that are a bit noisier compared to *UBY* translations. Additionally, machine translations for all words from an English word list were queried from the *Microsoft Bing* translation API. <sup>52</sup> In addition, a knowledge-free approach is tested by pairing all words from the English *Web1T* corpus with all words from the Spanish *Web1T* corpus. In order to limit the number of results to a manageable size, we only use every fifth word. The translation pairs always share at least one sense, but this is obviously not the case for the *Web1T* pairs, which will be unrelated in most cases.

In order to increase the ratio of possible cognates in the training data, the *XDice*-measure is used as a string similarity filter with a threshold of 0.425 on the translation pairs.<sup>53</sup> For the knowledge-free pairs, a stricter threshold of 0.6 is used to account for the lower quality.

For an impartial quality comparison, the number of training instances is first limited to 1,000, where (as shown above) the performance increases leveled off. The left part of table 5.4 shows the results for coverage and *MRR*. It can be seen that the results for the translation pairs extracted from *UBY*, *UWN* and *Bing* are only slightly inferior to the use of manually collected cognates for training. The small differences between the resources reflect the different level of linguistic control that has been applied in their creation. The knowledge-free pairs from *Web1T* yield drastically inferior results. We can conclude that training data consisting of selected cognates is beneficial, but that a high-quality list of translations in combination with a string similarity filter can also be sufficient and is usually

 $<sup>^{50}\</sup>mathrm{https://www.wiktionary.org}$ 

<sup>&</sup>lt;sup>51</sup>https://www.wikipedia.org

<sup>&</sup>lt;sup>52</sup>http://www.bing.com/translator

<sup>&</sup>lt;sup>53</sup>The threshold was selected by analyzing the collected cognate pairs. About ~80% of the cognate pairs exhibit an *XDice*-value above this threshold.

	Cov@5	MRR
Dice	.46	.21
XDice	.52	.25
LCSR	.51	.24
SpSim	.52	.22
СОР	.65	.54

Table 5.5: Comparison of different approaches for cognate production

easier to obtain. However, a string similarity filter can only be applied if the two languages share the same alphabet or if transliterations can be obtained (see section 5.2.6).

In a follow-up experiment, the full size of each training set is used. As expected, coverage and *MRR* both increase in all settings. Even with the knowledge-free training set that introduces many noisy pairs, acceptable results can be obtained. This shows that *COP* can be used for the production of cognates, even if no language-specific information beyond a lexicon list is available.

Comparison to previous work In previous work, the task was limited to cognate identification, i.e. the decision whether a candidate word pair is a cognate pair. Most approaches based the decision on the output of similarity measures (Kondrak and Dorr, 2004; Inkpen et al., 2005; Sepúlveda Torres and Aluisio, 2011; Montalvo et al., 2012). COP can now create a cognate list from a simple source word list because it directly produces the target cognate.

In order to compare the previous approaches to *COP*, we "produce" the target cognate for an English word by pairing it with all words from a list of Spanish words.<sup>54</sup> All resulting pairs are considered as candidate pairs for the cognate identification task.

COP is compared to four similarity measures that have been introduced for cognate identification and were described in section 5.2.3: the three frequently used string similarity measures (*LCSR*, *Dice*, and *XDice*), which performed well in Inkpen et al. (2005) and Montalvo et al. (2012), and the productive approach *SpSim* (Gomes and Pereira Lopes, 2011). For each source word, the candidate pairs are ranked according to the corresponding similarity score. The target word of the top-ranked pair is then considered to be the produced cognate. As the similarity measures often assign the same value to several candidate pairs, we get many pairs with tied ranks, which is problematic for computing coverage and *MRR*. Thus, pairs within one rank are randomized and the results are averaged over 10 randomization runs.<sup>55</sup>

<sup>&</sup>lt;sup>54</sup>In order to ensure a fair comparison, the Spanish word list that is also used as lexicon filter in *COP* is used as vocabulary.

<sup>&</sup>lt;sup>55</sup>The average standard deviation is 0.01.

Language pair	Source	Train	Test	Word list
en-es	Cognate data	2,403	327	91,655
es-en	UWN + XDice	10,531	327	203,031
en-de	UWN + XDice	7,944	100	40,794
en-ru	UBY + XDice	4,739	127	32,617
en-el	Bing + XDice	3,131	100	120,750
en–fa	Bing	20,461	241	84,628

**Table 5.6:** Source of the cognate training data for each language pair and size of training set, test set and word list.

Table 5.5 shows the results. The differences between the individual similarity measures are very small; the string similarity measures perform on par with *SpSim*. The low *MRR* indicates that the four measures are not strict enough and consider too many candidate pairs as sufficiently similar. COP clearly performs better than all other measures for both coverage and *MRR*. The results for the similarity measures are comparable to the knowledge-free variant of *COP* (Coverage: 0.47 and *MRR*: 0.33, compare table 5.4). Obviously, *COP* better captures the relevant cognate patterns and thus is able to provide a better ranking of the production list. However, it should be noted that similarity measures might also be a reasonable choice if translation resources are already available for the language pair and we only need to distinguish between cognates and non-cognates (Inkpen et al., 2005). Another advantage of *COP* is its applicability to language pairs with different alphabets (see the following section), whereas the similarity measures can only operate within one alphabet.

## 5.2.6 Multilinguality

The previous experiments showed that *COP* works well for the production of Spanish cognates from English source words. However, in language learning, all languages previously acquired by a learner need to be considered, which leads to a large set of language combinations. Imagine, for example, an American physician who wants to learn German. She has studied Spanish in school and the terminology in her professional field has accustomed her to Greek and Latin roots. When facing a foreign text, she might unconsciously activate cues from any of these languages.

In the following experiments, *COP* is tested for other languages with the same alphabet and across alphabets. In addition, we evaluate how well the cognates produced by *COP* correlate with human judgments.

Data The test data for all language pairs have been manually collected from web resources and have been proof-read by native speakers. For the training data, we used translation pairs from the lexical resource *UWN* for Spanish and German and from *UBY* for Russian. Unfortunately, *UWN* and *UBY* contain only a few examples for Greek and Farsi, so the *Bing* translations of English source words serve as training data for these languages. In order to improve the quality, we apply an *XDice* similarity filter on the resulting word pairs. The German words are filtered directly, the Russian and Greek words are first transliterated into the Latin alphabet. We asked native speakers to spot-check the transliteration results. As the Farsi reviewers considered the output of available transliteration engines for Farsi to be flawed, the training data for Farsi is not filtered. For each language, we also need a word list that is used to build a character-based language model for training the *COP* model and as the word filter for the produced cognates. Table 5.6 provides an overview of the datasets.

Same alphabet In the first experiment, the cognate production is tested in the reverse direction to produce English cognates from Spanish source words. The results in table 5.7 (upper part) show that *COP* works bi-directionally, as the scores for Spanish to English are comparable to those for English to Spanish. In addition, we train a model for another Western European language pair, namely English–German. The results are slightly worse than for English–Spanish which is surprising because English and German are considered to be more closely related. The worse results are due to our quite restrictive lexicon filter, which deletes a substantial number of the target cognates from the n-best list. For example, the German cognates *Karbunkel* and *Faksimile* are actually produced by the *COP* model for the English source words *carbuncle* and *facsimile*, but they are not contained in the word list and are therefore eliminated. We can improve the coverage with a less restrictive filter, but then the *MRR* drops because less words are deleted from the n-best list. In language learning scenarios, assuring high quality is usually considered to be most important; we therefore recommend opting for more restrictive filters.

**Cross-Alphabet** Previous approaches to cognate identification only operate on languages using the same alphabet. As *COP* is able to learn correspondences between arbitrary symbols, it can easily be applied on cross-alphabet language pairs. In the previous experiments,

<sup>&</sup>lt;sup>56</sup>en-de: http://www.german.about.com/library/blcognates\_\*.htm, en-ru: http://www.quizlet.com/281607/russian-cognates-flash-cards , en-el: http://www.everything2.com/title/Greek+Cognates, en-fa: http://www.ebookbrowse.com/farsi-or-persian-cognates-of-english-doc-d97194169, all accessed: February 11, 2013.

<sup>&</sup>lt;sup>57</sup>Threshold for German: 0.425, for Russian and Greek: 0.5

<sup>&</sup>lt;sup>58</sup>Using *ICU*: http://site.icu-project.org/, accessed: February 11, 2013.

<sup>&</sup>lt;sup>59</sup>de: http://www.htdig.org/files/contrib/wordlists/, ru: http://www.artint.ru/projects/frqlist/frqlist-en.php, el: word list extracted from Wikipedia by Torsten Zesch, fa: https://skydrive.live.com/?cid=3732e80b128d016f&id=3732E80B128D016F!3584, all accessed: February 11, 2013.

	Language pair	Cov@5	MRR
Como alphabat	en-es	.65	.54
Same alphabet	es-en	.68	.48
	en-de	.55	.46
	en-ru	.59	.47
Cross-alphabet	en-el	.61	.37
	en-fa	.71	.54

**Table 5.7:** COP results for other languages

cognate pairs that have exactly the same string representation have been excluded. For cross-alphabet pairs, this is not directly possible. Thus, the task is to tackle both standard transliteration (as in the English–Greek pair  $atlas-\acute{\alpha}\tau\lambda\alpha\varsigma$ )<sup>60</sup> and cognate production (as in  $archangel-\alpha\rho\chi\acute{\alpha}\gamma\gamma\epsilon\lambda o\varsigma$ )<sup>61</sup>.

COP is tested with Russian (ru), Greek (el), and Farsi (fa). The lower part of table 5.7 lists the results. Given that those language pairs are considered to be less related than English–Spanish or English–German, the results are surprisingly good. In particular, the production of Farsi cognates works very well even though the training data has not been filtered. Table 5.6 shows that the English–Farsi training data is much larger than the training data for the other language pairs. In the previous experiments, we have seen that bigger training datasets lead to better results. It seems that the size of the training data can balance the quality flaws in this case. For Greek, we observe a combination of high coverage and low *MRR*. This indicates that the gold word is produced by *COP* but other productions are ranked higher. A closer look at the errors reveals that *COP* often produces Greek words in several declinations (e.g. nouns in genitive case) which do not match the designated gold cognate and lead to a worse ranking of the correct target. This problem could be mitigated with a more restrictive lexicon filter.

From the results, we conclude that *COP* also works for other languages and even across alphabet boundaries. We are convinced that the settings should still be tuned to each language pair before applying the approach to real-world tasks. For example, the quality of the training data and the word lists differ strongly in our experiments (which can also be inferred from the varying sizes listed in table 5.6).

For an impression of the cognates produced by *COP*, we compiled a short list of international words that are likely to occur in all languages under study. Table 5.8 lists the top-ranked productions. Note that the gaps in the table are often cases where the absence of a cognate production is an indicator of *COP*'s quality. For example, the Greek words

<sup>&</sup>lt;sup>60</sup>The transliteration of άτλας is átlas.

<sup>&</sup>lt;sup>61</sup>The transliteration of αρχάγγελος is archággelos.

English	Spanish	German	Russian	Greek	Farsi
alcohol	alcohol	alkohol	алкоголь	αλκοολικό	الكل
coffee	café	_	кофей	_	قهوة
director	director	direktor	директор	_	دير
machine	machina	maschine	махина	μηχανή	ماشين
music	músico	musik	_	μουσικής	موسیقی
optimal	óptimo	optimal	оптимальный	_	مطلوب
popular	popular	populär	популярный	_	محبوب
theory	teoría	theorie	теория	θεωρία	تئورى
tradition	tradición	tradition	традиция	_	سنت

**Table 5.8:** Multilingual cognates for English source words produced by *COP* 

for director, popular, and tradition are not cognates of the English words but have a very different form. However, this is not true for the word coffee.

It should be noted that *COP* also produces false friends. Being able to produce both types can be useful for language learning scenarios. Cognates can facilitate comprehension processes, but false friends will lead to confusion. Cognates can thus be used for lexical simplification, whereas false friends can be used for generating challenging exercises that train learners to understand the different meaning between the word in their L1 and the false friend. In order to make use of both types in practical settings, a classification into cognates and false friends needs to be performed.

## 5.2.7 Classifying true and false cognates

In order to balance the cognate aspect, *COP* needs to be combined with a feature detecting false cognates (commonly known as false friends). *False friends detection* can be framed as inter-lingual semantic similarity. Determining the semantic similarity between two words is a well-known task in natural language processing (e.g. *car* and *vehicle* are closer than *car* and *banana*). For false friends detection, similarity needs to be determined across languages. Inkpen et al. (2005) simply check whether the two words are translations of each other in a dictionary.

Mitkov et al. (2008) classify cognate candidates into cognates and false friends using the multi-lingual resource *EuroWordNet* (Vossen, 1998). They apply path-based measures to determine whether two words are semantically similar. Candidates that are close in the resource are considered to be cognates, whereas distant words are classified as false friends. As this approach depends on the coverage of a multi-lingual resource, distributional methods have been tested as an alternative. The assumption is that words which are semantically similar tend to co-occur with the same words. Mitkov et al. (2008), Nakov (2009), and Ljubešic et al. (2013) all use approaches that aim at representing the meaning

of words by vectors of other words and experiment with different co-occurrence counts and vector representations. Each of the distributional methods for false friends detection still requires a translation step at some point. Either the candidate words or the distributionally close words need to be translated for the inter-lingual comparison. Recent work on cross-lingual word embeddings (Klementiev et al., 2012) can provide an alternative for determining semantic similarity for false friends detection, but the approach relies on parallel corpora. In language learning scenarios, high quality needs to be assured. For our text-completion datasets, the focus is on English-German cognates because they are the most common language pair in the exercises. As many resources are available for these two languages, we analyze a resource-based approach for false friends detection here. We use our COP model to produce German cognates for entries in the English word list and vice versa.<sup>62</sup> In the English-German direction, we produce at least one cognate for 14,386 source words and in the other direction for 7,178 source words. As each source word can have several cognates (see section 5.2.4), we focus on the top three productions for each word. It should be noted that COP sometimes only produces one or two potential cognates. Overall, we examine 16,539 potential cognates for German and 10,016 for English and split the pairs into cognates and false friends using two different resources.

Resources We access the multilingual resource *OntoWiktionary* through the *UBY* interface (Meyer and Gurevych, 2012). *OntoWiktionary* contains word senses for lexical entries in English and German and provides cross-lingual links between the entries. We consider all word pairs that share at least one sense as cognates, and all other word pairs as false friends. Words that cannot be found in the resource are labeled as "unknown". In addition, we use an approach for false friends detection that has been developed by the student Tim Feuerbach.<sup>63</sup> The approach relies on the dictionary *Ding* which has been constructed by Frank Richter.<sup>64</sup>

**Results** The results for classifying the *COP* productions into cognates and false friends are listed in table 5.9. It can be seen that the coverage of *OntoWiktionary* is not satisfactory for this task as 70–80% of the pairs cannot be resolved. The Feuerbach approach always returns a decision, so coverage is not an issue. The Ding resource also contains inflected forms which explains the higher ratio of cognates. Both resources have been constructed by humans, therefore the quality of identified cognates can be considered to be high. A first spot check reveals that many false friends such as *orgy-Orgel* (en: pipe organ),

<sup>&</sup>lt;sup>62</sup>The German word list we use is more controlled than the English one which explains the difference in size.
<sup>63</sup>We organized a shared task for false friends detection in the practice class of the lecture "Natural Language Processing and e-Learning" at *TU Darmstadt* in the summer term of 2015. The approach by Tim Feuerbach obtained the best results on the test set. It is available at <a href="https://github.com/timfeu/cognate-classifier">https://github.com/timfeu/cognate-classifier</a>, accessed: March 7, 2016

<sup>64</sup>https://www-user.tu-chemnitz.de/~fri/ding/, accessed: March 7, 2016

		en-de	de-en
	Source words	40,794	109,582
	Source words with <i>COP</i> productions	14,386	7,178
	<i>COP</i> productions with rank<=3	16,539	10,016
OntoWiktionary	Cognates	1,194 ( 7%)	1,132 (11%)
	False Friends	2,242 (14%)	1,881 (19%)
	Unknown	13,103 (79%)	7,003 (70%)
Feuerbach	Cognates	5,105 (31%)	3,811 (38%)
	False Friends	11,434 (69%)	6,205 (72%)

**Table 5.9:** Results of classifying *COP* productions into cognates and false friends using *OntoWiktionary* and the approach by Tim Feuerbach. Productions that are not contained in *OntoWiktionary* are labeled as "Unknown".

match-Matsch (mud), and sticky-stickig (stuffy) are correctly identified. However, a more detailed error analysis of the false friends decision of the two approaches reveals that the cross-lingual links are insufficient in both resources. Many word pairs that are obviously cognates such as duplicity-Duplizität, ecumenic-ökumenisch, and partially-partiell are also labeled as false friends because the resources simply do not contain the translation. Distributional approaches that work with large text corpora could help to overcome these coverage problems, but the output quality of distributional approaches is less controlled and they also require a translation step at some point in the process. We therefore decide to stay with the more conservative approach and use the output of the Feuerbach approach as cognate inventory for the experiments in chapter 7. For future work, multilingual sensedisambiguated word embeddings might be a more promising approach for a better trade-off between coverage and quality (Suster et al., 2016). It might also be helpful to reduce the words to their stems as these might be easier to compare across languages. For example, the cognate *japanisch* (Japanese) produced from the source *japan* is not completely correct, but it should probably not qualify as a false friend either as it will still lead the learners to the correct meaning.

## 5.2.8 Modeling human associations

In order to examine how well *COP* reflects human associations, we aim at producing cognates for the Czech words from the pilot study. COP is trained on a combination of manually collected Czech–German cognates<sup>65</sup> and translation pairs from *UBY* filtered with a very lenient string similarity filter ( $XDice \ge 0.2$ ). The number of training instances is rather

<sup>&</sup>lt;sup>65</sup>These cognates were extracted from http://www.phil.muni.cz/german/mediaev/histsem/dt-tsch-HS. htm and http://de.wikipedia.org/wiki/Tschechische\_Sprache#Deutsche\_Lehnw.C3.B6rter\_im\_Tschechischen, accessed: February 11, 2013.

Czech	Human associations (de)	COP productions (de)
nudle	Nudel (15)	nudel, nadel, ode
$\check{s}vagr$	Schwager (13)	sauger, schwager, berg
$\check{s}lak$	Schlag (12), Schlagsahne (3), schlagen (2)	stak
$br\acute{y}le$	Brille (12), brüllen (4)	brille,brie
cil	Ziel (9), $Himmel$ (2)	set, zelle, teller
$\check{z}old$	Sold(9), Zoll (5), Gold (2), verkauft (2), Schuld (2)	sold, gold, geld
$s\'al$	Salz (13), Saal (8)	set, san, all, saal
$ta\check{s}ka$	Tasche (8), Aufgabe (4), Tasse (4), Taste (2)	task, as, tick
$sk\check{r}i\check{n}$	Schrein (5), Bildschirm/Screen (3), schreien (2)	_
$fl \acute{e}tna$	Flöte (4), Flotte (4), Pfannkuchen (2), fliehen (2)	flut, $filet$
muset	Museum (11), müssen (3), Musik (3), Muse (2), Mus (2)	mus,most,muse,mit
$val\check{c}\imath k$	Walze (4), Walzer (3), falsch (2)	_
$tali\check{r}$	Taler (5), Teller (2), zahlen (3), teilen (2)	teller, taler, ader
$\check{s}unka$	schunkeln (2), Sonne (2), Schinken (1)	sun
knoflík	Knoblauch (11), knifflig (4), Knopf (1)	_

**Table 5.10:** *COP* productions for the words in the pilot study. *COP* only returns lowercased words by default.

small, as a language reform in the 19th century eliminated many Czech words with Austrian or German roots that were common in spoken Czech. Consequently, the model does not generalize as well as for other language pairs (see the column *COP Productions* in table 5.10). However, it correctly identifies cognates like *nudel*, *brille*, and *sold* which are ranked first by the human participants. As we argued above, *COP* also correctly produces some of the "wrong" associations, e.g. *gold* or *taler*. Thus, *COP* is to a certain extent able to mimic the association process that humans apply when identifying potential cognates.

Cognates in text-completion exercises In reading comprehension, cognates are known as facilitators because their meaning can be deduced from the form similarity to a word in the mother tongue (Ringbom and Jarvis, 2009). We therefore assumed that cognate gaps are easier to solve. However, the training data reveals that they are more likely to trigger production problems.

An indicator for production problems is the answer variety of a gap. If a gap has a high answer variety (i.e. the participants provided a high number of different answers), the learners struggled to produce the right answer correctly. The solutions for the 20 gaps with the highest answer variety (33 or more different answers) in the training data of the English C-tests all have a Latin stem: *appropriate*, *skeletons*, *tempting*, *extract*, *ancient*, *private*,

<sup>&</sup>lt;sup>66</sup>See for example http://mluvtecesky.net/en/introduction/country\_language/language\_history, accessed: March 3, 2016.

 $<sup>^{67}</sup>$ Coverage (0.4) and MRR (0.27) are not representative as the test set is too small.

design, concentrations, state-of-the-art, scientists, modern, examined, constant, essential, stable, entering, basis, synthetic, cost, demands. In contrast, the 20 gaps with the lowest answer variety (5 or less different answers) are very basic vocabulary: longer, coffee, coffee, in, water, very, give, you, for, people, living, other, number, water, water, from, over, you, over

The production problems might be related to the rare character combinations and the lower frequency of words with Latin stem. In addition, these words might not be part of the students' active vocabulary and are only guessed because they occur as cognates in the students' L1. This assumption is supported by the fact that many of the cognate answers contain the right stem but resemble orthographic principles from other languages, e.g. for skeletons we find \*skellets, \*skelleton(s), \*skelets, \*skelletts, \*skeletton(s), \*skeltons, \*skeletes, and \*skelette(s). A similar observation has been made by Nicolai et al. (2013). They identify misuses or misspellings of cognates as a feature for the task of natural language identification. We have a closer look at spelling difficulties in the following section.

### 5.2.9 Conclusions

In the previous sections we have introduced a new approach for character-based machine translation that requires only a small amount of training data and is able to produce cognates for many language pairs including languages with different alphabets. An inspection of the results indicates that this approach succeeds well in capturing morphological patterns. This finding has been used in subsequent works by Scherrer and Sagot (2014), Scherrer and Erjavec (2015), and Ling et al. (2015) and is considered to be beneficial for machine translation. The productive morphological processes could also be used to generate neologisms (e.g. for name creation). If few resources are available, *COP* can be a good starting point to obtain cognates. However, it requires additional editing steps to further improve the quality and to distinguish between true and false cognates. In practical language learning scenarios, we can usually rely on translation resources. In this case string similarity measures provide a good approximation for cognateness. However, the coverage can be improved if string similarity measures are coupled with measures capturing morphological processes like *SpSim* and *COP*.

# 5.3 Spelling Difficulty

The irregularities of spelling have been subject to debates for a long time in many languages. Spelling difficulties can lead to substantial problems in the literacy acquisition and to severe cases of dyslexia Landerl et al. (1997). Learning orthographic patterns is even harder for foreign language learners because the phonetic inventory of their mother tongue might

<sup>&</sup>lt;sup>68</sup>DE: Skelett, FR: squelette, ES: esqueleto, NL: skelet

be quite different. Thus, they have to learn both: the new sounds and their mapping to graphemes. English is a well-known example for a particularly inconsistent grapheme-to-phoneme mapping. For example, the sequence *ough* can be pronounced in six different ways as in *though*, *through*, *rough*, *cough*, *thought* and *bough*.<sup>69</sup>

Spelling problems in text-completion exercises In the training data of the exercises, we also find a plethora of spelling errors. The gap *appropriate* for example triggers 67 different answers, among them spelling errors such as \*appropriate, \*appropriate, \*appropriate. Spelling errors are usually considered as normal errors in text-completion exercises, because the generous time limit allows the students to revise their solutions (Raatz and Klein-Braley, 2002) and the line between a variant of the solution and a wrong solution cannot be clearly drawn. This makes *spelling difficulty* an important aspect of gap difficulty. In this section, we analyze whether the spelling difficulty of a word can be predicted. This contribution has been published in Beinborn et al. (2016). The data and the code is available at www.ukp.tu-darmstadt.de/data/spelling-correction/spelling-difficulty-prediction/.

Related work Most work related to spelling in natural language processing has focused on *error correction* (see section 2.2.3 and the overviews by (Ng et al., 2013, 2014)). Analyses of spelling problems have mainly been performed for the purpose of improving spelling corrections. Deorowicz and Ciura (2005) identify three types of causes for spelling errors (mistyping, misspelling and vocabulary incompetence) and model them using substitution rules. Toutanova and Moore (2002) use the similarity of pronunciations to pick the best correction for an error resulting in an improvement over state-of-the-art spell checkers. Boyd (2009) build on their work but model the pronunciation of non-native speakers, leading to slight improvements in the pronunciation-based model. The goal for this thesis is slightly different: we want to predict spelling difficulties for language learners.

Modeling the spelling difficulty of words could also have a positive effect on spelling correction because spelling errors would be easier to anticipate. Another important line of research is the development of spelling exercises. A popular recent example is the game *Phontasia* (Berkling et al., 2015). It has been developed for L1 learners but could probably also be used for L2 learners. In this case, the findings on cross-lingual transfer could be integrated to account for the special phenomena occurring with L2 learners.

## 5.3.1 Measuring spelling difficulty

Analyses of English spelling difficulties have a long tradition in pedagogical and psycholinguistic literature, but to the best of our knowledge the task of predicting spelling difficulty

<sup>&</sup>lt;sup>69</sup>IPA pronunciations according to https://en.wiktionary.org: /δου/, /θτυ/, /ιλf/, /kɔf/, /θɔt/, and /baυ/

has not been tackled before. In this section, we operationalize the analytical findings on spelling difficulty into features that can be derived automatically.

In general, three sources of spelling errors can be distinguished: i) errors caused by physical factors such as the distance between keys on the input device or omitted character repetitions, ii) errors caused by look-ahead and look-behind confusion (e.g. puclic-public, gib-big), and iii) errors caused by phonetic similarity of letters (e.g. vowel confusion vis-able-visible). Baba and Suzuki (2012) analyze spelling errors committed by English and Japanese native speakers using keystroke logs and find that the first two types are usually detected and self-corrected by the learner whereas phonetic problems remain unnoticed. In the text-completion exercises, the learners are encouraged to review their answers, so the focus should be on the type of spelling errors that usually remain undetected by learners.

Several analyses focus on phonetic ambiguity as an important factor of spelling difficulty. Frith (1980), for example, compares the spelling errors of good and poor readers and shows that the good readers only produce phonetic misspellings whereas the poor readers (which she called mildly dyslexic) often produce non-phonetic misspellings. As foreign language learners are not commonly dyslexic, our focus lies on the phonetic problems. Cook (1997) compares English spelling competence for L1 and L2 users. She confirms that the majority of spelling errors by all three groups (L1 children, L1 adults, L2 adults) are due to ambiguous sound–letter correspondences. Peereman et al. (2007) provide a very good overview of factors influencing word difficulty and also highlight the importance of consistent grapheme–phoneme correspondence.

In the following, we describe seven features that we implemented for spelling difficulty prediction: two word difficulty features (length and frequency) and five phonetic features (grapheme-to-phoneme ratio, phonetic density, character sequence probability, pronunciation difficulty and pronunciation clarity).

Traditional word features for spelling The two standard features word length and word frequency, that have been introduced in section 5.1, have proven to be good baselines for the spelling scenario. Medero and Ostendorf (2009) analyze vocabulary difficulty based on information found in Wiktionary and find that short length and high frequency are good indicators for simple words, confirming previous work. Put simply, the probability of producing an error is increased by the number of characters that need to be typed. For frequent words, the probability that the learner has been exposed to this word is higher and therefore the spelling difficulty should be lower. We determine the length of a word by the number of characters and the frequency is represented by the unigram log-probability of the word in the Web1T corpus (Brants and Franz, 2006).

**Orthographic depth** Dela Rosa and Eskenazi (2011) analyze the influence of word complexity features on the vocabulary acquisition of L2 learners and show that words which

follow a simple one-to-one mapping of graphemes to phonemes are considered to be easier than one-to-many or many-to-one mappings as in *knowledge*.<sup>70</sup> The *orthographic depth* can be expressed as the *grapheme-to-phoneme ratio* (the word length in characters divided by the number of phonemes). For English, we calculate the number of phonemes based on the phonetic representation in the *Carnegie Mellon University Pronouncing Dictionary*.<sup>71</sup> For Italian and German, a comparable pronunciation resource is not available. However, as the orthography of these two languages is more regular than for English, the pronunciation of a word can be approximated by rules. We use the grapheme-to-phoneme transcription of the text-to-speech synthesis software *MaryTTS* version 5.1.1 (Schröder and Trouvain, 2003) to determine the phonetic transcription for Italian and German. We will refer to transcriptions obtained from these resources as gold transcriptions.

**Phonetic density** The *phonetic density* has also been mentioned as a potential cause for spelling difficulty, but it has not yet been studied extensively (Joshi and Aaron, 2013). It is calculated as the ratio of vowels to consonants. Both extremes – words with very high density (e.g. *aerie*) and very low density (e.g. *strength*) – are likely to cause spelling difficulties.

Character sequence probability We assume, that the correspondence of graphemes and phonemes in a word is less intuitive if the word contains a rare sequence of characters (e.g. *gardener* vs. *guarantee*). To approximate this, we build a language model of character trigrams that indicates the *probability of a character sequence* using the framework *berkeleylm* version 1.1.2 (Pauls and Klein, 2011). The quality of a language model is usually measured as the perplexity, i.e. the ability of the model to deal with unseen data. The perplexity can often be improved by using more training data. However, in this scenario, the model is supposed to perform worse on unseen data because it should model human learners. In order to reflect the sparse knowledge of a language learner, the model is trained only on the 800–1000 most frequent words from each language. We refer to these words as the *basic vocabulary*.<sup>72</sup>

**Pronunciation Difficulty** Furthermore, we try to capture the assumption that a spelling error is more likely to occur if the grapheme-phoneme mapping is rare as in Wednesday. The sequence ed is more likely to be pronounced as in simple past verbs or as in Sweden.

English: http://ogden.basic-english.org

German: http://www.languagedaily.com/learn-german/vocabulary/common-german-words

Italian: https://en.wiktionary.org/wiki/Wiktionary:Frequency\_lists/Italian1000

All accessed: February 5, 2015

<sup>&</sup>lt;sup>70</sup>Grapheme length: 9, phoneme length: 5

<sup>&</sup>lt;sup>71</sup>http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed: December 5, 2015

<sup>&</sup>lt;sup>72</sup>We use the following lists:

We approximate this by building a phonetic model using *Phonetisaurus*, a tool that is based on finite state transducers which map characters onto phonemes and can predict pronunciations for unseen words.<sup>73</sup>

Analogous to the character-based language model, the phonetic model is also trained only on words from the *basic vocabulary* in order to reflect the knowledge of a language learner. Based on this scarce data, the phonetic model only learns the most frequent character-to-phoneme mappings and assigns higher *phonetic scores* to ambiguous letter sequences. We use this score as indicator for the *pronunciation difficulty*.

**Pronunciation clarity** Even if the learner experiences low pronunciation difficulty, she might still come up with a wrong pronunciation. For example, many learners are convinced that *recipe* should be pronounced / Isarp/. To model the discrepancy between expected and true pronunciation, we calculate the *Levenshtein* distance (Levenshtein, 1966) between the produced pronunciation by the phonetic model and the gold transcription as *pronunciation clarity*.

In summary, seven features have been implemented for the purpose of spelling difficulty prediction: two traditional features (length and frequency) and five phonetic features (orthographic depth, phonetic density, character sequence probability, pronunciation difficulty and pronunciation clarity). To the best of our knowledge, the task of predicting spelling difficulty is a new task. The orthographic depth and the phonetic density are common phonetic features that have been calculated before. The other three phonetic features are new contributions of this thesis. We use existing concepts from natural language processing in a new way to predict spelling difficulty. In particular, the idea of using extremely limited training data (and thus restricting the quality of the models) in order to simulate learner knowledge is a new approach.

## 5.3.2 Corpora

In order to evaluate the described model for predicting spelling difficulty, we need suitable data. For this purpose, we extract spelling errors from three corpora of annotated learner essays. The corpora contain annotations for a wide range of errors including spelling, grammar, and style categories. The EFC and the FCE corpus contain essays by learners of English and the Merlin corpus contains essays by learners of German and Italian.<sup>74</sup>

EFC The *EF-Cambridge Open Language Database* (Geertzen et al., 2012) contains 549,326 short learner essays written by 84,997 learners from 138 nationalities. The essays have been submitted to *Englishtown*, the online school of *Education First*. For 186,416 of these

<sup>&</sup>lt;sup>73</sup>http://code.google.com/p/phonetisaurus, accessed: January 20, 2015

<sup>&</sup>lt;sup>74</sup>It also contains essays by Czech learners, but this subset is significantly smaller than the ones for the other two languages and is therefore not used here.

essays, teachers provided correction annotations. We extract 167,713 annotations with the tag SP for spelling error. To our knowledge, this is by far the biggest available corpus with spelling errors from language learners.

FCE The second dataset is part of the *Cambridge Learner Corpus* and consists of learner answers for the *First Certificate in English (FCE)* exam (Yannakoudakis et al., 2011). It contains 2,488 essays by 1,244 learners (each learner had to answer two tasks) from 16 nationalities. The essays have been corrected by official examiners. We extract 4,074 annotations with the tag S for spelling error.

Merlin The third dataset has been developed within the EU-project MERLIN (Boyd et al., 2014) and contains learner essays graded according to the Common European Reference Framework (CEFR). The 813 Italian and the 1,033 German samples have been obtained as part of a test for the European language certificate (TELC). A subset of 752 German essays and 754 Italian essays was annotated with target hypotheses and error annotations by linguistic experts. We extract 2,525 annotations with the tag O\_graph from the German essays and 2,446 from the Italian essays. Unfortunately, the correction of the errors can only be extracted if the error annotation is properly aligned to the target hypotheses which is not always the case. We ignore the errors without correction, which reduces the set to 1,569 German and 1,761 Italian errors. In the following, we refer to the German subset as M-DE and the Italian subset as M-IT.

## 5.3.3 Spelling Error Extraction

As the corpora use different annotation formats and guidelines, we implement an extraction pipeline to focus only on the spelling errors. We apply additional pre-processing and compute the spelling error probability as an indicator for spelling difficulty.

**Pre-processing** We remove all spelling errors that only mark a change from lowercase to uppercase (or vice versa) and numeric corrections (e.g. 1 is corrected to one) as these are rather related to stylistic conventions than to spelling. We lowercase all words, trim whitespaces and only keep words which occur in a word list and consist of at least three letters (to avoid abbreviations like ms, pm, oz).

**Spelling error probability** For this evaluation, we take an empirical approach for quantifying error probability. A spelling error s is represented by a pair consisting of a misspelling e and the corresponding correction e. The error frequency  $f_e$  of a word e in the dataset e

 $<sup>^{75}</sup>$ Some corrections have two different tags; we only extract those with a single SP tag.

<sup>&</sup>lt;sup>76</sup>We use the word list package provided by Ubuntu for spell-checking: http://packages.ubuntu.com/de/precise/text/, languages wamerican, wngerman, wfrench, accessed: April 15, 2016

		EFC	FCE	M-DE	M-IT
Words	All	7,388,555	333,323	84,557	57,708
	Distinct	23,508	7,129	3,561	3,760
Spelling Errors	All	133,028	3,897	1,653	1,904
	Distinct	7,957	1,509	719	747
Ratio Errors/Words	Distinct	.34	.21	.20	.20

Table 5.11: Extracted words and spelling errors after pre-processing

is then determined by the number of times it occurs as a correction of a spelling error independent of the actual misspelling. The number of spelling errors  $S_D$  in the dataset is determined by summing over the error frequencies of all words in the dataset. To quantify the distinct spelling errors, we count all words with  $f_e \ge 1$  once.

$$s = (e, c) \tag{5.3}$$

$$f_e(w) = \sum_{s_i \in D} |w = c_i| \tag{5.4}$$

$$S_D = \sum_{w_i \in D} f_e(w_i) \tag{5.5}$$

The numbers of extracted words and errors are summarized in table 5.11. It can be seen that the EFC corpus is significantly bigger than the other corpora. The spelling errors in the EFC corpus are spread over many words leading to a higher ratio of erroneous words over all words.

The pure error frequency of a word can be misleading, because frequently used words are more likely to occur as a spelling error independent of the spelling difficulty of the word. Instead, we calculate the *spelling error probability* for each word as the ratio of the error frequency over all occurrences of the word (including the erroneous occurrences).

$$p_{err}(w) = \frac{f_{err}(w)}{f(w)} \tag{5.6}$$

The words are then ranked by their error probability to quantify spelling difficulty.<sup>77</sup> This is only a rough approximation that ignores other factors such as repetition errors and learner ability because detailed learner data was not available for all corpora. In future work, more elaborate measures of spelling difficulty could be analyzed (see for example Ehara et al. (2012)).

<sup>&</sup>lt;sup>77</sup>In the case of tied error probability, the word with the higher error frequency is ranked higher. In the case of an error frequency of zero for both words, the word with the lower correct frequency is ranked higher.

Corpus	high error probability	low error probability
EFC	$departmental,\ spelt,\ invincible$	boy, car, crime
FCE	synthetic, millennium, mystery	weeks, feel, rainbow
M-DE	tschüss, nächsten, beschäftigt	damit, machen, gekauft
M-IT	messagio, lunedí, caffè	$rossi,\ questo,\ tempo$

Table 5.12: Examples for high and low spelling error probability

Training and test data An inspection of the ranked error probabilities gives the impression that the spelling difficulty of a word is a continuous variable. However, the number of spelling errors is too small to sharply distinguish between a spelling error probability of 0.24 and of 0.26, for example. Instead, we only focus on the extremes of the scale. The *n* highest ranked words are considered as samples for high spelling difficulty and the *n* lowest-ranked words form the class of words with low spelling difficulty. As additional constraint, the errors should have been committed by at least three learners in the EFC dataset and by two learners in the other datasets. For the EFC dataset, we extract 500 instances for each class, and for the FCE dataset 300 instances. 200 instances (100 per class) are used for testing in both cases and the remaining instances are used for training. We find an overlap of 52 words that have a high spelling error probability in both English datasets. As the Merlin corpus is significantly smaller, we only extract 100 instances per class for German and Italian. 140 instances are used for training and 60 for testing. Table 5.12 provides some examples for high and low spelling error probabilities.

### 5.3.4 Experiments

The following experiments test whether it is generally possible to distinguish between words with high and low spelling error probability using the features described above. The classifiers are trained using *support vector machines* as implemented in *Weka* (Hall et al., 2009). The features are extracted using the *DKPro TC* framework (Daxenberger et al., 2014). Significance is calculated using McNemar's test (McNemar, 1947).

**Feature Analysis** In a first step, the predictive power of each feature is evaluated by performing ten-fold cross-validation on the training set. The results in the upper part of table 5.13 are quite similar for the two English corpora. Around 80% of the test words are classified correctly and the most predictive features are the word length and the pronunciation difficulty. It should be noted that the two features are correlated (Pearson's r: 0.67), but they provide different classifications for 131 of the 800 EFC instances in the cross-validation setting. The results for Italian are slightly worse than for English, but show the same pattern for the different features. For German, the features pronunciation difficulty and frequency perform slightly better than the length feature. The two features orthographic depth and

		EFC	FCE	M-DE	M-IT
	Random Baseline	.500**	.500**	.500**	.500**
	Orthographic Depth	.482**	.462**	.427**	.622**
	Phonetic Density	.483**	.349**	.564**	.508**
Individual	Character Sequence Probability	.706**	.642**	.736	.563**
Features	Pronunciation Clarity	.635**	.677**	.722	.683
	Pronunciation Difficulty	.792**	.792**	.828	.731
	Frequency	.634**	.742**	.778	.728
	Length	.809	.827	.747	.769
Combined	Length + Frequency + Pronunciation Diff.	.822	.832	.828	.792
Combined	All Features	.835	.847	.814	.778

**Table 5.13:** Feature analysis for spelling difficulty using 10-fold cross-validation. The prediction results are expressed as accuracy. Significant differences compared to the result with all features are indicated with \*\*(p<0.01).

	EFC	FCE	M-DE	M-IT
Random baseline	.500	.500	.500	.500
Length + Frequency + Pronunciation Difficulty	.840	.865	.766	.817
All Features	.840	.870	.800	.815

**Table 5.14:** Spelling difficulty prediction on the test set for both corpora. The prediction results are expressed as accuracy.

phonetic density are not predictive for the spelling difficulty and only perform on chance level for all four datasets. We additionally train a model build on the three best performing features length, frequency, and pronunciation difficulty as well as one using all features. It can be seen that the results improve slightly compared to the individual features. Due to the rather small datasets and the correlation between the features, the differences between the best performing models are not significant.

In general, the accuracy results are comparable across languages (78–85%) indicating that it is possible to distinguish between words with high and low spelling error probability. In the following, we test whether the models can generalize to the unseen test data.

**Prediction Results** After these analyses, the two combined models are evaluated on the unseen test data. The results in table 5.14 show that the models scale well to the test set and yield accuracy results that are slightly better than in the cross-validation setting. Again, the results of the two combined models are not found to be significantly different. There are two explanations for this. On the one hand, the test set is quite small (200 instances for English, 60 instances for German and Italian) which makes it difficult to measure sig-

Train Corpus Test Corp			Corpus		
		EFC	FCE	M-DE	M-IT
	# inst.	200	200	60	60
EFC	800	.840	.772	.703	.634
FCE	600	.764	.870	.767	.766
M-DE	140	.659	.829	.800	.796
M-IT	140	.397	.540	.780	.815

**Table 5.15:** Spelling difficulty prediction on the full set across corpora. The prediction results are expressed as accuracy. The number of instances is indicated in brackets for each dataset. The two classes are equally distributed.

nificant differences. On the other hand, this result indicates that length, frequency and pronunciation difficulty are very predictive features for the spelling difficulty and the other features only have insignificant effects. The finding that longer words are more likely to produce misspellings is not surprising. For deeper psycholinguistic analyses it might be useful to balance the spelling data with respect to the word length. In such a scenario, phonetic aspects would presumably become more important. However, as we want to model the probability that a learner makes a spelling error, we need to take the length effect into account as an important indicator.

Cross-corpus comparison The above results have shown that the prediction quality is very similar for the two English corpora. To analyze the robustness of the prediction approach, we compare the prediction quality across corpora by training on all instances of one corpus and testing on all instances of another corpus. We also include the German and Italian corpus to this cross-corpus comparison to evaluate the language-dependence of spelling difficulty.

The results in table 5.15 show that the accuracy for cross-corpus prediction generally decreases compared to the previous results of in-corpus prediction (which are listed in the diagonal of the result matrix), but still remains clearly above chance level for English and German. In contrast, training on the Italian corpus leads to bad results for the two English corpora. It is interesting to note that a model trained on the German spelling errors performs better on the FCE words than a model trained on the English errors from the EFC corpus. The FCE and the Merlin corpus have been obtained from standardized language examinations whereas the EFC corpus rather aims at formative language training. In the second scenario, the learners are probably less prepared and less focused leading to more heterogeneous data which could explain the performance differences across corpora.

Correct	Brazilian	Mexican	Chinese	Russian	German
attention	atention(27) attencion(10) atencion(3)	atention(13) attencion(1) attentio(1)	$attaention(1) \ atttention(1) \ -$	$attantion(5) \\ atantion(1) \\ atention(1)$	
departmental	departament(10) departamente(1) departaments(1)	$\begin{array}{c} department(1) \\ - \\ - \end{array}$	$deparment(2) \\ deparmental(1) \\ deprtment(1)$	_ _ _	_ _ _
hamburger	$hamburguer(2) \ hamburguers(2)$	hamburguer(2)	$hamburg \ hamburgs(1)$		
engineer	$engeneer (17) \ ingineer (2) \ ingener (2)$	$enginner(25) \ engeneer(8) \ engenier(4)$	$engneer(5) \\ engeneer(4) \\ enginner(3)$	$engeneer(14) \\ engeener(3) \\ ingener(2)$	$ingeneur(2) \ engeneer(2) \ ingeneer(2)$
marmalade					marmelade(3)

**Table 5.16:** Most frequent misspellings for selected examples

## 5.3.5 Error Analysis

For a more detailed analysis, we take a closer look at the mis-classifications for the EFC dataset. In a second step, we analyze spelling errors with respect to the L1 of the learners.

Mis-classifications The following words were classified as having high error probability, but have a low error probability in the corpus data: references, ordinary, universal, updates, unrewarding, incentives, cologne, scarfs, speakers, remained, vocals. It seems surprising that all those words should have a low error probability. A possible explanation could be that the words had been mentioned in the task description of the essays and are therefore frequently used and spelled correctly. Unfortunately, the task descriptions are not published along with the corpus and we cannot take this factor into account.

The words that were erroneously classified as words with low spelling error probability are generally shorter: *icy*, *whisky*, *cried*, *curry*, *spelt*, *eight*, *runway*, *tattoo*, *daughter*, *farmers*, *discreet*, *eligible*, *diseases*, *typical*, *gallery*, *genre*, *mystery*, *arctic*, *starters*, *stretch*, *rhythm*. In several cases, we see phenomena for which features are available, e.g. a low vowel-consonant ratio in *stretch* and *rhythm*, an infrequent grapheme-to-phoneme mapping in *genre*, a low character sequence probability in *tattoo*. Unfortunately, these features seem to be overruled by the length feature.

In other examples, we observe phenomena that are specific to English and are not sufficiently covered by our features such as irregular morphology (*icy*, *spelt*, *cried*). This indicates that features which model language-specific phenomena might lead to further improvements.

Influence of the L1 As phonetic features have a strong influence on spelling difficulty, we assume that the L1 of the learners plays an important role. For example, arctic is misspelled as \*artic, gallery as \*galery and mystery and typical are spelled with i instead of y. These misspellings correspond to the correct stem of the respective word in Spanish, Italian and Portuguese. In the following, we thus have a closer look at the influence of the L1.

The EFC corpus comprises essays from a very heterogeneous group of learners, but 71% of the annotated essays are written by learners from five nationalities, namely Brazilian, Chinese, German, Mexican, and Russian. For comparative analyses, we also extracted the spelling errors specific to each of these five nationalities. Table 5.16 shows anecdotal examples of cross-lingual influence on spelling difficulties. For the word attention, it can be seen that the Russian learners are tempted to use an a as second vowel instead of an e. For the Brazilian and Mexican learners, on the other hand, the duplication of the t is more problematic because doubled plosive consonants do not occur in the orthography of their mother tongue. L1-specific errors are often due to the existence of similar words – so-called cognates - in the native language of the learner. The word departmental is particularly difficult for Brazilian and Chinese learners. While most Brazilian learners erroneously insert an a due to the cognate departamento, none of the Chinese learners commits this error because a corresponding cognate does not exist. The Brazilian and Mexican misspellings of hamburger can also be explained with the cognateness to hamburguesa and hambúrguer respectively. A g followed by an e is pronounced as a fricative /x/ in Spanish and not as a plosive /g/. This indicates that the phonetic features should model the differences between the L1 and the L2 of the learner.

The word *engineer* provokes a large variety of misspellings. A common problem is the use of *e* as the second vowel, which could be explained with the spelling of the cognates (br: *engenheiro*, de: *Ingenieur*, ru: *инженер* transliterated as *inzhener*). However, the misspelling by the Mexican learners cannot be explained with cognateness because the Spanish spelling would be *ingeniero*. The spelling of *marmalade* with an *e* seems to be idiosyncratic to German learners.

The above analyses are only performed on an anecdotal basis and need to be backed up with more thorough experimental studies. The examples support the intuitive assumption that cognates are particularly prone to spelling errors due to the different orthographic and phonetic patterns in the L1 of the learner. This can explain the relation observed between cognates and productive difficulties in the text-completion exercises which was discussed in the previous section. The learners in the EFC corpus also differ in proficiency (e.g. German learners seem to be more advanced than Brazilian learners) which might also have an influence on the spelling error probability of words. However, it is complicated to disentangle the influence of the L1 and of the L2 proficiency based on the current data and we leave this analysis to future work.

### 5.3.6 Conclusions

We have seen that the traditional word features *length* and *frequency* are a good approximation for spelling difficulty, but they do not capture phonetic phenomena. The newly developed feature *pronunciation difficulty* can close this gap and complements the traditional features for spelling difficulty prediction. We conclude that the spelling error probability of a word can be predicted to a certain extent. The prediction results are stable across corpora and can even be used across languages. Further improvement could probably be reached by modeling language-specific features (e.g. morphology) and by taking the L1 of the learner into account. Our main goal is the prediction of exercise difficulty. We have seen that the orthographic depth and the phonetic density do not seem to be very relevant for spelling. In chapter 7, we evaluate whether the remaining five spelling features can contribute predictive power to exercise difficulty prediction.

# 5.4 Chapter Summary

This chapter has introduced measures for approximating the difficulty of words for language learners. In an overview of existing measures, traditional word features that measure word familiarity, morphological complexity, and syntactic behavior have been explained. In order to account for cross-lingual item transfer, the cognateness of words needs to be considered. We presented a new approach for cognate production that works for many languages including language pairs with different alphabets.

Insights into the training data of the exercise datasets have shown that productive difficulties resulting in spelling errors play an important role. We discussed the concept of spelling difficulty and identified the major causes. Existing features have been combined with three new phonetic features to predict the spelling difficulty of words in a corpus of learner essays. In chapter 7, we evaluate to which extent these word difficulty features contribute to the difficulty of exercises.

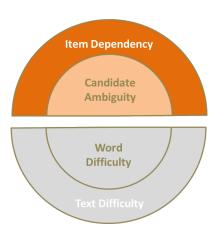
This chapter focused on local difficulty. An overview of the implemented word difficulty features can be found in the appendix in section B. However, all of the discussed word difficulty features can of course also be aggregated over all words in a text, if necessary. A corresponding readability feature could thus be derived by measuring the proportion of words in the text that exhibit a certain feature (e.g. the ratio of cognates over all content words in the text).

### CHAPTER 6

# The Influence of the Exercise Format

"I didn't fail the test, I just found 100 ways to do it wrong."

— Benjamin Franklin



**Figure 6.1:** The exercise format

The difficulty of an exercise is determined by the exercise content and the exercise format. The previous two chapters focused on the exercise content: the difficulty of the underlying text and the individual words in it. However, the same content can be transformed into exercises with different levels of difficulty depending on the exercise format. In this chapter, we analyze three phenomena that are related to the exercise format: the candidate space, the candidate ambiguity, and the item dependency.

In section 2.3.4, we have seen that the main format factors for text-completion exercises are the gap type, the answer format, and the deletion rate. These three factors determine the degree of information reduction that is applied on the original text. The learner needs to re-create the reduced information to solve the exercise.

Spolsky (1969, p. 6) describes information as "a reduction in uncertainty by eliminating certain possibilities". As a corollary, reduced information increases the number of possibilities. Natural language provides information using redundant ways so that native speakers can cope with a certain degree of redundancy reduction in noisy input. Language learners, on the other hand, perceive *redundancy reduction* directly as information reduction and therefore as an increase of possibilities. Spolsky shows that a higher degree of redundancy reduction leads to higher difficulty.

Previous work on text-completion exercises has shown that the gap type has an influence on the difficulty of an exercise. For C-test and X-text gaps, 50% of the word is deleted to create a gap and 100% for cloze test gaps. Köberl and Sigott (1994) additionally experiment with deleting 25% and 33% of the word. They find that a gap is more difficult if a larger proportion of the word is deleted. This is in line with the assumption that a higher degree of redundancy reduction leads to increased difficulty.

However, the authors show that the position of the deletion also plays a role. In their experiments, word-initial deletion (as in the X-test) leads to more difficult gaps than word-final deletion (as in the C-test). Consider the word <code>coruscate</code>, for example. For the C-test gap <code>coru\_</code>, the solution is the only candidate, but for the X-test gap <code>\_cate</code>, 36 candidates can be found. These findings lead to the assumption that the degree of redundancy reduction should not only be determined based on the proportion of deleted characters. Instead, we propose to measure the degree of redundancy reduction directly by the increase of possibilities, i.e. the average number of candidates for a gap. The first section presents an approach to calculate the average <code>candidate space</code> for each gap type.

The candidate space can be controlled by the answer format. For closed exercises, the size of the candidate space is constant at the number of provided answers. However, exercises with a constant number of candidates can still vary in difficulty. The raw number of candidates is not informative enough because it does not account for the suitability of the candidates. We need to estimate which candidates are more likely to compete with the right solution. In the second section, we discuss approaches for measuring this phenomenon of *candidate ambiguity*.

On a more global level, the deletion rate has an influence on the exercise difficulty. A lower deletion rate produces more gaps which results in a higher degree of redundancy reduction. The C-tests and X-tests are designed with a deletion rate of two, thus every second word is transformed into a gap. This increases the *item dependency* because the damaged context of a single gap can only be recreated by solving the surrounding gaps. In the third section, we perform an analysis of the dependencies between items and their predictability based on linguistic features.

The influence of the format factors might differ with respect to the target language of the exercise. In this chapter, all format analyses are compared for English, German and French.

Language	Words	Avg. word length
English (American)	99,171	$8.5 \pm 2.6$
French	139,719	$9.6 \pm 2.6$
German	332,263	$12.0 \pm 3.5$

Table 6.1: Vocabulary size and average word length for different languages

# 6.1 The Candidate Space

The candidate space is a good indicator for the degree of redundancy reduction that is caused by different gap types. It can be determined by calculating the candidates for each word in the vocabulary and then averaging the results. We compare the candidate space for the three languages English, French and German and the two test types C-test, X-test and cloze test. The results have previously been published in Beinborn et al. (2015a).

As vocabulary for the three languages, the word list package provided by Ubuntu for spell-checking is used.<sup>78</sup> In principle, any word list could be used, but we aimed at having a common source for all three languages. It should be noted that the size of the lists varies depending on the morphological richness of the language; the German list is more than three times bigger than the English one (see table 6.1). The second language-specific difference is the word length which is considerably higher for German. This is mainly due to the existence of compounds that concatenate two or more words into one.

A word in the vocabulary is a candidate for a gapped word in a C-test if it starts with the same prefix and has the correct length. As the prefix length is dependent on the word length (equation 6.1), the second constraint is redundant. The C-test candidate space CS for a gapped word g is then determined by all words w in vocabulary V that have the same prefix. If  $g = x_1 x_2 ... x_n$ :

$$prefix(g) = x_1...x_{n+2}$$
(6.1)

$$CS_g = |\{ \forall w \in V \mid prefix(w) = prefix(g) \}|$$
 (6.2)

$$\overline{CS} = \frac{\sum\limits_{w_i \in V} CS_{w_i}}{|V|} \tag{6.3}$$

The last equation determines the average candidate space for vocabulary V. For X-test candidates, we need to compare the end of the word instead of the prefix. Table 6.2 shows the average number of English, French and German candidates for the three test types.

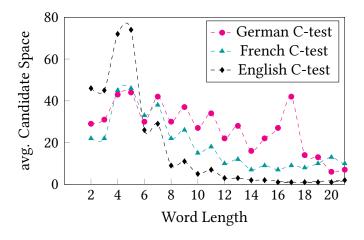
<sup>&</sup>lt;sup>78</sup>http://packages.ubuntu.com/de/precise/text/, languages wamerican, wngerman, wfrench, accessed: April 15, 2016

	Cloze	C-test	X-test
en	5	$20.2 \pm 29.0$	$102.5 \pm 219.4$
fr	5	$22.2 \pm 28.1$	$234.1 \pm 397.0$
de	5	$28.0 \pm 62.1$	$83.1 \pm 200.4$

**Table 6.2:** Average candidate space for different test types and different languages

It can be seen that the different size of vocabulary discussed above consequently leads to differences in the size of the average candidate space for the C-test. For German, the average number of candidates is thus higher than for French and English respectively. As expected, it can also be seen that the number of candidates for the X-test is higher than for the C-test as in the *coruscate* example above. However, it is interesting to observe that the candidate space for the X-test is not related to the vocabulary size. In this case, the candidate space for German is smaller than for the other two languages. This indicates that German word endings are more diverse.

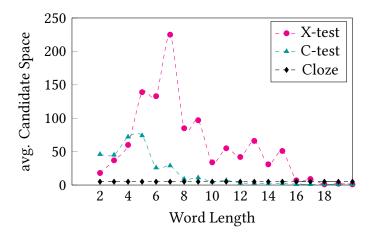
For all averages, the standard deviation is extremely high which shows that the candidate space varies strongly for the different words in the vocabulary. One important factor for this variation is the length of the word. If we remove one letter from the *coruscate* example and look at the C-test gap *cor*\_\_ instead, the candidate space increases from 1 to 36 candidates. We examine the influence of the word length in more detail, and calculate the average candidate space for all words with the same length. The vocabulary *V* in equation 6.3 is thus reduced to the words of the examined length.



**Figure 6.2:** Average candidate space for C-test gaps for different languages with respect to word length

**Candidate space for different languages** Figure 6.2 visualizes the average candidate space for the different languages with respect to word length. It can be seen that for English

the candidate space is maximized for extremely short words and decreases rapidly with increased word length. In comparison, the French and the German candidate spaces are more balanced: the number of candidates is lower for short words, but higher and more constant for longer words. This reflects the differences in average word length (see table 6.1) which are due to a more complex morphology in French and German.

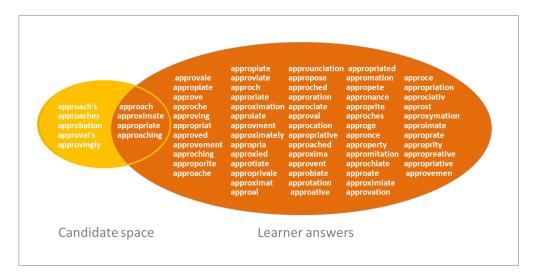


**Figure 6.3:** Average number of English candidates for different gap types with respect to word length

Candidate space for different gap types We also calculated the candidate space for the different gap types. Figure 6.3 visualizes the candidate space for the different gap types for English. The differences between the exercise types are similar for French and German. The number of candidates for our cloze test with five candidates (the solution and four distractors) is by definition always five. For open cloze tests, the computational candidate space for each gap would be equal to the size of the vocabulary. Compared to the C-test, the candidate space for the X-test is extremely large, in particular for words with medium length (five to nine characters).

Previous work has shown that the X-test is considerably more difficult than the C-test. From the results above it is tempting to conclude that this difference is due to the increased candidate space. However, following this hypothesis, the cloze tests should be fairly easy given the consistently small candidate space. The obtained error rates and the feedback of our test participants do not support this assumption. This gives rise to the idea that the candidate space considered by the learner differs from the computational one.

Candidate evaluation by learners When solving open formats, the learners cannot consider the full candidate space; only the words that are in the active vocabulary of the learner are accessible. In addition, the context can lead to priming effects and the test situation might alter the stress level of the participant and apply further restrictions.



**Figure 6.4:** The candidates for the gap *appro*\_\_\_: The computational candidate space is on the left, the answers provided by the students are on the right.

From the above arguments, one would expect that the learner's candidate space is smaller than the objective candidate space. However, we need to take into account that learners also consider wrong options, as in the example for the gap *appro* in Figure 6.4. The computational candidate space on the left consists of only 9 candidates, but the participants provided 68 different answers along with the solution appropriate (and only four of them intersect with the candidate space). It can be assumed that this does not even cover the complete candidate space for learners, as candidates that have been ruled out as incorrect by the learners are not accessible for us. The individual candidate space that is explored by a single learner is probably smaller than all the provided learner answers. Nevertheless, this example highlights the importance of modeling productive difficulties for test types with open answer format as discussed in chapter 5. In addition to the spelling variants, many wrong answers violate the length principle: 50% of the wrong answers in the English C-tests are actually longer or shorter than allowed. Weaker students tend to provide more false length answers (Pearson's r between overall score and number of false length answers: -0.67). However, the relative number of false length answers with respect to all wrong answers is higher for stronger students (Pearson's r between overall score and proportion of false length answers relative to all wrong answers: 0.3). Three main reasons for answers that violate the length principle can be identified. In many cases, the answer seems viable for the context and more easily accessible to the student than the solution.

(1) This global skill set ensures a bright future and an inter\_\_ and more enriching life path.

In addition to the solution interesting(29), the student also provided the answers international (48) and intercultural (11) quite often. Both seem contextually okay, and are semantically primed because they also occur at other positions in the paragraph. However, they

are both too long by one letter. A second reason is related to the morphological inflection; the student finds the correct lemma but fails to add the correct ending, i.e. an adjective is provided instead of an adverb, a singular is provided instead of a plural or an adjective is provided in the base form instead of the comparative form.

In this example, the students provide the answer hard (72) more often than the correct solution harder (48), although it is too short. In the third case, the student does not know the answer and provides just anything that remotely fits in order to not leave the gap blank.<sup>79</sup>

Bresnihan and Ray (1992) show that students perform better on the C-test if the length of the solution is graphically indicated by dashes or dots. This supports the assumption, that a violation of the length principle of a provided solution is often not noticed in the standard version of the C-test. As counting characters is a cognitive task that is not related to language proficiency, some test designers decide to relax the length principle and also accept longer or shorter answers if they are viable solutions. This could be modeled by adding a length tolerance to the calculation of the candidate space. However, adding a length tolerance would inflate the number of candidates that need to be evaluated.

For the closed cloze test, the candidate space is obviously constant, as it is determined by the number of provided answer options. The learners seem to consider even fewer candidates, on average only three of the five provided answers are actually selected. However, even the advanced learners commented that they experienced the task as very difficult. This indicates that the raw number of candidates is not a sufficient indicator for difficulty. It is also necessary to evaluate how well the candidate fits the gap and whether it competes with the correct solution.

<sup>&</sup>lt;sup>79</sup>We assume, that the student might be aware of the length violation in the third case, while it is not noticed in the other two cases. However, we have no evidence for that in the data and cannot distinguish between them.

# 6.2 Candidate Ambiguity

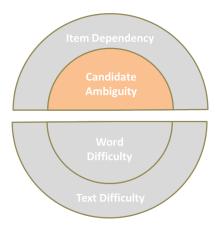


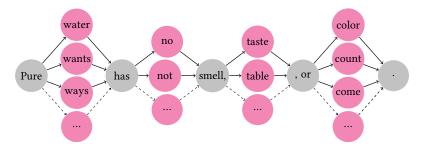
Figure 6.5: Candidate ambiguity: micro-level dimension of the exercise format

The main challenge for solving a text-completion exercise consists in identifying the most suitable candidate in the candidate space. Learners combine micro-level and macro-level processing strategies for this task: they evaluate the candidate with respect to the direct neighbors but also consider the sentential context (compare section 1.2). The suitability of the candidate with respect to the context is often called the *fitness* of the candidate (Sinha and Mihalcea, 2009).

Evaluating the probability that a word occurs in a certain phrase is a standard task in natural language processing (e.g. machine translation, speech recognition). For these tasks, human performance is usually considered to be the gold standard for the automatic approaches. The models are tuned and adjusted to reach human-like results. In learning settings, however, the human performance is flawed because of limited knowledge and lack of experience. In this section, a reverse approach is attempted: strategies from automatic solving are applied to model human difficulties. This approach is based on the assumption that exercises for which the automatic approaches fail indicate a deviation from the standard case. It can be assumed that specific knowledge is required to solve these exercises and that they will also be challenging for language learners who have limited knowledge.

## 6.2.1 Evaluating Micro-Level Ambiguity

Our computational evaluation of candidates in the micro-level context is mainly based on frequency effects, comparable to the evaluation of the context specificity in section 5.1.3. In chapter 5, only the correct solution of a gap was evaluated. To compare the fitness of the solution with the fitness of other candidates, it is necessary to compare the frequency effect for all candidates and create a ranking. The candidates are ranked according to their unigram, bigram and trigram frequency. The number of candidates above a threshold is



**Figure 6.6:** The search space for the sentence  $Pure\ wa\_$  has  $n\_$  smell,  $ta\_$ , or  $co\_$ . In this simplified example, the solution is always the topmost candidate and the displayed candidate space is reduced.

then taken as an indicator for the size of the candidate space for this gap.<sup>80</sup> A higher candidate space is expected to indicate higher difficulty. The rank of the solution among the candidates indicates the *candidate fitness* of the solution and the level of ambiguity between the candidates. If the solution is ranked lower than the competing candidates, the gap is expected to be more difficult to solve.

Another aspect leading to higher candidate ambiguity can be a high string similarity between a candidate and the solution. For example, many participants confused *base* and *basis* in the English C-tests. We calculate the maximum string similarity between the solution and each candidate based on the *longest common subsequence ratio* (Melamed, 1999) to account for this aspect of candidate ambiguity.

The approaches for evaluating micro-level ambiguity are integrated into the model for exercise difficulty. Their influence on difficulty prediction is evaluated in chapter 7.

## 6.2.2 Evaluating Macro-Level Ambiguity

For the evaluation of candidate fitness with respect to the macro-level context, we need to take the context into account. Consider the following cloze exercise:

(1) The stage lost a fine \_\_\_\_\_, even as science lost an acute reasoner, when he became a specialist in crime.[actor, estate, hunter, linguist, horseman]

The solution *actor* for the gap in the example is primed by the context *stage*. For the automatic evaluation of the context fitness of a word, two main approaches can be distinguished: language models and semantic relatedness. Both are usually applied to automatically solve tasks that require sentence completion. For the difficulty prediction scenario, we test whether wrong completion choices of the automatic approaches could predict solving

<sup>&</sup>lt;sup>80</sup>For the experiments, the threshold is set to 0 for simplification, but it could be tuned to the learner group. A threshold of 0 leads to longer processing times which might be inappropriate in real-life settings.

difficulties of humans. In this section, we focus on the solving methodology. The predictive power of the approaches for exercise difficulty is evaluated in chapter 7.

LM-based approach A probabilistic language model (LM) calculates the probability of a phrase based on the frequencies of lower order n-grams extracted from training data (Stolcke, 1994). This can be used to predict the fitness of a word for the sentential context. For example, Bickel et al. (2005) evaluate the use of probabilistic language models to support auto-completion of sentences in writing editors. In the completion scenario, only the left context is available, whereas the text-completion exercises also provide the right context of the gap. Zweig et al. (2012) thus model the problem of solving cloze tests as a lexical substitution task. The part to be substituted is a gap and the set of *substitution candidates* is already provided by the exercise candidates which only need to be ranked.

The language model fitness of candidate c for sentence s is determined by the logprobability of the language model for the sentence in which the gap is replaced with c. For solving a cloze exercise, the candidate with the maximum fitness is selected. For the half-open exercise formats C-test and X-test, modeling the task is more complicated because the sentential context is not static. It needs to be determined by solving the surrounding gaps. This can lead to strong dependencies between items, e.g. solving a subsequent item is facilitated if the previous one has been solved correctly. The full search space of possible sentences with all candidate combinations for each gap grows exponentially with the number of gaps in the sentence (see Figure 6.6). We thus use a pruning step after each gap that scores the generated sub-sentences using a language model and only keeps the *n* best. For our experiments, the threshold n is set as follows:  $n = min(10000, 10^{\#Gaps})$ . This threshold was determined by the assumption that each gap should add at most ten candidates to the combinatorial space. The maximum of 10,000 options is set to facilitate the computation for very long sentences. For most sentences, the maximum is not reached. The generated sentences are ranked by their log-probability in the language model and the highest-ranked sentence is chosen. Each gap is then solved with the candidate that occurs in this chosen sentence. If several sentences obtain the same probability, one is chosen at random. This strategy is repeated ten times and the results are averaged.

In the following experiment, we use 5-gram language models that are trained on monolingual news corpora using *berkeleylm* with *Kneser-Ney smoothing*. <sup>81</sup> Zweig et al. (2012) train a large n-gram model on 19th century novels from *Project Gutenberg*. They use texts from the same domain to reduce the perplexity of the model for the *Sherlock Holmes* sentences. However, as language learners are not very likely to have read 19th century literature, we train our language model on more general news sentences from the *Leipzig* collection (Quasthoff et al., 2006) to simulate learner knowledge.

<sup>&</sup>lt;sup>81</sup>http://code.google.com/p/berkeleylm, accessed: January 4, 2015

Semantic relatedness approach Language models cannot capture relations between distant words in the sentence. To account for this constraint, Zweig et al. (2012) include information from latent semantic analysis (Deerwester et al., 1990). For this method, every word is represented by a vector of related words that is calculated on the basis of training data. The *semantic relatedness* between two words can then be expressed by the cosine similarity of the two vectors.

Zweig et al. (2012) calculate the relatedness of a candidate for a sentence based on a latent semantic analysis index of the domain-specific corpus of 19th century novels. As argued above, learners of English probably do not have experience with texts from this era. We therefore approximate their general domain knowledge with an *explicit semantic analysis* index (Gabrilovich and Markovitch, 2007) calculated on *Wikipedia*. Similar to Zweig et al. (2012), the candidate fitness is determined by summing over the cosine similarity between the vector representation  $\vec{c}$  of the candidate c and the vector representation  $\vec{w}$  of every content word c in the sentence. The *semantic relatedness* (*semRel*) of candidate c with word c and the *semantic fitness* of candidate c for sentence c are calculated as follows:

$$semRel(c, w) = \frac{\vec{c} \cdot \vec{w}}{\|\vec{c}\| \|\vec{w}\|}$$
 (6.4)

$$semFitness(c, S) = \sum_{w_i \in contentWords(S)} semRel(c, w_i)$$
 (6.5)

A gap is solved by choosing the candidate with the maximum fitness. The semantic relatedness cannot be calculated for function words because they do not carry any semantic content. In addition, most semantic approaches can only be calculated for lemmas. Due to these two restrictions, the semantic relatedness approach cannot be used for solving C-tests and X-tests because it does not provide any information for choosing function words and inflections.

Zweig et al. (2012) combine the two approaches to maximize the solving ability for cloze tests. As our focus is not on reaching maximum accuracy but on modeling the solving abilities of human learners, we did not optimize our models for 19th century texts. We trained them on recent texts from the news domain instead. To the best of our knowledge, results for automatic solving of C-tests and X-tests have not yet been reported. In order to get an impression of the solving ability of the approaches above and to compare them to the human performance, we discuss the accuracy below. The results are compared to a baseline that always selects the most frequent candidate without considering the context and to the average human results. The accuracy of human solving is provided by the mean test score of all participants (see section 3.4.1) and should not be compared across test types

<sup>&</sup>lt;sup>82</sup>The index was retrieved from https://public.ukp.informatik.tu-darmstadt.de/baer/wp\_eng\_lem\_nc\_c.zip, accessed: November 25, 2015.

	Human	Baseline	LM-based	Semantic
C-test en	.68	.11	.76	-
C-test fr	.48	.10	.79	-
C-test de	.45	.09	.76	-
X-test de	.64	.09	.73	-
Cloze en	.70	.21	.28	.32

**Table 6.3:** Solving accuracy for the different candidate evaluation strategies

as the participant groups had different backgrounds and different language proficiency (see table 3.4). The results of the semantic approach are only provided for the cloze test because it cannot be calculated for function words and inflected words which are very frequent in C-tests and X-tests.

**Results** The accuracy of the automatic solving strategies and the average human performance in table 6.3 shows that the LM-based solving strategy strongly outperforms the baseline and can also beat the average human learner for the open test formats.

Even the large candidate space of the X-test can be disambiguated quite well. However, all results are significantly below the human upper bound of native speakers (which is close to 1.0 for all exercises). This leads to the assumption that failure of the automatic approaches could be a good indicator for difficult gaps.

For the cloze tests, the candidate ambiguity seems to be more challenging. The LM-based candidate evaluation only performs slightly better than the baseline due to the fact that the context frequency of the candidates has been controlled in this dataset. The semantic relatedness approach works better, but also fails to select the correct candidate in most cases. This is probably due to the specific domain of the cloze tests. *Sherlock Holmes* novels are fictional and often contain outdated vocabulary. The semantic relatedness calculated on *Wikipedia* does not seem to reflect this domain very well.

Not surprisingly, our results for the cloze tests are worse than those obtained with models trained on task-specific data in previous work.<sup>83</sup> However, the focus is not on developing a perfect solving method, but on modeling the difficulty for the learner. A question is less likely to be solved if the context fitness of a distractor is rated higher than that of the solution. Consequently, the solving approaches are transformed into features for difficulty prediction that return the rank of the solution among the candidates. An overview of the implemented features can be found in the appendix in section C.

<sup>&</sup>lt;sup>83</sup>The best approach by Zweig et al. (2012) reaches an accuracy of 0.52.

# 6.3 Item Dependencies

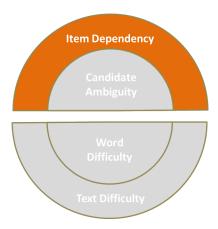


Figure 6.7: Item dependencies: macro-level dimension of the exercise format

The previous section has shown that the context plays an important role for the evaluation of the candidates. For many gaps, the direct micro-context can provide sufficient information for the solution, but in some cases the whole sentential context needs to be taken into account. In C-tests and X-tests, every second word is transformed into a gap and thus the context is restricted. In order to solve a gap, the participants need to restore the context by solving the surrounding gaps. This process can create dependencies between gaps as in the following example:

(1) Led by ambitious private companies, w\_\_ are entering the early stages of the migration of o\_\_ species away from Earth and o\_\_ adaption to entire new worlds. (we, our, our)

The first gap is ambiguous in the local context and can either be filled with we or who. If the participant makes the wrong choice, the correct solution our for the other two gaps is hard to determine. Many researchers acknowledge that processing on the sentence or the text level is more challenging than micro-level processing (Babaii and Ansary, 2001; Sigott, 2006). We therefore assume that gaps which are highly dependent on other gaps are more difficult to solve because they require macro-level processing. In this section, we analyze approaches for modeling and predicting item dependencies.

### **6.3.1 Determining Dependencies**

In previous work, dependencies between items have usually been examined based on the  $Q_3$  statistic (Yen, 1984). In this section, we follow the explanation by Krampen (2014) which provides a good overview. For the  $Q_3$  model, we first estimate a standard Rasch-model (Rasch, 1960) for the data to determine the person parameters (see section 3.4.2 for a more

detailed explanation). For each person<sub>v</sub>-item<sub>i</sub> combination, the difference e between the observed value X and the value expected by the model E(X) can then be calculated (equation 6.6). These differences are called *residuals*. In order to determine the dependencies between pairs of items i and j, the Pearson correlation r of the residuals of the two items is calculated (equation 6.7) which is called the  $Q_3$  model.

$$e_{vi} = X_{vi} - E(X_{vi}) (6.6)$$

$$Q_{3_{ij}} = r_{e_{vi}e_{vj}} (6.7)$$

In the case of *local item independence*, this correlation should be close to zero. The correlation is directional, i.e. a positive value indicates positive dependence between the two items and vice versa. Two items have been considered as dependent items in previous work if they exhibit an absolute correlation value higher than 0.2 (Krampen, 2014; Yen, 1984).

Assessing local dependence for items in C-tests has received a lot of interest from the psychological perspective. This is due to the fact that the standard *Rasch model* assumes local independence between the items. Intuitively, the C-test violates this assumption and thus requires more sophisticated models (i.e. testlet or copula models). Schroeders et al. (2014) analyze local dependencies for German C-tests but find that they have only little effect on the models (p. 414):

Due to the small amount of LID [ = local item dependence] in the observed data, it was not possible to demonstrate the pros and cons of the different models that take into account the inter-dependencies compared to the Rasch model.

Eckes and Baghaei (2015, p. 93) come to similar conclusions:

Therefore, it seems safe to conclude that the degree of local dependence present in the C-test studied here should not be a matter of much concern.

In both cases, the authors seem to be surprised by their findings. In contrast, Harsch and Hartig (2010) and Krampen (2014) find moderately high dependencies between English C-test items and also examine potential causes. They find that the highest dependencies between gaps can be found for neighboring gaps and for cases in which the same word is transformed into a gap twice. They raise the hypothesis that dependencies between items are often due to linguistic phenomena. For example, in idioms like knock on wood or semantically close words like skeletons and bones, it can be assumed that the second gap is easier to solve if the first gap has been solved correctly.

The opposing findings above could indicate that the phenomenon of item dependencies differs for English and German. In the following experiments, we examine item dependencies for the English, German and French C-test data and for the German X-test data and analyze whether it is possible to predict local dependence based on linguistic features. The

	Dependent Pairs	Positive r	Negative r		
Ctest_en	5.9	61.7	38.3		
Ctest_de	7.3	72.8	27.2		
Ctest_fr	21.9	59.9	40.1		
Xtest_de	14.9	76.9	23.1		

**Table 6.4:** Percentages of pairs of dependent gaps in training data and the percentages of negative and positive residual correlations in these dependent gap pairs

	Neighbors	Same Word
Ctest_en	24.7	7.1
Ctest_de	13.3	1.5
Ctest_fr	11.8	1.6
Xtest_de	16.7	2.0

**Table 6.5:** Percentages of neighboring gaps and word repetitions in the dependent pairs with positive residual correlation

previous analyses are all based on small datasets with very few texts compared to the larger datasets in this thesis.

Results Table 6.4 shows how the dependencies between gaps vary for the different exercise datasets. The French C-tests and the X-test trigger a higher ratio of dependent gaps. This indicates that the participants need to take more context into account to solve them. For all four datasets, the residual correlation of dependent gaps is more frequently positive than negative. Harsch and Hartig (2010) only analyzed positive correlations as item dependencies because they indicate that solving the first gap facilitates solving the second gap or that failing the first gap makes it harder to solve the second gap. We follow this interpretation and focus on the positively related gaps.

Harsch and Hartig (2010) claim that most dependencies can be explained by the fact that the two gaps are either neighboring gaps or refer to the same word. The results in table 6.5 show that these two phenomena explain only 30% of the dependencies for English and less than 20% for the other datasets. A first spot-check of the data for more sophisticated linguistic reasons for the dependencies does not reveal any obvious causalities. We run a second analysis to examine whether the detected dependencies remain stable across samples.

## 6.3.2 Dependencies Across Samples

In the German dataset, two texts have been answered by ten different samples of learners which allows us to compare the reliability of measures across learner samples (see sec-

Sample	1	2	3	4	5	6	7	8	9	10
Text 3	27	67	31	20	12	9	63	11	19	17
Text 8	15	52	10	16	15	11	30	15	15	8

**Table 6.6:** Number of dependent gap pairs (out of 380 pairs) for German C-tests across different samples

tion 3.4). For each learner sample, we determine the  $Q_3$  statistic for the 380 pairs of gaps in each text. The results in table 6.6 show that the number of dependent gap pairs varies strongly across samples. There does not exist a single pair of gaps that is classified as dependent in every sample. On the contrary, there are cases in which a pair occurs with negative residual correlation in one sample and with positive residual correlation in another. These results indicate that the measured dependencies between gaps are more dependent on the sample than on the underlying text. Section 3.4 showed that the error rates for the gaps in these texts remain quite stable across the samples. The samples are thus more homogeneous with respect to difficulty than with respect to gap dependencies.

## 6.3.3 Predicting Dependencies

According to the results above, we conclude that dependencies cannot be predicted based on linguistic features in our datasets, at least not according to the recommended  $Q_3$  measure. However, we have also seen that dependencies between gaps do occur quite often, in particular for French and German. As a consequence, we approximate the influence of potential dependencies on the difficulty using a few more shallow features.

A sentence that contains many gaps is usually more complex and a gap occurring towards the end of a sentence, is more likely to be influenced by incorrectly reconstructed context. The position of the gap, the number of gaps and the number of preceding gaps in the sentence and in the paragraph are thus calculated as features for difficulty prediction. In order to account for word repetitions, a boolean feature indicates whether the same solution also occurs in another gap. The *neighbor effect* can be approximated by the probability of the trigram spanning the previous gap and the current gap and the trigram spanning the current gap and the following gap.

For English, we additionally check for gaps with the prefix *th* because they enable many reference words such as *this*, *that*, *there*, *then*, *these*, *those*, *they*, and *their*. The student needs to perform co-reference resolution in order to select the correct word. These *referential gaps* usually cannot be solved on the basis of the micro context. For French and German, we did not identify a corresponding prefix with comparable referential ambiguity. An overview of the implemented features can be found in the appendix in section D.

# 6.4 Chapter Summary

In this chapter, we analyzed the format dimensions candidate ambiguity and item dependency. We computed the candidate space of different exercise types and compared it across three languages. From an analysis of the candidate space for language learners, we concluded that the absolute size of the candidate space is less relevant than the ambiguity of the candidates. We discussed measures for evaluating micro-level and macro-level ambiguity. The two macro-level measures have been adapted from existing sentence-completion approaches. The solving capabilities of the approaches are comparable to human learners. This observation leads to the assumption that the candidate evaluation strategies could be used to predict learner difficulties. This aspect is evaluated in the following chapter.

Previous work regarding item dependencies in text-completion exercises led to the assumption that dependencies are caused by linguistic features. We analyzed the dependencies in our datasets and showed that the common measure for item dependency does not produce stable results across samples. This indicates that a relation between item dependency and linguistic features cannot be confirmed because the linguistic features remain constant for each sample. As this finding contradicts intuitive assumptions, more research especially in the area of psychometrics is required. It is likely that item dependencies are more closely related to the sample than to the text. As a compromise, we model item dependency with rather shallow gap position features and account for neighbor effects. In the following chapter, the predictive power of format factors on the difficulty of items in text-completion exercises will be evaluated in more detail.

Chapter 6. The Influence of the Exercise Format

#### CHAPTER 7

# **Exercise Difficulty Prediction**

"It is a test of true theories not only to account for but to predict phenomena."

William Whewell

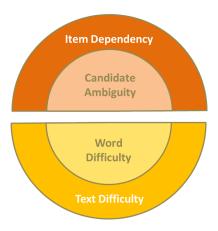


Figure 7.1: Exercise difficulty

In the previous chapters, the difficulty of text-completion exercises has been analyzed from several perspectives that are now combined in a four-dimensional difficulty model. This chapter contains experiments that evaluate whether the model can predict the difficulty of gaps in unseen exercises. The model is evaluated for five datasets covering different languages and different test types. The datasets are described in chapter 3. The experiments have previously been described in Beinborn et al. (2014b) and Beinborn et al. (2015a).

For a more detailed understanding of the difficulty of the task, we first present the results of a human annotation study. In the following sections, we analyze whether automatic methods can reach the predictive power of humans when assessing the difficulty of

135

unknown C-tests, X-tests, and cloze tests. We start with the C-test because it is the most frequent exercise type in the datasets.

# 7.1 Human Difficulty Prediction

Matching the difficulty of an exercise to an intended target group is a tedious and time-consuming task for teachers if they rely solely on their teaching experience. As they already know the correct solutions, they might not always be able to anticipate the confusion an exercise might cause for learners. In order to examine the human ability to predict the difficulty on the item-level, we conducted an annotation study using the English C-test training data. In section 4.2.3, we have seen that the teachers misjudged the mean difficulty of the tests and did not arrange them in an ascending order of difficulty as intended. However, following the discussion in chapters 4 and 5, the mean difficulty might just not be the appropriate concept because it neglects the variance between the items. We asked three teachers to assign a difficulty category to each gap in the training data of the English C-tests according to the following scheme:

```
    Very easy gap (error rate ≤ 0.25)
    Easy gap (0.25 < error rate ≤ 0.5)</li>
    Medium gap (0.5 < error rate ≤ 0.75)</li>
    Difficult gap (error rate > 0.75)
```

A gap is annotated correctly if the human-assigned class matches the actual error rate as defined in chapter 3.

**Participants** The participants are three English teachers from the language center at *TU Darmstadt* who are experienced test designers. All of them are female and have at least nine years of English teaching experience. Two of them are native speakers of English and one is German. The teachers were already familiar with the dataset because they had selected the texts for the student placement test.

**Results** The three annotators obtain comparable accuracy: each of them correctly predicts approximately 50% of the gaps (see Table 7.1). We find no obvious bias in the annotations, difficulty is both under- and overestimated. If we combine the human prediction by taking the median of the three annotators, 53.4% are annotated correctly.<sup>84</sup> These results show that even experienced teachers are not able to foresee all factors that influence the difficulty of a gap.

<sup>&</sup>lt;sup>84</sup>In case of a tie between two classes due to a missing annotation from one annotator, the higher class is assigned.

	A1	A2	A3	Median A1-A3
Correct prediction	200	209	192	213
Over-estimation	90	99	83	101
<b>Under-estimation</b>	107	89	118	84
NA	2	2	6	1
Accuracy	.50	.52	.48	.53
	A1, A2	A2, A3	A1, A3	All three
Fleiss' K	.31	.37	.39	.36

**Table 7.1:** Results of the human annotation

In order to analyze whether the mis-predictions can be explained by structural patterns, we additionally calculate the *inter-annotator agreement* of the human annotation based on the *Fleiss'*  $\kappa$ . This metric estimates the degree of agreement compared to pure chance agreement in case of random annotation. It is calculated using version 0.84 of the *irr*-package for R (Gamer et al., 2012). As the three annotators are quite homogeneous in experience, the low agreement in the last row of table 7.1 was not expected by them. The Fleiss'  $\kappa$  for the three annotators is 0.36, the pairwise comparison ranges from 0.31 to 0.39. A closer look at the data reveals that only for 38.6% of the gaps, all three annotators agreed with each other. For only 25.3% of the gaps, all three annotators agreed with each other and with the actual measured error rate. For comparison, we also evaluated the annotation as a binary task, i.e. we merged the easy classes 1 and 2 and the hard classes 3 and 4. In this case, the Fleiss'  $\kappa$  rises from 0.36 to 0.47. Each annotator annotates 77% of the gaps correctly when applying the more coarse-grained scheme. As this is still not very accurate, we have a closer look at the class distribution.

			Hun	nan N	Media	an
		Instances	1	2	3	4
Actual Class	1	172	118	48	6	0
	2	107	24	59	23	1
	3	77	7	34	29	6
	4	41	2	8	26	7

**Table 7.2:** Confusion matrix for human median prediction

Table 7.2 is a confusion matrix for the prediction when taking the median of the annotators. It shows that the annotators struggle in particular with predicting the most difficult gaps (class 3 and 4). In general, the mediocre human accuracy and the low inter-annotator agreement support the assumption that human difficulty prediction is quite subjective and

indicates that reliable objective prediction methods can prove helpful. In the following sections, we evaluate the performance of an automatic prediction approach.

Note that the automatic prediction is compared with the actual error rates, not with the classes annotated by the humans. Thus, it is theoretically possible to outperform human performance with automatic methods and provide a potentially useful support tool for test designers.

# 7.2 Predicting C-Test Difficulty

In this section, we present the results for gap difficulty prediction for English, French, and German C-tests.

## 7.2.1 Experimental Setup

In a first step, the most suitable parameters for the prediction experiments are determined on the training data (section 7.2.2 and section 7.2.3). As the datasets are rather small, the settings are evaluated with *leave-one-test-out cross-validation*. This means that a classifier is trained on all C-tests except one and the resulting model is then used to predict the difficulty of the gaps of the excluded test. This procedure is repeated for every test in the dataset. As already explained in chapter 4, the results of the prediction are only calculated once over the collected predictions of each fold as recommended by Forman and Scholz (2010).

Data We use the C-test datasets described in chapter 3. Table 3.1 and 3.4 provide the main statistics for the datasets. We first evaluate different task setups and perform feature analysis on the training data. The predictive power of the difficulty model is then evaluated on the held-out test data (section 7.2.4). To approximate the performance on bigger training sets, we finally conduct leave-one-test-out cross-validation on the combined set of training and test data.

Cross-validation In Beinborn et al. (2014b), the leave-one-out cross-validation experiments were performed on single gaps, i.e. the classifier was trained on all gaps except one and tested on the remaining gap. This setup was chosen to increase the training data, but it has an important weakness. As one text contains several gaps, there might occur an overfitting effect to the text properties. In addition, it is an unlikely scenario that a test designer wants to predict the difficulty of a single gap while already knowing the difficulty of the remaining gaps in the text. In Beinborn et al. (2015a), we thus switched to a more realistic evaluation setting with leave-one-out cross-validation on tests instead of gaps which is also applied for the experiments in this thesis.

Level	Dimension	Feature	C-test en	C-test de	C-test fr	X-test de	Cloze en
		AvgSentenceLength	✓	✓	✓	✓	✓
		AvgWordLengthInCharacters	✓	✓	1	✓	✓
	Text Difficulty	AvgWordLengthInSyllables	1	✓	✓	1	1
	Text Difficulty	ChunksPerSentence	1	✓	✓	1	1
0		VerbVariation	✓	✓	✓	✓	✓
$\check{\mathbb{A}}$		TypeTokenRatio	✓	✓	✓	✓	✓
MACRC		LeftTrigramLogProbability	✓	✓	✓	✓	
Σ		NumberOfGapsInCoverSentence	✓	✓	✓	✓	
		NrOfPrecedingGaps	✓	✓	✓	✓	
	Item Dependency	NrOf Preceding Gaps In Cover Sentence	✓	✓	✓	✓	
		OccursAsGap	✓	✓	1	✓	
		PositionOfGap	<b>√</b>	✓.	<b>✓</b>	<b>√</b>	1
		RightTrigramLogProbability	✓	✓	<b>√</b>	✓	
		COPCognate	✓	✓		✓	
		GapIsADJ	1	✓	✓	1	
		GapIsADV	✓	✓	✓	✓	
		GapIsART	✓	✓	✓	✓	
		GapIsCONJ	✓	✓	1	✓	
		GapIsNN	✓	✓	✓	✓	
		GapIsNP	<b>√</b>	✓.	<b>✓</b>	<b>√</b>	
		GapIsPP	<b>/</b>	✓.	✓.	<b>/</b>	
		GapIsPR	<b>/</b>	<b>√</b>	<b>√</b>	<b>/</b>	
		GapIsV	<i>,</i>	✓	✓	✓	
		IsAcademicWord	<i>,</i>	,			,
		IsCompound	<i>\'</i>	<b>√</b>	<b>✓</b>	<b>✓</b>	✓
		IsCompoundBreak	<b>√</b>	<b>√</b>	<b>√</b>	<b>√</b>	
		IsDerivedAdjective	<i>\</i>	<b>√</b>	<i>\</i>	<b>√</b>	
		IsFunctionWord	<i>\</i>	1	<b>√</b>	✓ ✓	
	Word Difficulty	IsInflectedAdjective	<b>V</b>	٠.	<b>√</b>	<b>√</b>	
		IsInflectedNoun IsInflectedVerb	<b>V</b>	1	<b>√</b>	<b>√</b>	
		IsLemma	./	<b>√</b>	/	/	
		IsReferentialGap	./	•	•	v	
		IsSyllableBreak	1	✓	1	/	
$\circ$		IsWordWithLatinRoot	1	•	·	•	
<u>K</u>		LMProbability	/	✓	✓	1	/
MICRO		LMProbabilityOfPrefix	1	/	/	-	-
$\mathbf{z}$		LMProbabilityOfPostfix				1	
		LMProbabilityOfSolution	1	✓	/	1	
		LeftBigramLogProbability	✓	✓	✓	1	1
		LengthOfSolutionInCharacters	✓	✓	✓	✓	✓
		LengthOfSolutionInSyllables	✓	✓	✓	✓	✓
		NrOfUbySenses	✓	✓		✓	1
		OccursAsText	✓	✓	✓	✓	
		PhoneticDifficulty	✓				
		PhoneticClarity	✓				
		PosProbability	✓	✓		✓	✓
		RightBigramLogProbability	✓	✓	1	✓	✓
		TrigramLogProbability	1	✓	✓	1	1
		UbyCognate	<b>√</b>	✓.	<b>✓</b>	<b>√</b>	
		UnigramLogProbability	✓	✓	✓	✓	<b>✓</b>
		BigramSolutionRank	✓	✓	✓	✓	✓
		EsaSolutionRank					✓
		LmSolutionRank	✓	✓	✓	✓	✓
		MaxStringSimWithCandidate	✓	✓	✓	✓	✓
	Candidate Ambiguity	NrOfBigramCandidates	✓	✓	✓	✓	
	Candidate Ambiguity	NrOfCandidates	✓	✓	✓	✓	
		NrOf Trigram Candidates	✓	✓	✓	✓	
		NrOfUnigramCandidates	✓.	✓.	✓	✓	
		TrigramSolutionRank	✓.	✓.	✓,	<b>√</b>	<b>√</b>
		UnigramSolutionRank	<b>√</b>	<b>√</b>	<b>√</b>	✓	<b>√</b>
		Number of Features	59	54	51	54	23

**Table 7.3:** Features for the difficulty prediction experiments

Features In Beinborn et al. (2014b) and Beinborn et al. (2015a), we used a large set of readability features (28) because they have been identified as important indicators for difficulty in previous work. As readability features are averaged over all elements in the text, the values of these features are constant for all the gaps of one test. Readability features are thus not very informative for the current task because we analyzed that the gap difficulty varies strongly also within texts (see chapter 4). In some cases, the readability features might even have a negative effect. In the error analysis of Beinborn et al. (2015a), we noticed that the classifier produced negative difficulty estimates for all gaps in a test. As the error rate cannot fall below zero in practical settings, this prediction result indicates that the classifier interprets the gaps as significantly easier than all gaps seen in training. This effect was provoked because the underlying text exhibited very high readability compared to the texts in the training data.

The experiments in chapter 4 have shown that the correlation of readability features with the mean difficulty is very low for all C-tests in the training set. We therefore only use a few standard readability features for the following experiments. For the other three dimensions, all described features that are applicable have been chosen.

Table 7.3 provides an overview of the features included in the difficulty model. Missing features for German and French are due to missing resources. Detailed descriptions of the features can be found in the appendix. The features are extracted using *DKPro TC* (Daxenberger et al., 2014) and the results are calculated using *support vector machines* in *Weka* (Hall et al., 2009).

## 7.2.2 Regression vs Classification

The difficulty of a gap can be expressed on a continuous scale (e.g. the error rate) or on a discrete *classification* scale as in the scheme for the human annotation study in section 7.1. We assumed that the difficulty classes would be easier to interpret for the teachers than the fine-grained numerical scale. However, as the actual error rates are continuously distributed, gaps that are close to the class boundaries are more likely to be mislabeled in this setting. Therefore, we also test *regression* prediction using the actual error rates instead of the class labels. As we examine a linear relationship using the same scale, we use *Pearson's r* between the error rates and the predicted difficulty as evaluation metric. *Spearman's r* models monotonic relationships which would be less accurate for this task, but the differences are negligible.

In order to get a better impression of the usefulness of the prediction, we additionally calculate the *quadratic weighted*  $\kappa$  which considers the classes on an interval scale. If the actual class is 2, a prediction of class 3 or 1 is considered to be slightly better than a prediction of class 4. We use the same implementation of the measure as in Zesch et al. (2015). To compare the performance for the classification and regression task, we also calculate the

			Clas	sificat	tion		Regress	ion
		P	R	$F_1$	q.w. $\kappa$	r	RMSE	q.w. $\kappa$
	Majority/mean baseline	.19	.43	.26	.00	.00	.25	.00
	Sigott baseline	.25	.40	.29	.00	.12	.26	.11
en	Frequency baseline	.19	.43	.26	.00	.39	.24	.28
	Difficulty model	.43	.46	.44	.36	.50	.23	.44
	Human median	.56	.53	.54	.59	_	_	.59
	Majority/mean baseline	.10	.29	.14	.00	.00	.26	.00
de	Sigott baseline	.42	.36	.33	.20	.39	.24	.22
ue	Frequency baseline	.31	.38	.34	.45	.54	.22	.43
	Difficulty model	.43	.43	.43	.49	.60	.21	.52
	Majority/mean baseline	.09	.30	.14	.00	.00	.27	.00
c.,	Sigott baseline	.34	.43	.37	.45	.53	.23	.43
fr	Frequency baseline	.35	.44	.38	.55	.59	.22	.51
	Difficulty model	.45	.45	.45	.56	.67	.21	.60

**Table 7.4:** Results for leave-one-test-out cross-validation on the training set for regression and classification prediction (both trained with support vector machines). Classification results are the weighted average of precision (P), recall (R) and  $F_1$ -measure over all four classes. Regression results are the Pearson correlation (r) and the root mean squared error (RMSE). The quadratic weighted K is calculated for both modes to enable comparison.

quadratic weighted  $\kappa$  for the regression predictions by mapping the numerical predictions back into classes according to the scheme explained in the previous section.

**Results** Table 7.4 shows the results for leave-one-test-out cross-validation on the training data in the classification and the regression setup. The  $\kappa$  values enable the comparison of the regression and the classification approach. It can be seen that our model performs slightly better in the regression setup. In chapter 3 we have analyzed that the error rates are continuously distributed. We will therefore focus on regression prediction for the remainder of the thesis. The decision to model the problem as a regression task is probably not the last word on the subject. If bigger datasets become available, more fine-tuning of the learning parameters can probably lead to improved performance and more informed decisions. The choice between classification and regression also depends on the intended use case of the developed model (see chapter 8).

In table 7.4, we compare our model against the human performance for English and three baselines. A naïve one that predicts the majority class for classification and the mean value for regression, a frequency baseline that only considers the unigram probability of the solution, and the Sigott-baseline which uses the features proposed by Sigott (1995): the frequency of the solution, the word class of the solution, and the constituent type of the phrase containing the gap. Sigott (1995) determines the frequency of the solution word

using counts from the *SUSANNE* corpus, <sup>85</sup> but in our experiments the more representative *Web1T* corpus is used for all frequency calculations (Brants and Franz, 2006). It can be seen that the majority/mean baseline yields bad results for all three languages. This can be explained with the observation that the difficulty is almost continuously distributed. The majority/mean baseline always returns the same difficulty value for every gap, which has a negative effect on the quadratic weighted  $\kappa$ . In the following, the results for each language are described in more detail.

**English** For English, the difficulty model clearly outperforms all baselines. The quadratic weighted  $\kappa$  indicates that the results for the Sigott-baseline are not useful. The frequency baseline performs slightly better with a quadratic weighted  $\kappa$  of 0.28. The full model yields a Pearson correlation of 0.50 between predicted and true error rates and a  $\kappa$  of 0.44. This can be seen as a useful approximation of difficulty, but the averaged prediction of the human experts is more accurate.

**German** For German, the baselines yield more reasonable results. In particular, the frequency baseline is quite strong with a Pearson correlation of 0.54 which is better than the best result for English. The prediction performance of our difficulty model outperforms the baselines considerably with a Pearson correlation of 0.60 and a  $\kappa$  of 0.52.

**French** For French, we obtain by far the best results. Both the Sigott-baseline and the frequency baseline already provide a good approximation for the difficulty. The full difficulty model yields a Pearson correlation of 0.67 and a quadratic weighted  $\kappa$  of 0.60 which is even better than the human difficulty prediction for English.

**Discussion** In summary, our model outperforms the baselines for all three languages and the accuracy of the difficulty prediction is comparable to humans. We have seen that the prediction works better for the French and German dataset although more resources are available for English. These differences seem to be strongly connected to the relation between word frequency and item difficulty. For French and German, word frequency alone is already a good approximation for item difficulty. For English, this is not the case and there are at least two possible explanations for this.

First, the frequency effect might be concealed by candidate ambiguity. In section 6.1, we have seen that the candidate space for short English words is considerably larger than for short French and short German words. In section 5.1, we have discussed that high-frequency words are usually shorter in English. These two findings indicate that gaps with high-frequency solutions have a higher candidate ambiguity in English. Re-consider the referential gap th : many of the possible answers (e.g. this, that, there, then, these, those,

<sup>&</sup>lt;sup>85</sup>http://www.grsampson.net/RSue.html, accessed: December 7, 2015

	en	de	fr
all features	.50	.60	.67
only micro	.48	.59*	.68**
only macro	.36**	.36**	.32**
without text difficulty	.47**	.60	.68*
without word difficulty	.38**	.37**	.34**
without candidate ambiguity	.47**	.60	.66
without item dependency	.50	.61**	.67

**Table 7.5:** Regression results for different feature groups. Significant differences of Pearson's r compared to the result with all features are indicated with \*(p<0.05) and \*\*(p<0.01).

*they*, *their*) are highly frequent, but choosing the adequate solution for the context from the 62 available candidates will be difficult for the learner.

Second, the background knowledge of the learners and the learning experiences might be more heterogeneous than for the other two languages and therefore more difficult to model. English influence is ubiquitous in entertainment, research, and media. Learners of English are thus more likely to have encountered words from the target language than learners of French or German. They might have developed expert vocabulary in a certain domain of interest (e.g. computer technology), but still not be familiar with basic words from other domains. This effect is related to the topic bias which is discussed in more detail in chapter 8.

In the following, we perform a more detailed analysis of the automatic prediction approach to better understand the influence of the different difficulty factors. For clarity, we do not add the results for RMSE and quadratic weighted K to the tables. However, they can be inferred from the provided scatter plots of the predictions.

## 7.2.3 Feature Analysis

Having established a fixed evaluation setting, we analyze the performance of the different feature groups. The significance of the differences to the results with all features are calculated using the paired.r function in version 1.4.5 of the psych-package for R (Revelle, 2014).

**Processing Levels** The results in the upper part of table 7.5 show that the gap difficulty is mainly determined by the features representing micro-level processing. This is not surprising as these features are calculated for each gap, whereas many of the macro-level features are constant for all gaps in the paragraph. The predictive power on the micro-level of our approach is a strong improvement over a previous prediction approach by Klein-Braley (1984) that only attempted to predict paragraph difficulty (compare chapter 4). For English

Dimension		en	de	fr
Text Difficulty		AvgSentenceLength AvgWordLengthInSyllables ChunksPerSentence	AvgSentenceLength AvgWordLengthInCharacters ChunksPerSentence TypeTokenRatio	TypeTokenRatio
	Word Class	GapIsV	GapIsADJ GapIsADV IsFunctionWord	GapIsArt GapIsNP
Word Difficulty	Production	IsCompound NrOfSenses OccursAsText Phonetic Similarity	IsLemma COPCognate LMProbability NrOfSenses	IsCompound LMProbabilityOfPrefix LMProbabilityOfSolution OccursAsText
	Familiarity	UnigramLogProbability	UnigramLogProbability	LengthOfSolutionInSyllables UnigramLogProbability
	Context	LeftBigramLogProbability	TrigramLogProbability	LeftBigramLogProbability RightBigramLogProbability
Candidate Ambiguity		NrOfBigramCandidates NrOfCandidates TrigramSolutionRank UnigramSolutionRank	NrOfBigramCandidates	NrOfBigramCandidates NrOfCandidates NrOfUnigramCandidates BigramSolutionRank
Item dependency		OccursAsGap IsReferentialGap LeftTrigramLogProbability RightTrigramLogProbability	PositionOfGap	NrOfGapsInSentence RightTrigramLogProbability
		18	15	17

**Table 7.6:** Selected features based on the *WrapperSubsetEval*-evaluator with the *SMOreg*-classifier and *BestFirst*-search as implemented in *Weka* (Hall et al., 2009).

and French, the results for the prediction using only the micro-level features are even better than the ones obtained with the full model.

**Dimensions** The lower part of table 7.5 shows the results for *ablation tests* with the four feature dimensions. It can be seen that the word difficulty features are by far the most important ones. Leaving out the other dimensions has almost no effect on the result. This supports previous theoretical research claiming that the solution word itself and its micro context are most relevant for the solving processes (Sigott, 2006). However, the word difficulty dimension is also the largest dimension (see table 7.3). The differences in the results could thus simply be caused by the varying number of features. We thus elaborate on this issue and perform feature selection.

**Feature Selection** We perform feature selection on the training set using the *Wrapper-SubsetEval*-evaluator with the *SMOreg*-classifier and *BestFirst*-search as implemented in *Weka* (Hall et al., 2009). Table 7.6 lists the selected features for each language. It can be seen that although the number of features is reduced by more than 50%, all dimensions are represented in the selected features. This could be explained by the observation that some

		en	de	fr
	all features	.37	.64	.69
Train-Test	only micro	.39	.61**	.70
	selected features	.39	.61**	.72
	all features	.48	.62	.67
LOO-all	only micro	.48	.59**	.70**
	selected features	.50	.62	.71**

**Table 7.7:** Prediction results on the test set (Train-Test) and leave-one-test-out cross-validation on all C-tests (LOO-all) with all features, only micro-level features and selected features. Significant differences of Pearson's r compared to the result with all features are indicated with \*(p<0.05) and \*\*(p<0.01).

features within one dimension tend to be correlated. We conclude that the dimensions in our model represent the factors that have an influence on the difficulty of text-completion exercises quite well. It should be noted though that the number of selected micro-level features is approximately twice as high as the number of macro-level features. This supports the impression from the ablation tests that micro-level features and in particular the word difficulty dimension are the most important factor for difficulty. The following section examines how well the model built on the training set scales for prediction on the test set.

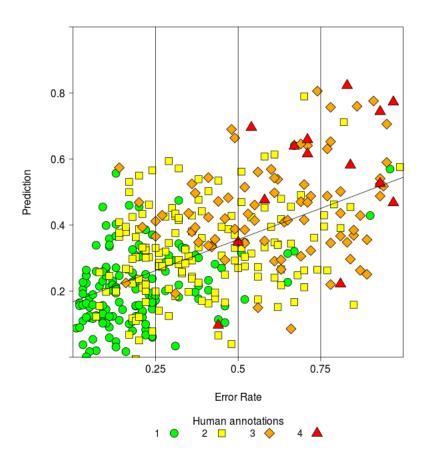
#### 7.2.4 Prediction Results

The upper part of table 7.7 shows the prediction results for the test set. In addition, the lower part reports results for performing leave-one-test-out cross-validation on the union of the training and the test set to analyze the performance of the model with a larger amount of training data. For French and German, the results are even better than the ones on the training dataset. It can also be seen that the prediction quality of the train-test setting reaches the same level as the leave-one-test-out cross-validation on the full set. This indicates that the model generalizes very well for these two languages. For English, the results are worse. In particular, the results of the train-test setting are very low. This indicates that the content of the training set might differ more strongly from the test set than in the other two languages. We will analyze this in more detail in the following section.

Interestingly, the differences between the different feature sets are only marginal and are not significant in most cases. The results show that satisfactory prediction results can also be obtained with a smaller subset of the features. For English and French, the results

<sup>&</sup>lt;sup>86</sup>It should be noted that in this setting, the training data on which we perform the feature selection is also included in the test setting. The results of the model with selected features for the leave-one-test-out validation on all tests should therefore be interpreted with caution. We only include them for completeness.

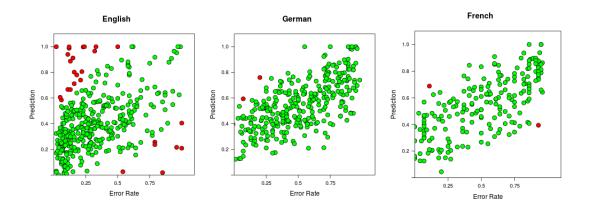
improve slightly when using only the selected features. With larger datasets, more fine-tuning of the features and their parameters would be possible, but in our case this would lead to over-fitting. In general, the results confirm the assumption that micro-level features and in particular word difficulty features are most predictive for gap difficulty. We discuss this observation in more detail in section 7.6 with respect to all datasets. As the prediction results for C-tests reveal some weaknesses in the approach, we take a closer look at the errors in the following section.



**Figure 7.2:** Prediction results for English C-tests using leave-one-test-out cross-validation on the training set. The symbols indicate the difficulty class that was annotated by the human experts.

## 7.2.5 Error Analysis

For a better understanding of the relation between human and automatic predictions, the plot in figure 7.2 combines both results for English. The position in the plot indicates the relation between the true error rate and the automatic prediction, and the symbol represents the corresponding human annotation. The vertical lines indicate the class boundaries. In



**Figure 7.3:** C-test prediction results for the train-test setting using all features. Instances with an absolute difference of predicted and actual error rate  $\geq 0.5$  are colored red.

the case of perfect annotation, only green circles would be found in the first quartile, yellow squares in the second, orange diamonds in the third, and red circles in the fourth quartile. In the case of perfect automatic prediction, the regression equation would have a slope of 1 and pass through the point of origin. The plot reveals that the regression equation predicts the right tendency, but tends to underestimate difficult gaps. The human annotators perform well for the easiest gaps (class 1, green circle), but also struggle with the prediction of the more difficult gaps.

As additional information, figure 7.3 visualizes the predictions for all three languages in the train-test setting. It can be seen that the better prediction quality for German and French is reflected in a small number of prediction outliers (red circles). In contrast, the low prediction quality for the English test set results in a high number of outliers. Figure 7.2 showed that under-estimation of hard gaps was a problem in the training set. For the prediction of the test set, over-estimation of the difficulty is more severe. The quality differences between the prediction for English and for the other two languages are not easy to explain. As an introspective speculation, we assume that the background knowledge of the participants and their learning experience are more heterogeneous for English. Unfortunately, the datasets do not contain enough information about the participants to verify this assumption. We have also seen that the results for English increase strongly for the leave-one-test-out cross-validation on all tests compared to the train-test setting. This indicates that the test set contains some unseen idiosyncratic properties that are smoothed out in the merged set. We have a closer look at the outliers in the train-test setting to understand the most frequent causes for wrong predictions.

**Over-estimation** Three structural causes for over-estimation of difficulty can be found: priming effects, named entities and hyphenated words. Priming effects occur in all three

languages when a seemingly difficult word is provided correctly by the participant because it is already mentioned beforehand in the text. This phenomenon is captured by the feature <code>OccursAsText</code>, but cases of priming occurred only seldom in the data. The other two causes for over-estimation are only found in the English data. For three gaps, the solution consists of a named entity (<code>Mahatma Gandhi</code>, <code>Nelson Mandela</code> and <code>Deutsche Bahn</code>) which is known to the participants, but does not occur in the vocabulary of the model. A feature for entity recognition exists, but the training data did not contain any named entities, so the model does not account for them. A similar problem occurs with hyphenated words (<code>colour-blind</code>, <code>non-violent</code>, <code>carbon-free</code>). It can be seen, that the teachers deviated from the original length constraint here and applied it only on the second part of the compound. As these kind of compounds have not been seen in the training set, our approach estimates the difficulty for the word <code>carbon-free</code>, while it should rather consider only the difficulty of the word <code>free</code>.

Compounds, named entities and priming effects could all be better integrated into our model. With a larger amount of training data containing these phenomena, the approach should then be able to adjust the difficulty predictions for these cases accordingly.

Under-estimation In the few cases where the model strongly under-estimates the difficulty of a gap, the high error rate is usually caused by local ambiguity that misleads the participants to choose the simpler alternative. For English, *the* is provided instead of *this*, *let* instead of *leave*, and for French, *ces* is provided instead of *car*. This preference should be captured by ranking the candidates according to their unigram probability, but apparently this ranking feature is not weighted strongly enough based on the training data. For the *leave/let* example, the provided solution violates the length constraint because it is too short. It is therefore not included in the computational candidate set. A solution could be to relax the length constraint and increase the candidate space for the model.

Two cases of under-estimation are caused by a common spelling difficulty that should be covered by our model. For a German gap, the participants provide  $da\underline{s}$  instead of  $da\underline{s}\underline{s}$ , and for English,  $o\underline{f}$  is provided instead of  $o\underline{f}\underline{f}$ . The solutions are basic vocabulary and are very short words, therefore the spelling ambiguity does not receive enough weight.

Interestingly, the analysis of the outliers revealed a few errors in the test design for English. The teachers had to encode the solution of the gaps manually in the web interface which led to some typing errors. In these cases, the error rate accumulates to 1.0, although many participants provided the correct solution. As the input to the prediction system contains the correct solution, the difficulty estimate is lower than the observed error rate. Our difficulty approach is the first step towards automatic design and selection of suitable C-tests as is further discussed in chapter 8. A higher amount of automation in the process will help to minimize the effect of human error and lead to cleaner data which can in turn then improve the prediction quality.

#### 7.2.6 Comparison to related work

The partial publication of our results in Beinborn et al. (2014b) encouraged further C-test experiments. Svetashova (2015) works with 20 English C-test paragraphs from placement tests at the *University of Tübingen* and conducts a large range of analyses that shed more light on the difficulty construct. In elaborate clustering experiments, she compares error rates and IRT parameters and also considers the difficulty of the text passage containing the item.

In contrast to our work, she incorporates the proficiency level of the participant directly into the difficulty measure by analyzing performance profiles (see section 3.4). Svetashova's analyses show that the items in her dataset should best be represented by four classes: easy items in easy texts, easy items in difficult texts, difficult items in easy texts and difficult items in difficult texts. This classification goal differs from our approach where the text containing the item is integrated as a dimension in the difficulty model and not interpreted as a prediction label.

For difficulty prediction, Svetashova approximates the item difficulty with a larger range of readability features than in our experiments (including sentence-level readability) and she also adds psycholinguistic word ratings from the MRC database (Gilhooly and Logie, 1980). Her results show that the readability classification of easy and difficult texts works well, but the distinction between easy and difficult items is harder. When using the same difficulty measure as in this thesis, she obtains results that are comparable to the ones in Beinborn et al. (2014b) for classification and slightly better for regression.

The findings by Svetashova and our analyses in chapter 4 indicate that a two-step classification which first identifies the macro-level difficulty of the C-test and then zooms in on the difficulty of the individual items might be a reasonable approach for future work. Another aspect of this idea is discussed in section 7.6.

# 7.3 Predicting X-test Difficulty

In this section, we evaluate the difficulty prediction approach for the German X-tests. X-tests are considerably more difficult than C-tests and can be used to assess more proficient learners up to native speakers.

## 7.3.1 Experimental Setup

The experimental setup for the X-tests is comparable to the C-tests. We perform leave-one-test-out cross-validation on the different tests and report the same evaluation metrics. The feature set is the same as for the German C-tests. However, a few adjustments are necessary because the test designers for the X-test explicitly encode several possible solutions in a list. The word difficulty features are calculated for the first element of the list as it is the preferred

	X-test de	Pearson's r
	mean baseline	.00**
	frequency baseline	.12**
	Sigott baseline	.18**
LOO-train	all features	.33
	only micro	.28
	only macro	.32
	without text difficulty	.28
	without word difficulty	.35
	without candidate ambiguity	.32
	without item dependency	.38
Train-Test	all features	.34
LOO-all	all features	.35

**Table 7.8:** Regression results for leave-one out cross-validation on the X-test training set with different feature groups. Significant differences of Pearson's r compared to the result with all features are indicated with \*(p<0.05) and \*\*(p<0.01). The lower part of the table contains the results for the train-test setting and for leave-one-test-out cross-validation on the union of training and test set.

solution. For the candidate features, the rank of the highest ranked solution is returned. In addition, some features need to be adjusted to account for the fact that the second part of the word is now provided as a hint.

#### 7.3.2 Prediction Results

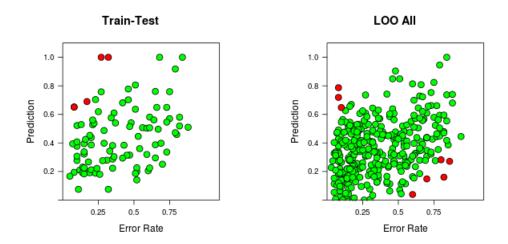
The prediction quality for the X-tests is significantly lower than for the C-tests. The results in table 7.8 show that our approach can beat all baselines, but the correlation is not as high as for the C-tests. In addition, the prediction varies only insignificantly for the different feature groups. This indicates that the learning effect of the different features is rather small. More detailed feature selection experiments are therefore not meaningful.

The results for the train-test setting and for the leave-one-test-out cross-validation on all tests are also on a surprisingly low level compared to the results for the C-test. In the following error analysis, we aim at a better understanding of these performance differences.

## 7.3.3 Error Analysis

The X-test dataset differs from the C-test datasets with respect to three characteristics: the size, the test type, and the participants' background.

Size The X-test dataset consists of only 14 tests (348 gaps) in total which is a very small number for any machine learning approach. Given this small size, the lower prediction quality is generally not surprising. The visualized predictions in figure 7.4 show that the



**Figure 7.4:** X-test prediction results for the train-test and the LOO-all setting using all features. Instances with an absolute difference of predicted and actual error rate  $\geq 0.5$  are colored red.

quality is not necessarily as bad as the numbers indicate. We observe only very few outliers and the general tendency of the prediction is correct, but the variation is too high. We see that with more data in the leave-one-test-out cross-validation on all tests, the spread of variations is slightly adjusted towards the middle. This leads to the assumption that the prediction quality will improve with a larger amount of training data. Bigger datasets would then also enable a more fine-grained feature analysis and parameter fine-tuning.

Test type For the X-test, the second part of the word is provided as hint. To date, there exist only very few research analyses for this test type (Cleary, 1988; Köberl and Sigott, 1994; Scholten-Akoun et al., 2014) and none of them take a computational perspective. The main line of argumentation for the increased difficulty of the X-test contrasts the focus on lexical-semantic elements in the X-test as opposed to morphological and syntactic elements in the C-test. In order to properly select a solution, the participants need to take more macro-level context into account. For C-test difficulty prediction, the strongest predictive power was gained from the micro-level features. A better approach for X-tests might be the evaluation of semantic relations on the macro-level. A first approximation to this phenomenon is the feature that ranks the candidates based on their semantic relatedness to the context (which was introduced in chapter 6.2). However, we did not apply it to the X-test because it is not defined for function words.

**Participants** The participants for the German X-test dataset are all native or near-native speakers of German. As a consequence, the learners should not have any problems with word familiarity in most cases. An analysis of the outliers indicates that the difficulty of compounds like *Lernstoff*, *Unterrichtssituation*, *Motivationsformen* and *Lehrkonzept* is

slightly over-estimated. Compounding is a creative process leading to an infinite number of potential words in German. The compounds described above have a very low frequency, however, the parts of the compound are rather frequent, particularly for students enrolled in education studies. Therefore, they can easily derive the compound from the context. The frequency effect which was quite strong for the C-tests is thus less relevant for this dataset. This supports the hypothesis described above that the semantic candidate ambiguity should be studied in more detail for the X-test format. In order to enable more elaborate experiments for the analysis of features and learner factors, larger datasets with X-test data would have to be created.

# 7.4 Predicting Cloze Test Difficulty

Cloze tests are the most popular type of text-completion exercises. As the open cloze test is too ambiguous to allow automatic scoring, most computer-assisted learning environments use a closed version of cloze exercises with multiple choice options for the answer. The format of the cloze exercises thus differs significantly from the C-test and the X-test. This section analyzes whether our prediction approach can also be applied to cloze tests.

## 7.4.1 Experimental Setup

The experimental setup for the difficulty prediction of cloze exercises is comparable to the previous experiments, but the feature set is smaller. Many of the features for C-tests and X-tests are not applicable for the cloze dataset. The cloze questions consist of individual sentences with a single gap, therefore dependencies between items cannot occur and all features related to item dependency are irrelevant. We also do not need to check for syllable and compound breaks because cloze gaps do not include a hint for the solution. The size of the candidate space is constant because the participants always see five answer options and the POS-tag for gaps is controlled, so we can omit the features related to these phenomena. We can also ignore all features related to production problems because the cloze test is a recognition task.

#### 7.4.2 Prediction Results

The cloze dataset is even smaller than the X-test dataset; only 100 gaps are available for training and again 100 for testing. The results in table 7.9 show that the small training size impedes a comprehensive feature analysis on the training set. The performance of different feature groups varies, but the differences are not significant due to the small dataset. Only the difficulty prediction based solely on macro-level features can be ruled out. This is in line with the findings for C-tests that micro-level features better capture the difficulty of items. For cloze tests, this effect seems to be even stronger. As the cloze exercises in our data

	Cloze-test en	Pearson's r
	mean baseline	.00**
	frequency baseline	.22**
LOO-train	all features	.41
	only micro	.36
	only macro	19**
	without text difficulty	.40
	without word difficulty	.19**
	without candidate ambiguity	.38
Train-Test	all features	.26
	only Micro	.26
LOO-all	all features	.37
	only Micro	.35

**Table 7.9:** Regression results for leave-one-test-out cross-validation on the cloze training set with different feature groups. Significant differences of Pearson's r compared to the result with all features in the corresponding setting are indicated with \*\*(p<0.01). The lower part of the table contains the results for the train-test setting and for leave-one-test-out cross-validation on the union of training and test set.

consist only of a single sentence, this is not surprising. However, previous work focused on readability as an indicator for cloze difficulty (Skory and Eskenazi, 2010; Klein-Braley, 1984).

The results for the leave-one-test-out cross-validation on the training set show the same tendencies as for the C-test. The micro-level features and in particular the word difficulty features are most predictive for cloze item difficulty. With a Pearson correlation of 0.41, the results are clearly worse than those obtained for the C-test.

The poor prediction results on the test set reveal that the model is over-fitted to the training data. When the size of the training data is doubled in the setup for leave-one-test-out cross-validation on the full set, the prediction results improve again but remain on the same low level as the results for the X-test.

It should be noted that the macro-level features perform even worse than in the previous experiments. This shows that the relation between readability and the difficulty of text-completion exercises seems to be less relevant for multiple-choice cloze exercises. The difficulty of an item depends less on the available context and more on the competition between the provided answer options. We have a closer look on the prediction quality and identify the prediction outliers in the following section.

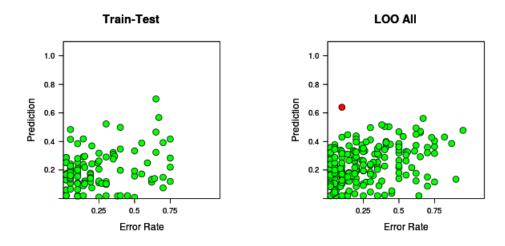


Figure 7.5: Cloze test prediction results for the train-test setting and leave-one-out cross-validation on all items using all features. Instances with an absolute difference of predicted and actual error rate  $\geq 0.5$  are colored red.

#### 7.4.3 Error Analysis

Figure 7.5 shows the relation between the actual error rate and the predictions based on the full model for the train-test setting and the leave-one-test-out cross validation on all instances (LOO All). It can be seen that strong outliers are not the problem. In general, the model consistently under-estimates the difficulty. In table 3.4, we have seen that the mean error rate in the cloze training data is 0.28 with a standard deviation of 0.23. In the test data, we find similar values: a mean error rate of 0.26 with a standard deviation of 0.22. This range of difficulty can explain why the model almost only predicts difficulty values below 0.6. However, in practical settings information about the extreme values on the scale (the easiest and the most difficult gaps) is usually most interesting because extreme items are not very discriminative and should be used with caution. In future work with larger datasets, it would be useful to focus more on the extremes of the difficulty scale.

Another aspect that could contribute to the consistent under-estimation of the model is the guessing factor. The cloze questions have five answer options, so each participant has a 0.2 probability to select the right answer simply by chance. As each gap has only been answered by 20–22 participants, the guessing factor might have a strong effect on the error rate for some gaps in the training data leading to error rates that are not a good representative of the difficulty. However, the extreme outliers in the prediction are also influenced by other factors. In the following examples, the knowledge of the learner is not modeled appropriately. The number of times a candidate was selected by the participants is indicated in brackets.

- (1) When I saw him that afternoon so enwrapped in the \_\_\_\_ at St. James's Hall, I felt that an evil time might be coming upon those whom he had set himself to hunt down.
  - [music (3), doorway (8), store-room (1), stream (0), shadows (9)]
- (2) For half an hour I sat with \_\_\_\_ ears. [straining (1), extended (4), widened (10), closed (5), bent (1)]
- (3) They meet, they write, there is a complete \_\_\_\_ between them.
  [understanding (8), problem (0), repose (0), explanation (1), conversation (12)]

The first example has an error rate of 0.9 because the participants favor the candidates shadows and doorway over the solution music. This can be explained by a lack of world knowledge: the participants probably do not know that St. James's Hall was a popular concert hall. The model does not properly capture this lack of knowledge and the semantic and the language-model based candidate ranker both rank the correct solution first. In example 2, the distractors widened, closed, and extended are all selected more often by the participants than the correct solution straining which is chosen only once. In example 3, the distractor conversation is selected more often than the solution understanding. In both cases, the participants do not seem to be familiar with the collocation that is evident for native speakers. These three examples highlight the importance of modeling language learners differently than native speakers as discussed in chapter 4.

In the following two examples, the participants' choices are spread over several candidates probably due to a lack of comprehension of the context.

- (4) He had trained it, probably by the use of the \_\_\_\_ which we saw, to return to him when summoned.

  [milk (6), sarcasm (2), shrubs (3), guilt (3), emotions (6)]
- (5) To do this was quite beyond my power, and I could only hope that by \_\_\_\_ I might in ten years' time save enough to enable me to put up my plate.

  [sunset (1), excavation (6), torture (2), economy (5), pondering (6)]

In the first example, the reference it cannot be resolved. If the participants had known that the antecedent is a snake, selecting the solution milk would probably have been easier. Reference resolution has not been an issue in the training data and milk is a very easy word, therefore the model underestimates the difficulty.

In the second example, the context actually provides cues for the solution *economy* such as *save enough* and *ten years' time* so that the candidate rankers rank it first. However, the phrase *to put up one's plate* is extremely rare and might have confused the participants.

We have seen that the readability features had not contributed significantly to the difficulty of the training data. These two examples indicate that the complexity of the sentence does have an effect on the difficulty of a cloze exercise. However, the choice of distractors can conceal the sentence complexity. If the distractors are very improbable, the participants can guess the correct solution even if they do not understand the meaning of the sentence. This highlights the importance of a combined evaluation of content and format factors for difficulty prediction as in our model.

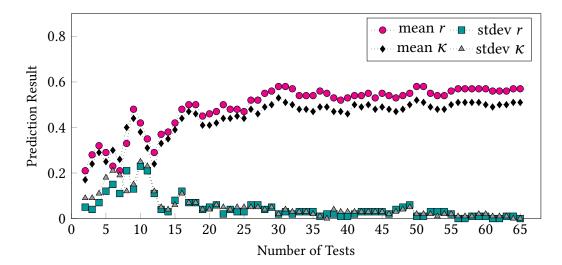
## 7.4.4 Comparison to related work

In section 2.3.4, we described a previous attempt to cloze difficulty prediction by Hoshino and Nakagawa (2008). Unfortunately, their original train and test datasets are not available anymore. Nevertheless, we attempt to compare our approach to their results. They used 208 cloze exercises from a TOEIC preparation book and extracted the 80 easiest (error rate smaller than 0.361) and the 80 hardest questions (error rate higher than 0.5) from it. The dataset that we received from Ayako Hoshino contains 702 questions. For each question, the solution is presented with three distractors. If we extract the easy and hard questions following their approach, we receive 213 hard and 213 easy questions. As the original subset cannot be determined anymore from the provided information, we use the full set. We run the same feature set as for the Sherlock Holmes dataset and perform the binary classification in a ten-fold cross-validation setting as Hoshino and Nakagawa (2008). Without further tuning or feature selection, our approach reaches an accuracy of 67.84% using support vector machines as implemented in Weka (Hall et al., 2009). The best classifier reported by Hoshino and Nakagawa (2008) reaches an accuracy of 60.63%. This shows that our approach outperforms the state-of-the-art approach on comparable data although their cloze questions differ significantly from the ones in our dataset. For the majority of the TOEIC questions, all four candidates have the same word stem and only differ in their grammatical categories. This aspect of grammatical candidate ambiguity has not been targeted by our candidate rankers, but it is implicitly captured by the language model approach. With a stronger focus on grammatical difficulty, even better results are probable for this dataset.

# 7.5 Learning Curve

The error analyses of the different prediction tasks indicate that the size of the training set plays an important role for the prediction quality. In order to verify this argument, we calculate a learning curve. We use the training data of the German C-tests because it has the largest size and is thus most suitable for analyzing size effects.

**Experimental setup** Using the standard feature set described above, we perform leave-one-test-out cross-validation with increasing sizes of the training set. For every n from 2 to 66, a subset of n C-tests is picked randomly from the training data and we calculate the result for leave-one-test-out cross-validation. In order to account for outliers, we repeat



**Figure 7.6:** Learning curve for leave-one-test-out cross-validation with increasing size of training data for the German C-tests. The prediction results are reported as Pearson's r and quadratic weighted K and are averaged over three runs.

this process for three runs. In figure 7.6, the results for Pearson's r and quadratic weighted  $\kappa$  are averaged over the three runs for each n.

Results The results of the learning curve experiment show that with a training set of less than 20 tests, the standard deviations are quite high for the three runs. This indicates that the prediction results are not very stable for these sizes. Satisfactory prediction quality can only be reached with a training set of at least 30 tests. Only the French and the German C-tests fulfill this criterion for the half-open exercise types (see table 3.1). This explains the slightly inferior results for the other exercise types. In our experiments, we did not attempt to tune prediction parameters because it is only meaningful with datasets that are several magnitudes bigger than the ones that are currently available. This leads to the optimistic expectation that strong performance improvements are possible in the future based on the findings in this thesis.

The learning curve also shows that the Pearson correlation and the quadratic weighted  $\kappa$  are highly correlated. This finding and the visualization of the predictions in the regression plot strengthens the impression of high prediction quality that is not always guaranteed by high correlation values alone.

## 7.6 The Influence of the Exercise Text

The prediction experiments for all datasets have shown that the features related to the global properties of the underlying text (the *macro-level* features) are not very predictive of the item difficulty. This contradicts previous work and also the subjective intuition that

completing a complex text is more difficult than completing an easy text. However, we can find an obvious explanation for these findings.

All texts in our datasets have been carefully selected (and in some cases also modified) by a group of experts. The C-tests and the X-tests have been used in authentic test settings; thus the quality of the test had to be guaranteed. As a consequence, all texts in the datasets are of high quality and adequate difficulty and the variance of readability is rather low. If the learners had been confronted with a wider range of texts, readability effects would probably be more evident. Imagine, for example, completing a C-test that is based on a patent description or on a stock report. We assume that this would be an almost impossible challenge even for native speakers.

In the experiments, we have predicted the difficulty for already existing exercises. As a next step, we aim at generating new exercises that match the quality expectations of experienced test designers. To reach this goal, we need to carefully select adequate texts. As the readability factors are only insufficiently represented by the training data, we need to apply additional readability filters *before* applying our difficulty model. We will explain this process in more detail in the following chapter.

# 7.7 Chapter Summary

In this chapter, the computational realizations of the four difficulty dimensions have been combined into a single model to predict the item difficulty of text-completion exercises. The approach has been evaluated on five datasets covering three languages and three test types. The best results have been obtained for German and French C-tests. The prediction quality for X-tests and cloze tests is lower. Detailed error analyses and the calculation of a learning curve have shown that these performance differences are mainly due to the size of the datasets.

We have shown that automatic difficulty prediction is possible with satisfactory quality. Based on the findings in this thesis, it can be expected that automatic prediction approaches can reach even higher quality in the future if the parameters of the model are fine-tuned on bigger datasets.

#### CHAPTER 8

# **Difficulty Manipulation**

"Conquering any difficulty always gives one a secret joy, for it means pushing back a boundary-line and adding to one's liberty."

- Henri-Frédéric Amiel

Generating a text-completion exercise is supposed to be a simple, automatic procedure that transforms a text into an exercise. However, the responsible designers of placement tests report that the majority of exercises are not usable in practical settings because the difficulty is not appropriate. The main challenge lies in selecting a suitable text basis for the exercise. The educators need to analyze many text candidates and adjust them manually to fit their needs.

In the previous chapter, we have seen that the prediction approach presented in this thesis can predict the difficulty with acceptable accuracy. This chapter aims at using the difficulty prediction for further automation of exercise selection and manipulation. Following the difficulty model, we distinguish between the manipulation of content and the manipulation of format factors.

In a first step, we implement a web demo that automatically generates a test from a text and predicts its difficulty. It is comparable to a rough difficulty estimate by a teacher and can help in the manual content selection process. In a second step, we develop an automatic approach for content selection and evaluate it with human experts. In addition, we provide a synonym dataset for lexical substitution that is enriched with word difficulty information. For the format factors, we focus on the manipulation of distractors because the ambiguity between the solution and the distractors is the main source of difficulty for cloze exercises.

159

# 8.1 Manipulating Exercise Content

In chapter 7, we have seen that the difficulty of an exercise is mainly determined by the exercise content. Exercise difficulty can thus be manipulated by changing the content. Texts from educational resources guarantee high text quality and adequate topics, but are often perceived as artificial due to oversimplification. Authentic texts, on the other hand, are more likely to contain linguistic flaws, inadequate content or constructions that are too difficult for the learners. In section 2.2.1, we have seen that text simplification approaches are not yet very robust. Fortunately, textual material is available in abundance and most language learning scenarios are not limited to using one specific text. As a consequence, manipulating exercise content can be interpreted as a content selection task.

Although the difficulty of selecting suitable content is commonly acknowledged, there exist no guidelines for the quality assurance process. Klein-Braley (1997) admits that text selection cannot be a random process because the text needs to be self-contained. In personal communication, educators explained that they rely on common sense, intuition, and experience to find and adapt a suitable text. In addition, they adhere to the following rules of thumb:

- The exercise should contain a topic from a general domain that does not require any expert knowledge to avoid a domain bias.
- Texts that contain many named entities, numeric expressions, or subjective arguments are typically not usable because they require factual background knowledge.
- The difficulty of the exercise should be appropriate.

As our approach for difficulty prediction can serve as a fast diagnostic tool for the last condition, we implemented it as a web demo.

#### 8.1.1 Web Demo for Test Generation

The web demo is available at: http://spz-etest.ukp.informatik.tu-darmstadt.de:9000/ctest It displays a simple input window for text (see Figure 8.1). A generation function transforms the text into a C-test that can be directly used with students. This service has already been provided by previous tools,<sup>87</sup> but our demo offers an important additional feature: a second output mode provides a difficulty visualization for the gaps.

Each gap is presented with a label and the color of the label indicates the difficulty. For the color-coding, the prediction values are mapped into difficulty classes according to the scheme in section 7.1. Gaps that have a predicted difficulty smaller than 0.25 are visualized with a green label and gaps with a predicted difficulty between 0.25 and 0.5 with a yellow label. In chapter 7, we have seen that the occurrence of hard gaps is rather infrequent. In

<sup>&</sup>lt;sup>87</sup>http://lingofox.dw.com/index.php?url=c-test, accessed: December 2, 2015

# **Difficulty Prediction for C-tests**

This demo visualises our research on the difficulty prediction for C-tests. You can insert a text, generate a C-test, and predict the difficulty of each gap. Please note: if you change the language, predicting the difficulty will be slow for the first attempt because new models need to be loaded. Subsequent predictions should be faster.



Figure 8.1: Web demo for difficulty prediction

order to improve the usability of the tool, we merge the difficulty classes three and four and visualize all gaps with a predicted difficulty higher than 0.5 with a red label. The demo is currently realized only for the C-test because it is the most popular exercise type for the co-operating educational institutions (the language center at *TU Darmstadt* and the *TestDaF* institute). It could easily be extended to X-tests and cloze tests.

The web demo serves as a first direct tool for teachers that can already alleviate the content selection. The next step consists in further automation of the content selection process. The main advantage of automatic evaluation over human evaluation is the processing speed. Using the methods developed in this thesis, we can examine a large set of texts from a corpus and determine their suitability for language learning exercises. In the following, we describe a test selection pipeline for English C-tests that could easily be adapted to other languages and to the generation of other test types.

## 8.1.2 Selection Pipeline

The following pipeline can be applied to any text corpus. The corpus is pre-processed using the *Stanford* segmenter and named entity recognizer (Manning et al., 2014), and the *TreeTagger* POS-tagger (Schmid, 1999).

**Text reduction** As C-tests consist only of a single paragraph with three to five sentences, the texts in the corpus first need to be reduced to the required size. A very rigid approach that only selects the first paragraph of each article is chosen for that goal. If the first paragraph is too short, subsequent paragraphs are added until reaching the desired length. Paragraphs are not split to make sure that the text is coherent and comprehensible. All paragraphs from the middle or the end of the articles are ignored to avoid unresolvable backward references.

Quality check In a subsequent quality check, strict filters are applied on the selected paragraphs. Texts with named entities or capitalized words are ignored because the named entities might be unknown to the learner. Only one numeric expression is allowed for each text to avoid overly technical content. In addition, we also filter texts that contain first person pronouns and quotation marks because they tend to be more subjective. The length reduction and the filters are very restrictive to assure high quality tests. It is very likely that suitable texts are also discarded due to these checks. If quality is already assured by a more careful corpus selection or by a human post-editing step, the filters could be readjusted to be more lenient.

Difficulty Prediction After the corpus has been reduced to a subset of suitable texts, the remaining texts are transformed into C-tests. The difficulty of each gap is then predicted based on the full model for English. Depending on the final application, the difficulty of the results can be manipulated by filtering out tests with too many (or too few) difficult gaps. For the current pipeline, the tests are written in an *HTML* output format that visualizes the difficulty of the gaps with colors (as in the output of the web demo in figure 8.1). If a more fine-grained difficulty analysis is required, it is also possible to directly output the regression value. This could be useful if the language proficiency from the target learner group differs from the learner group in the learned model. For example, if the test designers anticipate higher proficiency, they could simply subtract a constant from the regression prediction to receive a more probable estimate.

#### 8.1.3 Evaluation of Content Selection

For a better understanding of the usability of the test selection process, human experts are asked to evaluate the output.

**Experimental setup** The generation pipeline employs several quality filters. However, the usability of the output depends to a large extent on the quality of the selected corpus. In order to reduce the topic bias, the text corpus should contain texts from a general domain. The most popular English corpora contain news texts reporting on current events. Unfortunately, news texts are time-dependent and often require background knowledge that

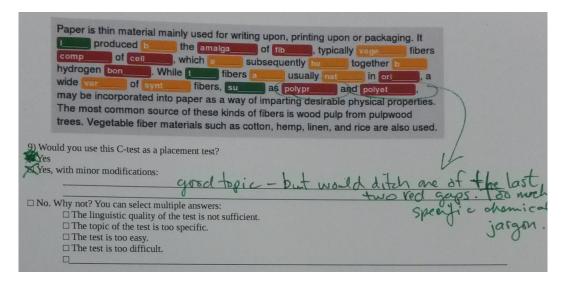


Figure 8.2: Example for human evaluation

language learners might not have. Instead, we use a sub-corpus of *Wikipedia* consisting of 2,615 articles that were selected by humans as being of high quality and importance.<sup>88</sup> From these articles, the pipeline extracts 28 texts to be of sufficient linguistic quality, generates corresponding C-tests, and predicts the difficulty. In the evaluation, we aim at determining the usability of the resulting C-tests for real scenarios.

**Participants** The participants are three English teachers from the language center at TU Darmstadt who are experienced test designers. All of them are female and have at least nine years of English teaching experience. Two of them are native speakers of English and one is German. The same teachers also participated in the annotation study in section 7.1.

**Evaluation setup** Each of the 28 C-tests is presented including the difficulty visualization (see figure 8.2). The teachers are informed that the difficulty estimates are predicted automatically and might not be accurate. For each C-test, the teachers are asked the following questions:

Would you use this C-test as a placement test?

- (1) Yes.
- (2) Yes, with minor modifications:
- (3) No.

If the teachers select *No*, they are asked to elaborate on the reasons.

Why not? You can select multiple answers:

<sup>&</sup>lt;sup>88</sup>Available from http://www.evanjones.ca/software/wikipedia2text.html, accessed: November 19, 2015

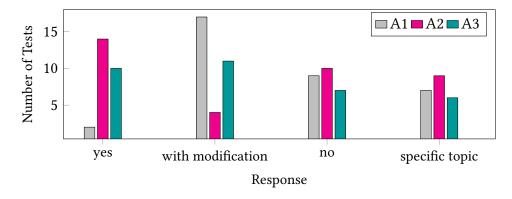


Figure 8.3: Results of the human evaluation for text selection

- (1) The linguistic quality of the test is not sufficient.
- (2) The topic of the test is too specific.
- (3) The test is too easy.
- (4) The test is too difficult.
- (5) \_\_\_\_\_

An additional question asked whether the rejected test could be used for another group of learners, but only one of the teachers actually answered this question. It is therefore omitted from the evaluation.

**Results** As can be seen from the plot in figure 8.3, the agreement for this task is very low. The *Fleiss'*  $\kappa$  is only 0.17 for the three annotators. However, the agreement for rejecting a test is quite high. If the two labels *Yes* and *Yes*, *with minor modifications* are merged, the Fleiss'  $\kappa$  increases to 0.41. We can see that the annotators approach the task with different granularity. A1 recommends to edit most C-tests, whereas A2 is willing to accept 14 tests directly. This shows that the test selection is a very subjective task even for experienced teachers.

In addition to the three labels for the decision, figure 8.3 also plots the number of texts for which the teachers selected the option *The topic of the test is too specific*. It can be seen, that this explanation accounts for almost all of the rejected tests. The tests were either considered to be too religious or philosophical (*Creationism*, *Free Will*) or too technical (*Power Plugs*, *Mass–Energy Equivalence*). As the placement test is taken by a large group of students with very heterogeneous backgrounds, religious topics might introduce a domain bias or might even be considered as offending. All three evaluators have a background in the humanities. This might cause a bias towards rating topics as being too technical. As the linguistic quality was never criticized, it seems that our strict quality filters are working well. For future applications, it is important to also account for topic preferences. This could either be realized by selecting a more suitable text corpus or by adding topic modeling to the quality filters.

Tavakoli et al. (2011) show that the learner's familiarity with the text genre and domain has an effect on the exercise difficulty. For self-directed learning, this indicates that the learner can cope with higher difficulty if the exercise is situated in her domain of expertise. The learners who contributed to the thesis datasets had very heterogeneous backgrounds, so it is not possible to account for individual preferences. We only approached a common trait of the group by modeling the cognateness of the words with German for the English and French datasets and with English for the German datasets because the majority of the learners know these languages. In order to adapt the difficulty model to a specific target group or to an individual learner, the difficulty model needs to be coupled with a learner model that accounts for individual differences such as domain knowledge and learning styles.

The evaluation further shows that only one of the unaccepted tests was rated as too difficult by A3 and two were rated as too easy by A2. The teachers aim at a wide range of difficulty because they want to account for many student levels. Section 4.2.2 described that the teachers aim at arranging five tests with ascending difficulty. Comments like *Maybe, as second text of 5* indicate that they directly try to classify the texts into one of the positions. The teachers were also encouraged to mark problematic gaps as can be seen in figure 8.2. These markings and the free comments indicate that the teachers actually considered the difficulty prediction as an important guideline and focused more on the gaps highlighted in red, see the examples below. The annotators were the same as those that struggled to estimate the gap difficulty correctly in section 7.1. This could explain why they are likely to rely on the prediction.

Teacher comments for modifications:

- Use as low-medium by taking out/replacing two of the red gaps
- Maybe. Use for hardest text but take out/replace a couple red gaps, probably too tricky overall.
- Take out red gap and use as mid-intermediate.
- Have no idea what "compu" is.
- Find synonym for compulsory.

**Overall evaluation of usability** After evaluating all 28 C-tests, the teachers were asked for an overall rating of the usefulness of the approach.

How did you experience the difficulty predictions?	ow did you experience th	ne difficulty predictions?	
--	--------------------------	----------------------------	--

- (1) They were helpful.
- (2) They were distracting.
- (3) They were not helpful, but also not distracting.
- (4)

Two teachers stated that the difficulty predictions were useful and one of them chose the open option and commented that *they were quite accurate*. The teachers were also asked if they would prefer more fine-grained information about the difficulty. All three agreed that the current format is sufficient. For the last three questions, we allowed free text answers. All teachers agreed that the system is useful and that not knowing the answer actually helped with the evaluation of the tests. They were positive about the approach and would like to use it in practice, see the comments below:

Do you think the automatic test generation is useful?

- Yes, will save loads of time guessing/estimating which texts are easier/harder + putting them in the right order.
- · Yes.
- Yes, as a basis for generating texts/tests.

In this evaluation scenario, you do not directly see the answers for the gaps. How does this affect your evaluation of the quality and the difficulty of the test?

- Not at all. I need to see if I can answer the red gaps.
- This is positive. It gives a more accurate view of authentic test conditions + difficulty.
- Better for me this way. Easier to assess the difficulty. I find it much easier to modify texts that are already gapped than to start from scratch with gapping so this program would be very useful.

Do you have any other comments?

- We would definitely use the programme for our placement tests, I think.
- I think this could be a very useful tool for the SPZ.<sup>89</sup>
- We generally stick to "general" scientific texts + avoid the humanities.

The overall rating was thus very positive. The last comment points to the problem with the selection of topics that had already been mentioned above.

One of the teachers complained that the gap dres was labeled as being easier than the gap vin in an article about sushi, although dressing has a lower frequency. However, it is a cognate and quite frequent in German, while the word vinegar does not have any corresponding German forms and might actually be more difficult. This example supports the assumption that the teachers cannot always anticipate all difficulty factors.

## 8.1.4 Manipulating Word Difficulty

The free comments in the teacher evaluation showed that most modifications of the generated C-tests would consist in substituting a word that is considered to be too difficult with a simpler synonym. This is a common task in natural language processing which is known as *lexical simplification*. In his overview of simplification approaches, Siddharthan (2014)

<sup>&</sup>lt;sup>89</sup>SPZ is an abbreviation for Sprachenzentrum meaning language center.

claims that "the ability to map words to meaning is the main difference between poor and skilled readers" and that the word choice is therefore the major source for text complexity. Most lexical simplification systems are build around the assumption that more frequent words are easier (De Belder and Moens, 2012). Unfortunately, the degree of simplification that is obtained by replacing a word with a synonym is difficult to measure. In several annotation studies, human evaluators were asked to rate the complexity of a word in context. The agreement values for this task are commonly rather low (Specia et al., 2012). In addition, the human raters are usually native or near-native speakers of the language with standard reading abilities. This conveys the implicit assumption that the typical target groups for lexical simplification such as children, dyslexic readers and non-native speakers all have the same simplification needs. In the chapters about text and word difficulty (4 and 5), we have seen that specific factors need to be taken into account for language learners.

Manipulating recognition difficulty Language learners have the benefit that they can rely on their knowledge of at least one additional language (their L1) when learning a foreign language. This has the effect that words which have *cognates* in the known language(s) facilitate the comprehension for the learner. This finding provides a new means for manipulating the difficulty of recognition tasks which has not been considered in previous work on simplification. In the lexical substitution dataset by De Belder and Moens (2012), for example, the word severely is annotated as being more difficult than the potential substitutes dramatically and critically. For German learners, the latter two words are cognates which facilitates the inference of meaning while severely has no corresponding cognate translation. Complex words can thus be substituted with a synonym that is a cognate for the learner to alleviate reading comprehension. In analogy, reading tasks can be complicated by using a higher ratio of words which are idiosyncratic in the target language so that their meaning cannot be guessed by the similarity to known words. This concentration on unfamiliar word stems can lead to a more focused increase of vocabulary. In order to highlight the differences between the languages that the learner already knows and the target language, the usage of false friends can be helpful. However, in this scenario it is important that the context explicitly marks the different meaning of the word so that the learner realizes the contrast.

Manipulating production difficulty The analysis of the datasets used in this thesis in chapter 5 has shown that although cognates facilitate *recognition*, they also lead to *production* errors due to two reasons. Cognates are more difficult to spell because they often have Latin roots that do not necessarily fit the regular spelling patterns of the target language. In addition, the learners might not explicitly know the word in the target language and erroneously apply the spelling of the cognate in their L1.

		Frequency		Cogn	ateness	Spellin	g Difficulty
speculation	venture	2,395,509	8,064,420	true	false	high	high
apathy	numbness	607,273	478,109	true	false	high	high
rival	competitor	3,495,097	2,746,376	true	false	low	high
attraction	magnet	4,490,790	2,454,624	true	true	high	low

**Table 8.1:** Lexical substitution examples

Synonym list In order to account for the phenomena mentioned above, we provide a resource that can be used for more targeted lexical substitution. We extract synonyms from *UBY* (Gurevych et al., 2012) and enrich each word with information about cognateness and spelling difficulty based on the automatic methods described in chapter 5. In order to support language recognition, words can then be substituted with synonyms that are cognates. Similarly, words can be substituted with synonyms that are easier to spell to facilitate language production. However, the substitutions should be based on word sense disambiguation to assure that the meaning is appropriate. Strict meaning preservation is not necessarily required in language learning but completely meaningless phrases should be avoided.

In the four examples in table 8.1, the first word is always a cognate in German. The first two examples come from different frequency bands. Speculation and venture are both very frequent, but it can be assumed that speculation is easier to understand for German students. Apathy and numbness are both infrequent and are probably not in the active vocabulary of beginning and intermediate learners, but the meaning of apathy can be derived due to the cognateness. In the last two examples, the focus is on the spelling difficulty. Rival and magnet are both easier to spell than their synonyms. If teachers are aiming to reduce the probability of spelling errors, substitutions can be performed according to this criterion. Especially the last example also highlights the importance of word sense disambiguation. Only a few contexts allow to substitute attraction with magnet, thus, lexical substitution should not be performed without word sense disambiguation.

#### 8.1.5 Conclusions

The evaluation of our approach for test generation provides a promising outlook. The inconsistent difficulty predictions by the teachers described in section 7.1 have shown a need for more objective difficulty estimates. The evaluation reveals that our tool can provide a solution that receives high acceptance and is considered useful by the participating practitioners. In order to use the automatic test generation directly without a human quality control, topic preferences need to be taken into account. For the modification of generated tests, the teachers often provided only small changes concerning the substitution of indi-

vidual words. We generated a synonym list that is enriched with cognateness and spelling difficulty information that can be used for learner-focused lexical substitution.

# 8.2 Manipulating Exercise Format

The difficulty of the same exercise content varies depending on the selected exercise format. In section 2.3.4, we discussed the influence of the deletion rate and the gap type on the difficulty. A smaller deletion rate means a higher number of gaps on less text which leads to higher item dependency and higher difficulty. The C-tests and the X-test are generated with the lowest possible deletion rate of two, i.e. every second word is transformed into a gap. The gap type can differ with respect to the portion of the word that is deleted (50% for C-tests and X-tests, 100% for cloze tests) and the position of the deletion (the last part of the word for C-tests and the first part of the word for X-tests). In general, higher deletion portions are more difficult and deleting the beginning of the word is more difficult than deleting the end (Köberl and Sigott, 1994).

The difficulty of exercises is also influenced by the available candidates. A lower number of possible candidates leads to lower difficulty because the probability of guessing the right solution is increased. For closed cloze exercises, the candidate space is fixed by the number of provided distractors. In this case, the most important factor for the difficulty is the ambiguity between the distractors and the solution. Distractors that can easily be ruled out facilitate an exercise, whereas tempting distractors confuse the learner and make the exercise more difficult (see section 7.4). Cloze difficulty can thus be manipulated by selecting distractors that increase (or decrease) the candidate ambiguity.

In order to analyze this claim, we focus on the cloze exercises with the highest and lowest difficulty. We aim at adjusting the difficulty of these examples towards the middle by finding substitution candidates for the distractors that alter the candidate ambiguity. Distractor selection is a complex task (see section 2.2.2) and a large-scale evaluation study is out of the scope of this thesis. We focus on a few selected examples as a proof of concept instead.

**Example selection** We extract the *easiest gaps* with an error rate  $\leq 0.1$  from the cloze dataset and only select those for which the two ranking features (the semantic score and the language model probability, see chapter 6) rank the solution on rank 1. For the resulting 14 questions, it is quite likely that the low difficulty was caused by the low *candidate ambiguity*. In addition, we extract the *hardest gaps* (error rate  $\geq 0.5$ ) and select the six gaps for which all candidates are ranked higher than the solution.

The ranking features evaluate the *context fitness* of the candidates based on a language model and on the semantic relatedness. If the context fitness of the solution is lower than that of the candidates, the difficulty of the gap is increased. In order to manipulate the diffi-

culty, we aim at increasing the candidate ambiguity for the easy questions and decreasing it for the hard questions.

## 8.2.1 Manipulating Distractors

The process of manipulating the distractors consists of two phases: generating substitution candidates for the existing distractors and evaluating their effect on the candidate ambiguity.

Generating substitution candidates The process of distractor selection is challenging because it needs to be assured that the distractor is not a proper solution for the gap. In addition, the cloze dataset was built with the requirement that all distractors should be grammatical in the context of the gap (Zweig and Burges, 2011). For the original dataset, these conditions had been assured by human post-editors who validated the distractors.

We aim at manipulating the candidate ambiguity without violating these constraints. We thus try to replace a distractor with words that are grammatically and semantically close to the human-validated original. This closeness can be approximated using a distributional thesaurus that groups words together which often occur within the same context. For each candidate of the selected questions, we extract 15 potential replacements that are distributionally close to the source from the *JoBimText* application (Biemann and Riedl, 2013). We use a pre-computed model based on dependency parses. As this model only contains lemmas, a pre-computed trigram-based model serves as a fall-back solution for finding substitutions if the original distractor is inflected. We then filter the obtained replacements and remove all words that are distributionally close to the solution. In addition, all replacements that are equal to one of the original distractors are also removed.

Evaluating substitution candidates Candidate ambiguity is evaluated using the two ranking features based on the semantic relatedness and the language model probability introduced in section 6.2. For the easy questions, the candidate fitness of the replacement candidate should be higher than that of the solution in order to increase the candidate ambiguity and thus increase the difficulty. As the experiments have shown that the semantic ranker tends to be a better estimate for the candidate ambiguity of the cloze exercises (see section 6.2), we select the substitution candidate that maximizes the semantic relatedness. If no candidate is ranked higher than the solution, the process is repeated with the language model score. Using this process, 37 substitutions for the 56 distractors of the 14 easy cloze exercises are identified. For 19 distractors, no suitable substitution could be selected. The following easy gap serves as an example for the proposed substitutions:

<sup>&</sup>lt;sup>90</sup>http://maggie.lt.informatik.tu-darmstadt.de:10080/jobim, accessed: December 7, 2015, used models: stanford, trigram

It was not until we had reached home that I began to \_\_\_\_\_ the true state of affairs. Original distractors: [realize, haunt, ravage, undermine, outrun]
Proposed substitutions: haunt/influence, ravage/cause, outrun/defeat

For the hard questions, the goal is to reduce the candidate ambiguity without using absurd solutions that are easily identifiable as incorrect. We apply the same process, but the inverse requirement for the ranking. Only nine substitutions could be identified for the 24 distractors of the six difficult exercises. It is particularly interesting that for three of the six exercises no substitution candidate could be selected. The process of generating substitution candidates terminated successfully, but all potential substitution candidates were ranked lower than the solution. This indicates that the context fitness of the solution is already very low and the cloze exercise might simply be difficult to solve independent of the distractors.

In order to assure that the distractor constraints are not violated by the substitutions, two native speakers were asked to proof-read the 17 exercises for which substitutions could be generated.

## 8.2.2 Quality Evaluation of Distractors

We present the solution, the original distractors and the generated distractors of the 17 cloze exercises as equal candidates to two native speakers of English. The difference between the candidates is not indicated. For each of the candidates, the evaluators A1 and A2 are asked to judge in two steps whether the candidate fits the gap i) grammatically and ii) semantically. The goal is that all candidates are grammatical, but only the solution to the cloze exercise should fit semantically. The annotators were instructed to apply strict judgments. They should judge the *grammaticality* of a candidate independent of its meaning. They should mark a candidate as *semantic fit* if they could easily construct the meaning of the sentence including the candidate. A candidate that is a semantic fit can thus be considered as a solution for the exercise.

In all cases, the annotators correctly identify the solution to fit grammatically and semantically. However, for the distractors the two annotators disagree considerably regarding grammaticality (Cohen's  $\kappa$ : 0.05) and semantic fit (Cohen's  $\kappa$ : 0.7)) as can be seen in the tables 8.2 and 8.3.

Grammaticality A1 marks all candidates (original and generated) as grammatical except for one generated substitution. A2 follows a stricter line and labels 33 candidates as ungrammatical including several of the original distractors. She seems to consider additional aspects such as selectional preferences for her judgment. As the original distractors had already been labeled as grammatical by other human annotators in the study by Zweig and Burges (2011), we only rely on the judgment by A1 for grammaticality. Thus, only one of

	Easy Questions			Hard Questions			
	Instances	A1	A2	Instances	A1	A2	
Solution	14	14	14	3	3	3	
Original Distractors	56	56	43	12	12	10	
Generated Substitutions	36	35	19	8	8	6	

**Table 8.2:** Evaluation of grammaticality: number of candidates that were labeled as grammatical by the annotators. In the ideal case, all candidates should be grammatical.

	Easy Questions			Hard Questions			
	Instances	A1	A2	Instances	A1	A2	
Solution	14	14	14	3	3	3	
Original Distractors	56	1	4	12	9	2	
Generated Substitutions	36	3	3	8	4	2	

**Table 8.3:** Evaluation of semantic fit: number of candidates that were labeled as semantic fit for the sentence. In the ideal case, only the solutions should be labeled as semantic fit.

the generated substitution candidates (*criticism*) can be considered strictly ungrammatical for the context (it should be the plural *criticisms*).

**Semantic fit** Even the original set of candidates seems to contain distractors that are solutions for the exercise. For the easy questions, A1 labeled one of the 56 original distractors as semantic fit and A2 four others. Three out of the 37 generated substitutions are labeled as semantic fit by both annotators and are therefore invalid distractors.

For the three hard exercises, both annotators label two out of twelve original distractors to fit semantically. As these two distractors belong to different exercises, at least two of the hard exercises are highly ambiguous even for native speakers. A1 perceives even higher ambiguity and labels nine of the original distractors to fit semantically. This explains the high error rates for these exercises. It also shows that the candidate rankers capture the right tendency by ranking the solution on rank 5.

As a consequence, the substitution candidates for ambiguous distractors are of course also likely to be ambiguous. The results for the three hard exercises are inconclusive. For the first exercise, all generated distractors are considered to fit semantically into the context by both annotators, for the second none of the generated distractors fit semantically and for the third exercise the annotators disagree.

Consequences The above results show that the substitution generation approach preserves the grammaticality of the candidates in almost all cases. Avoiding that the generated substitutions fit semantically is slightly more complicated, particularly if this constraint was

already violated by the original distractor. In future approaches, a more rigorous validity check using textual entailment approaches needs to be applied (Zesch and Melamud, 2014). In order to avoid ambiguity, the generated substitutions are slightly adjusted before presenting them to non-native speakers. The ungrammatical distractor *criticism* is changed into the grammatical plural *criticisms*. The current pipeline did not include a proper agreement check, but this could easily be added. For the easy exercises, three of the generated substitutions need to be removed because both annotators agree that they are correct answers.

- (1) He was a solicitor and was \_\_\_\_ my room as a temporary convenience until his new premises were ready.
- (2) The probability was, therefore, that she was \_\_\_\_ the truth, or, at least, a part of the truth.
- (3) They undoubtedly showed that the \_\_\_\_ was much deeper than was at first conjectured.
- (4) I stared at it \_\_\_\_, not knowing what was about to issue from it.

The substitution *holding* for the distractor *serving* is excluded for exercise 1, *learning* (for *surveying*) is excluded from exercise 2, and earth (for *moon*) from exercise 3. One of the hard exercises (exercise 4) is excluded completely because the original and the generated distractors are too ambiguous. The context for this exercise is not very restrictive and allows too many answers. These amendments result in 16 exercises with on average three substituted distractors.

## 8.2.3 Difficulty Evaluation of Distractors

The exercises with the substituted distractors are set up in a web survey similar to the ones in the previous cloze study described in section 3.3.1. The exercises were answered by 20 non-native speakers of English who had not taken part in the previous study. The majority of the participants were German and their proficiency level was comparable to the participants of the previous study.

Results The intended adjustment of the difficulty could not be confirmed by the study. Only for two of the fourteen easy exercises and for one of the two hard exercises, a change in the error rate could be observed. They are listed below, the first two are the easy exercises that have become more difficult and the last one is the hard exercise that has become easier. For the other exercises, the difference between the error rate of the original and the manipulated exercises was smaller than 0.2.

(1) The stage lost a fine \_\_\_\_\_, even as science lost an acute reasoner, when he became a specialist in crime.

Presented options: [scientist (5), hunter (0), actor (15), land (0), horseman (0)]

Performed substitutions: linguist/scientist, estate/land

Original error rate: 0.05, new error rate: 0.25

- (2) A few good \_\_\_\_ and the reputation which I had won in the hospital brought me rapidly to the front, and during the last few years I have made him a rich man. Presented options: [knives (1), criticisms (3), cases (13), dinners (2), landlords (0)] Performed substitutions: taunts/criticisms

  Original error rate: 0.1, new error rate: 0.35
- (3) That cold, \_\_\_\_, ironical voice could belong to but one man in all the world.

  Presented options: [relaxed (9), hungry (1), rented (0), disorganized (0), incisive (10)]

  Substitutions: serene/relaxed, self-contained/rented, inflexible/disorganized

  Original error rate: 0.82, new error rate: 0.5

It can be seen, that the increased error rate can be directly attributed to the substitution *scientist* in example (1) and *criticisms* in (2). In example (3), replacing the originally tempting distractor *self-contained* with *rented* caused the biggest effect. The substitution *relaxed* did not decrease the difficulty. For these three exercises, the difficulty manipulation worked well, but it did not have an effect on the other eleven exercises. We discuss potential reasons for this below.

**Discussion** We identify three possible explanations for the observation that only few difficulty changes could be observed in the study. First, the selection of substitution candidates with the distributional thesaurus leads to a very limited candidate set. If the original distractors were not well chosen, substitutions that occur in the same context are equally bad. The candidate rankers would probably perform better when selecting the substitution from a bigger set of candidates. However, in this setting, better quality assurance using textual entailment approaches as in Zesch and Melamud (2014) is required to eliminate substitution candidates that solve the gap correctly.

Second, our candidate rankers do not capture selectional preferences of verbs very well which can result in rather unsuitable distractors, e.g. the distractor hanging for the context I rushed out, \_\_\_\_ loudly for my stepfather is not distracting from the correct solution calling because it is not likely to occur with loudly nor with a for complement. Out of the eleven easy questions for which no manipulation effect was observed, eight use verbs as distractors. The selectional preferences of the distractors thus need to be more controlled. However, the more the distractor matches the context, the more likely it is an actual solu-

tion. Meeting this fine line to generate difficult cloze distractors is thus a very challenging task.

Third, the intended increase of difficulty might actually be realized, but the effect is not measurable. A similar argument has already been discussed in Zesch and Melamud (2014). In case the manipulated distractors do distract the learner more but the correct solution is selected anyway, the difference is not evident. In order to avoid this for future studies, the learners could be asked to indicate their certainty for each exercise. All participants in the study have been learning English for at least five years (60% for more than 10 years). The results thus might have also been caused by a floor effect,i.e. the exercises are generally too easy for the participants.<sup>91</sup>

## 8.3 Chapter Summary

In this chapter, the findings for difficulty prediction have been used for the goal of difficulty manipulation. A web demo that visualizes the output of test generation and difficulty prediction for any text has been implemented and is publicly available at http://spzetest.ukp.informatik.tu-darmstadt.de:9000/ctest. For difficulty manipulation of C-tests, a pipeline that selects texts from a corpus, generates tests, and predicts the difficulty has been developed. In an evaluation study, the output of the pipeline was rated by teachers. More than half of the generated tests were considered to be usable in real-world placement tests and the overall ratings were promising. For future improvements of the approach, topic selection and lexical substitution should also be considered. To this end, a resource consisting of synonyms enriched with cognateness and spelling difficulty annotation was created and made publicly available. 92 For difficulty manipulation of cloze tests, the selection and manipulation of distractors is crucial. We propose to substitute existing distractors to adjust the difficulty of extremely hard and extremely easy exercises. The substitutions are generated using a distributional thesaurus and their candidate fitness is evaluated using the candidate ambiguity rankers described in chapter 6. The grammaticality and the semantic fit of the substitutions were evaluated by two native speakers. The results show that grammaticality of the substitutions can mostly be preserved. Avoiding that the substitutions fit semantically is more complicated. The exercises were also tested with language learners, but the manipulation effect on the difficulty could only be shown for a few examples. The distractor manipulation could be improved by modeling selectional preferences of verbs and applying less restrictive candidate generation in future work.

<sup>91</sup>http://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-1698-3\_230, accessed: January 27, 2016

 $<sup>^{92}</sup>$ At http://www.ukp.tu-darmstadt.de/data/c-tests/difficulty-prediction-for-language-tests

## Chapter 8. Difficulty Manipulation

#### CHAPTER 9

## **Conclusions**

"Die Schwierigkeiten wachsen, je näher man dem Ziele kommt."

Johann Wolfgang von Goethe

The final chapter summarizes the findings of the thesis and provides an outlook to promising research directions that could arise from this work.

## 9.1 Summary

This thesis has shown that it is possible to predict the difficulty of text-completion exercises automatically with high accuracy reaching human performance levels. The presented approaches are already considered as useful by educational practitioners for designing language proficiency tests. With further engineering and fine-tuning, the developed technology can also be used to directly generate exercises for self-directed learning scenarios without human intervention. In the following, the main findings of each chapter are summarized and their potential implications for future work are pointed out.

Chapter 2 provided an overview of the state of the art in computer-assisted language learning and educational natural language processing with two main conclusions. First, text-completion exercises are a popular choice for language proficiency tests. They combine educational quality expectations with technological practicality. Second, an approach for measuring and adapting exercise difficulty automatically is required to take exercise generation to the next level.

Chapter 3 presented the new difficulty model that forms the theoretical basis for this thesis. The model interprets exercise difficulty as a combination of content and format factors and is further categorized into the four dimensions of text difficulty, word difficulty,

177

candidate ambiguity, and item dependency. The goal for the thesis was refined as the development of computational measures for each of the dimensions and their integration into an approach for difficulty prediction. For the empirical basis of the thesis, five difficulty datasets have been introduced. A data-driven comparison of two existing theories led to the decision for the error rate as indicator of difficulty.

Chapter 4 provided an analysis of the text difficulty dimension. A readability experiment has shown that the claims in previous work about the predictive power of existing readability features for the difficulty of text-completion exercises are not supported by our data. A psycholinguistic comparison of L1 and L2 readability indicates that the traditional view on readability fits the holistic comprehension and interpretation approaches by native speakers, but does not necessarily match the more conscious processes applied by language learners. As a consequence, computational approaches need to focus more on local difficulties. This is an important finding for future work in readability, and it has already been considered by Pilán et al. (2014).

Chapter 5 introduced measures for approximating the difficulty of words for language learners. We developed an approach for alphabet-agnostic cognate production to account for cross-lingual item transfer. It uses character-level machine translation to model morphological processes across languages. This approach has recently also been adopted by Scherrer and Sagot (2014), Scherrer and Erjavec (2015), and Ling et al. (2015) and is considered to be beneficial for machine translation. In order to successfully integrate the cognate approach into real-world applications, quality control and the accurate distinction between true and false friends need to receive more attention. To account for productive word difficulties, we developed an approach for measuring spelling difficulty that relies on new phonetic features. For the evaluation of the approach, spelling errors have been extracted from a corpus of learner essays. We made the datasets for the cognate and spelling experiments publicly available. In future work, the insights into word difficulty and the corresponding features can also contribute to related tasks such as spelling correction, lexical substitution, writing assistance, and natural language identification.

Chapter 6 focused on the format dimensions candidate ambiguity and item dependency. For the evaluation of candidate ambiguity, two measures for automatic solving of exercises were analyzed and adapted to model the difficulty for language learners. The idea of limiting the training data of computational models to simulate the limited knowledge of learners is a new twist that runs counter to the general assumption that bigger is better. An analysis of item dependencies across samples has shown that existing measures do not capture the

<sup>&</sup>lt;sup>93</sup>An overview page with all links can be found at http://www.ukp.tu-darmstadt.de/research/past-projects/exercise-difficulty-for-language-learning/

intuitive concept of item dependency properly. The assumption that item dependencies are caused by linguistic features was not borne out by our data and has therefore not been captured in computational measures. Instead, rather shallow position and neighbor effects were included. The findings could explain inconclusive results by Schroeders et al. (2014) and Eckes and Baghaei (2015). However, the hypothesis that item dependencies are related to linguistic features is too strong to be ultimately rejected and remains a research topic for psychometricians. Future research will hopefully shed light on this issue.

Chapter 7 contains the main contribution of this thesis. The four difficulty dimensions are combined into a single model to predict the item difficulty of text-completion exercises. The evaluations show that the automatic prediction reaches the performance of a single human expert if enough training data is available. As our approach is trained on real error rates and not on human annotations, more fine-tuned approaches should be able to outperform human experts. The automatic difficulty prediction has three advantages over the judgment by human experts. It is based on objective measurable properties, it is consistent, and it is faster.

Chapter 8 analyzed applications of the difficulty prediction approach for the goal of difficulty manipulation. We implemented a web demo and a novel approach for test generation and selection from a corpus that was positively evaluated by language teachers. Further modification and adaptation of exercises can be improved with our generated resource for lexical substitution targeted at language learners. For the difficulty manipulation of cloze tests, an approach for distractor generation and evaluation has been developed that is based on the implemented candidate rankers. The evaluation with users has shown that this approach works for some examples, but needs to be further refined and combined with other approaches for distractor generation. In particular, progress in the automatic analysis of semantic relations and selectional preferences of verbs can lead to improvements for this task.

In this thesis, we have developed computational features that are motivated by psycholinguistic theories. In future work, the computational approaches can in turn contribute empirical results to improve theories about language learning.

## 9.2 Outlook

Most of the measures developed in this thesis profited from natural language processing techniques that have been developed for other tasks. Accordingly, our contributions to difficulty prediction can hopefully also lead to further improvements in related tasks. Word difficulty measures can be exploited to improve readability and simplification systems. The

analyses on cross-lingual transfer can lead to improved natural language identification and exercise generation that considers the background knowledge of the learner. The computational approaches for evaluating candidate ambiguity for language learners might foster improvements in learner-specific sentence completion and other writing assistance tools.

Computer-assisted language learning is a truly interdisciplinary research area attracting interest from educational researchers, psychologists, linguists, and of course computer scientists. It is our strong conviction that real progress can only be made when knowledge from the different disciplines is consolidated. This thesis has contributed to interdisciplinary collaboration by modeling psycholinguistic findings as computational measures. The analysis of item dependencies is strongly influenced by psychometric approaches and the results have already generated interest from psychological researchers and raised new research questions. Making our approach available as a web demo was greatly appreciated by educational researchers and practitioners. Integrating their feedback could lead to technical improvements for future projects in educational natural language processing.

The main contribution of this thesis is the implementation of an approach for automatic difficulty prediction of exercises. The consequent next step after automatic difficulty prediction is tackling more focused exercise generation.

Improving exercise generation In the past, exercise generation approaches have been built to explore technical feasibility rather than to match educational needs. In order to support learning, the generated exercises need to fit the learner's proficiency and should be targeted at the learning goals. In chapter 8, we have already explored first attempts towards informed difficulty manipulation. Our difficulty model provides the opportunity to anticipate the difficulty of exercises without testing them directly on learners. The approaches for content selection, lexical substitution, and distractor manipulation all need to be explored further and will benefit from technological advances. Content selection needs to be controlled by topic and genre modeling, lexical substitution by word sense disambiguation, and distractor manipulation by textual entailment approaches to generate better exercises that are targeted towards the learners' needs.

The direct reference points for future work regarding the individual measures have already been pointed out above. The following paragraphs discuss two more general lines of research that could be pursued from here by taking contrasting perspectives.

Widening the focus In many parts of this thesis, we have focused on very specific phenomena. The high prediction quality has been confirmed for several datasets, but all of them contain text-completion exercises in Western-European languages. An interesting next step would be to widen the focus and extend the approach to languages from other language families and other exercise types. Open exercises, task-based and communicative

learning, and interactive approaches have a higher reputation from an educational point-of-view than the more static exercises that are favored by computer scientists. In future work, the difficulty concept thus needs to be adapted to account for exercises outside the text-completion paradigm. The findings of this thesis could be combined with recent progress in the field of essay grading and short answer scoring. In addition, the approaches need to be integrated with speech technology to address the full skill set of language learning.

With the ongoing trends of massive open-online courses, computer-based assessment, and big data analytics, more and in particular bigger datasets will become available. The example-based error analyses in this thesis can then be replaced with large-scale parameter tuning to adjust the difficulty concept to the specific tasks at hand and further improve the performance of exercise difficulty prediction.

Narrowing the focus In contrast to the line of research described above, the opposite direction also promises interesting insights. In the thesis, we have seen that the background of the learner plays an important role, for example, for cross-lingual transfer. Our datasets contain information from many heterogeneous learners at one point in time. Difficulty is thus interpreted as an aggregated mean over all learners and is approximated by linguistic properties of the exercise. For future work, it would be interesting to evaluate the difficulty of an exercise with respect to the learner's ability. For this kind of research, longitudinal performance data from more focused target groups or even from individual learners is required. In order to properly predict learner performance, the difficulty model then needs to be combined with a learner model that represents the learner's proficiency and also other characteristics such as personal preferences, learning style, and domain knowledge. The main challenge here consists in dynamically updating the difficulty estimate of an exercise according to the learner's progress and to the learning goals. Another important factor that we neglected in this thesis is the role of the teaching style. Different educational theories and cultural preferences lead to different didactic approaches which in turn have an influence on the competence of the learner and the experienced difficulties. Being able to generate exercises with fine-grained adaptivity that can be used directly for individual language training will strongly increase the acceptance and also the educational benefit of technological approaches.

Supporting multilingualism is a high priority for the *European Union*: "One of the EU's multilingualism goals is for every European to speak 2 languages in addition to their mother tongue." This ambition is based on the motivation that language skills foster inter-cultural understanding in our modern, heterogeneous societies. We hope that the results of this thesis can contribute to the integration of technology into language learning scenarios and lead to more flexible learner support.

<sup>94</sup>See http://europa.eu/pol/mult/index en.htm, accessed: December 17,2015

## Index

ablation test, 144 cognate production (COP), 28, 89 accuracy (A), 58 cognateness, 85 achievement test, 10 cognates, 75, 85, 167 anchor test, 50 coherence, 67 answer, 35 cohesion, 67 answer format, 17, 39 comprehension exercises, 14 average sentence length, 38 computer-adaptive testing, 13 computer-based testing, 13 basic vocabulary, 107 conceptual mapping, 73 concreteness, 81 C-test, 32, 33 construct validity, 11, 35 C-test paragraph, 48 content factors, 44 CALL: computer-assisted language learncontent validity, 11 ing, 12 content word, 66 candidate, 35 context fitness, 169 candidate ambiguity, 3, 44, 45, 169 context specificity, 84 candidate fitness, 124, 125 convergent validity, 11 candidate space, 45, 118 coverage, 90 CEFR, 9, 10 criterion, 11 CEFR manual, 11 criterion-referenced testing, 10 character sequence probability, 107 cross-lingual transfer, 74 character-based machine translation, 89 classical test theory, 51 deletion rate, 32, 39 classification, 140 Dice, 88 closed exercises, 17 cloze exercise, 16, 24 difficulty model, 44 distractor, 24, 35 cloze test, 32 co-occurrence score, 38 distractor generation, 25 coefficient of variation, 55 distributional hypothesis, 83

#### INDEX

early closure, 38 inter-annotator agreement, 137 educational objectives, 14 interpretation exercises, 16 EFC corpus, 27, 108 item dependency, 3, 37, 44, 118 error correction, 27, 105 item design, 40 error detection, 27 item response theory (IRT), 37, 51 error rate, 52 item transfer, 74 essay grading, 29 item-level difficulty prediction, 39 exercise, 35 L1, 27, 72 exercise content, 3, 37 L1 acquisition, 72 exercise difficulty, 23, 25 L2, 27, 72 exercise format, 3, 37, 39 L2 learning, 72 exercise generation, 20 L2 readability, 72 exercise quality, 23, 25 language model, 64 explicit semantic analysis, 127 language model fitness, 126  $F_1$ -measure, 58 language proficiency, 31 face validity, 11 LCSR, 88, 125 false friends, 85 leave-one-out cross-validation, 59 false friends detection, 100 leave-one-test-out cross-validation, 138 FCE corpus, 28, 109 length principle, 40, 122 Fleiss' kappa, 137, 164 lexical association, 73 format factors, 44 lexical chain, 67 formative assessment, 10 lexical diversity, 65 frequency, 64 lexical fossilization, 79 function word, 66 lexical inferencing, 83 lexical richness, 38 gap, 35 lexical simplification, 166 gap type, 39 linguistic exercises, 14 global difficulty, 44 local difficulty, 44 grammar exercise, 22 local independence, 53 grammatical cloze exercise, 25 local item independence, 130 grammaticality, 171 grapheme-to-phoneme ratio, 107 macro-level processing, 5, 36, 44, 157 MALL: mobile-assisted language learning, half-open exercises, 17 12 high-stakes test, 10 manipulation exercises, 16 hint, 35 mean reciprocal rank (MRR), 90 human upper bound, 47

mean test score, 51

MERLIN corpus, 27, 109

micro-level processing, 5, 36, 44

ICALL: intelligent computer-assisted lan-

guage learning, 18

multiple choice, 17 reduced redundancy, 118 multiple response, 17 reduced redundancy principle, 31 redundancy, 31 natural language identification, 27 referential gap, 132 natural language processing (NLP), 18 regression, 140 neighbor effect, 132 relaxed scoring, 33 noise, 31 reliability, 11, 35 noticing exercise, 23 revised hierarchical model, 73 RMSE, 143 objectivity, 10 root mean square error (RMSE), 59 odd-one-out question, 16 open exercises, 17 selective attention, 16 orthographic depth, 107 self-directed learning, 13 overgenerate-and-rank, 21, 24 semantic fit, 171 semantic fitness, 127 P-value, 51 semantic relatedness, 127 Pearson correlation (Pearson's r), 59, 140 short answer scoring, 30 phonetic density, 107 shuffle exercise, 16, 23 phonetic score, 108 significance, 59 placement test, 10 simplification, 19 polysemy, 81 solution, 35 poverty of the stimulus, 73 Spearman correlation (Spearman's r), 140 precision (P), 58 spelling difficulty, 105 prefix, 35 spelling error probability, 110 prefix deletion test, 34 SpSim measure, 89 production, 17, 36, 167 string similarity, 88 production exercises, 16 substitution candidates, 126 proficiency classification, 29 summative assessment, 10 proficiency test, 10 support vector machines, 111, 140, 156 prompt, 29 syntactic complexity, 83 pronunciation clarity, 108 system transfer, 74 pronunciation difficulty, 108 target hypothesis, 28  $Q_3$ -model, 129 test, 35 question generation, 22 test score, 51 Rasch model (1PL), 52, 130 text difficulty, 3, 44, 62 readability, 19, 62 text-completion exercise, 24, 31 recall (R), 58 tribond question, 16 recognition, 17, 36, 75, 167 TTR. 68

recognition exercises, 16

two-parametric logistic model (2PL), 53

#### INDEX

type-token ratio, 38, 65

unitary trait hypothesis, 32

validity, 11, 35 verb variation, 66 vocabulary exercise, 22

web interface, 48 word class, 38, 83

word difficulty, 3, 44 word frequency, 38, 106 word length, 106

X-test, 32, 34 XDice, 88 XSL, 68

zone of proximal development, 62

# **List of Figures**

1.1	Thesis model for exercise difficulty	4
1.2	Cloze exercise	4
3.1	Screenshot of the web interface	49
3.2	Sample item characteristic curves	53
3.3	Sorted difficulty parameters	54
3.4	Stability of difficulty measures	56
3.5	Standard error of error rates	57
4.1	Text difficulty: macro-level dimension of the exercise content	61
4.2	Paragraph difficulty	71
4.3	Visualisation of error rates	71
4.4	Language representation in the minds of language learners	74
5.1	Word difficulty: micro-level dimension of the exercise content	79
5.2	Cognate production workflow	89
5.3	COP learning curve	94
6.1	The exercise format	117
6.2	Candidate space for different languages	120
6.3	Candidate space for different test types	121
6.4	Learner candidate space	122
6.5	Candidate ambiguity: micro-level dimension of the exercise format	124
6.6	Visualisation of the search space for C-tests	125
6.7	Item dependencies: macro-level dimension of the exercise format	129
7.1	Exercise difficulty	135
7.2	Comparison of human and automatic difficulty predictions	146
7.3	C-test predictions	147

## List of Figures

7.4	X-test predictions	151
7.5	Cloze test predictions	154
7.6	Learning curve	157
0 1	Web dame for difficulty and disting	1/1
0.1	Web demo for difficulty prediction	101
8.2	Example for human evaluation	163
8.3	Results of the human evaluation for text selection	164

# **List of Tables**

2.1	Overview of text-based exercises	15
3.1	Overview of exercise datasets	47
3.2	English C-test data	49
3.3	Difficulty statistics for German gaps	55
	Mean error rate and standard deviation for each training dataset	58
4.1	Readability scores for sample texts	64
4.2	Correlation of readability features with mean difficulty	69
4.3	Mean difficulty prediction	70
5.1	Results of the cognate study	86
5.2	Training lists for minimal model	92
5.3	Parameter selection for <i>COP</i>	93
5.4	Influence of data quality and size	95
5.5	Results for cognate production	96
5.6	Cognate datasets for different languages	97
5.7	COP results for other languages	99
5.8	Multilingual cognates	100
	Results for false friends detection	102
5.10	COP productions for pilot study	103
5.11	Extracted words and spelling errors after pre-processing	110
5.12	Examples for spelling error probability	111
5.13	Feature analysis for spelling difficulty	112
5.14	Results for spelling difficulty prediction	112
5.15	Results for cross-corpus comparison	113
5.16	Most frequent misspellings for selected examples	114
6.1	Vocabulary sizes for different languages	119

## List of Tables

6.2	Average candidate space	120
6.3	Solving results for candidate evaluation strategies	128
6.4	Pairs of dependent gaps	131
6.5	Neighboring gaps and word repetitions in dependent pairs	131
6.6	Dependent gap pairs across samples	132
7.1	Results of the human annotation	137
7.2	Confusion matrix for human median prediction	137
7.3	Features for the difficulty prediction experiments	139
7.4	Regression and classification prediction of C-test difficulty	141
7.5	C-test regression results for different feature groups	143
7.6	Selected features	144
7.7	Prediction results for C-tests	145
7.8	X-test regression results for different feature groups	150
7.9	Cloze test regression results for different feature groups	153
8.1	Lexical substitution examples	168
8.2	Evaluation of grammaticality of the substitution candidates	172
8.3	Evaluation of semantic fit of the substitution candidates	172

## **Bibliography**

- Abraham, R. G. and Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76(4):468–479. 38, 45
- Adamson, G. W. and Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10(7):253–260. 88
- Agarwal, M. and Mannem, P. (2011). Automatic gap-fill question generation from text books. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64. Association for Computational Linguistics. 25
- Agarwal, M., Shah, R., and Mannem, P. (2011). Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics. 21
- Ai, R., Krause, S., Kasper, W., Xu, F., and Uszkoreit, H. (2015). Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In *Proceedings of the Seventh International Joint Conference on Natural Language Processing*, pages 26–33. Asian Federation of Natural Language Processing. 22
- Al-Khalifa, H. S. and Al-Ajlan, A. (2010). Automatic readability measurements of the arabic text: an exploratory study. *The Arabian Journal for Science and Engineering*, 35(2C). 67
- Aldabe, I., de Lacalle, M. L., Maritxalar, M., Martinez, E., and Uria, L. (2006). ArikIturri: an automatic question generator based on corpora and NLP techniques. In Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 584–594. Springer. 23, 25, 26
- Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 2(13):219–227. 32, 33

- Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics. 20, 67
- Amaral, L. A. and Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *The Journal of the European Association for Computer Assisted Language Learning (ReCALL)*, 23(01):4–24. 12, 18
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing (U.S.) (1999). *Standards for educational and psychological testing*. American Educational Research Association. 10
- Anagnostou, N. K. and Weir, G. R. (2006). From corpus-based collocation frequencies to readability measure. In *ICT in the Analysis, Teaching and Learning of Languages*, pages 33–46. 84
- Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v.2. Journal of Technology, Learning and Assessment, 4(3). 13, 29
- Baba, Y. and Suzuki, H. (2012). How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 373–377. Association for Computational Linguistics. 106
- Babaii, E. and Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle? *System*, 29(2):209–219. 5, 34, 36, 44, 45, 129
- Babaii, E. and Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, 34(4):586–600. 36, 66
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing*, 17(1):1–42. 32
- Becker, J. (2004). Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT). PhD thesis, Freie Universität Berlin, Germany. 52
- Becker, L., Basu, S., and Vanderwende, L. (2012). Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 742–751. Association for Computational Linguistics. 24

- Beinborn, L., Zesch, T., and Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the 1st Workshop on NLP for computer-assisted language learning*, volume 80, pages 11–19. Linköping University Electronic Press, Linköping Universitet. 8, 62
- Beinborn, L., Zesch, T., and Gurevych, I. (2013). Cognate production using character-based machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891. Asian Federation of Natural Language Processing. 8, 85, 88, 89, 222
- Beinborn, L., Zesch, T., and Gurevych, I. (2014a). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–529. 8, 85
- Beinborn, L., Zesch, T., and Gurevych, I. (2014b). Readability for foreign language learning: The importance of cognates. *International Journal of Applied Linguistics*. 8, 46, 62, 135, 138, 140, 149
- Beinborn, L., Zesch, T., and Gurevych, I. (2015a). Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications held in conjunction with NAACL 2015*, pages 1–11. 8, 34, 46, 119, 135, 138, 140
- Beinborn, L., Zesch, T., and Gurevych, I. (2015b). Factors of difficulty in German language proficiency tests. Book of Abstracts: Language, Learning, Technologgy Conference. 8
- Beinborn, L., Zesch, T., and Gurevych, I. (2016). Predicting the spelling difficulty of words for language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications held in conjunction with NAACL 2016*, pages 73–83. The Association for Computational Linguistics. 8, 105
- Benjamin, R. G. (2011). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88. 63
- Bensoussan, M. and Ramraz, R. (1984). Testing EFL reading comprehension using a multiple-choice rational cloze. *The Modern Language Journal*, 68(3):230–239. 33
- Berkling, K., Pflaumer, N., and Coyplove, A. (2015). Phontasia—a game for training German orthography. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, pages 1874–1875. 105
- Bernhard, D., De Viron, L., Moriceau, V., Tannier, X., et al. (2012). Question generation for French: collating parsers and paraphrasing questions. *Dialogue and Discourse*, 3(2):43–74. 21

- Berthele, R., Peyer, E., and Kaiser, I. (2011). On abduction in receptive multilingualism. evidence from cognate guessing tasks. *Applied Linguistics Review*, 2:191–220. 87
- Bick, E. (2004). Grammar for fun: IT-based grammar learning with VISL. Copenhagen studies in language, 30:49. 23
- Bickel, S., Haider, P., and Scheffer, T. (2005). Predicting sentences using N-gram language models. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing HLT '05*, pages 193–200. Association for Computational Linguistics. 126
- Biemann, C. and Riedl, M. (2013). Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95. 170
- Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, pages 395–479. 52, 53
- Blake, R. J. (2011). Current trends in online language learning. *Annual Review of Applied Linguistics*, 31:19–35. 12
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay Company. 14, 15
- Boyd, A. (2009). Pronunciation modeling in spelling correction for writers of English as a foreign language. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop and Doctoral Consortium*, pages 31–36. Association for Computational Linguistics. 105
- Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., and Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC*). 27, 109
- Brants, T. and Franz, A. (2006). Web 1T 5-gram corpus version 1.1. *Linguistic Data Consortium*. 24, 65, 93, 106, 142, 220, 223, 224
- Bresnihan, B. and Ray, S. (1992). C-tests and the usefulness of non-linguistic instructions. In Grotjahn, R., editor, *Der C-Test. Theoretische Grundlagen und praktische Anwendungen 1*, pages 193–216. Brockmeyer, Bochum. 40, 123
- Brew, C. and McKelvie, D. (1996). Word-pair extraction for lexicography. *Proceedings of the second international conference on new methods in language processing*, pages 45–55. 88

- Broerse, A. C. and Zwaan, E. (1966). The information value of initial letters in the identification of words. *Journal of Verbal Learning and Verbal Behavior*, 5(5):441–446. 34, 40
- Brown, J. C., Frishkoff, G. A., and Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826, Morristown, NJ, USA. Association for Computational Linguistics. 22, 24, 25, 80, 83
- Brown, J. D. (1989). Cloze item difficulty. Journal of the Japan Association for Language Teaching, 11:46-67. 11, 32, 38, 40, 44, 61, 83
- Brown, J. D. (1998). An EFL readability index. *Journal of the Japan Association for Language Teaching*, 20(2):7–36. 72
- Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990. 64, 65
- Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.
- Burrows, S., Gurevych, I., and Stein, B. (2014). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117. 30
- Carrell, P. L. (1987). Readability in ESL. Reading in a Foreign Language, 4(1):21-40. 72
- Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: a review. *International Journal of Bilingualism*, 7(1):71–87. 75
- Chali, Y. and Hasan, S. A. (2015). Towards topic-to-question generation. *Computational Linguistics*, 41:1–20. 21, 24
- Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, 10(2):157–187. 36
- Chen, C.-Y., Liou, H.-C., and Chang, J. S. (2006). FAST: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL Interactive presentation sessions*, pages 1–4. Association for Computational Linguistics. 25

- Chen, W., Aist, G., and Mostow, J. (2009). Generating questions automatically from informational text. In *Proceedings of the 2nd Workshop on Question Generation (AIED 2009)*, pages 17–24. 21
- Chinnery, G. (2006). Going to the MALL: mobile assisted language learning (Emerging technology). *Language Learning & Technology*, 10(1):9–16. 12
- Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press, Cambridge, MA. 73
- Cleary, C. (1988). The C-test in English: left-hand deletions. *RELC Journal*, 19(2):26–35. 34, 151
- Cohen, A. D. (1984). The C-Test in Hebrew. Language Testing, 1(2):221-225. 36
- Coleman, M. and Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284. 63
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135. 20, 61, 62
- Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462. 63, 65
- Colton, S. (2002). Automated puzzle generation. In *Proceedings of the AISB Symposium on AI and Creativity in the Arts and Science*, pages 99–108. 16
- Cook, V. (1997). L2 users and English spelling. *Journal of Multilingual and Multicultural Development*, 18(6):474–488. 106
- Cook, V. J., Long, J., and McDonough, S. (1979). First and second language learning. In Perren, G., editor, *The Mother Tongue and Other Languages in Education*, pages 7–22. Routledge. 72, 76
- Council of Europe: Language Policy Division (2009). Relating examinations to the common european framework: a manual. http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL\_en.pdf. last accessed: October 18, 2015. 11
- Council of Europe: Language Policy Division (2011). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press. 9
- Crossley, S. A., Louwerse, M. M., McCarthy, P. M., and McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30. 20

- Crossley, S. A., Salsbury, T., and McNamara, D. S. (2012). Predicting the level of language learners using lexical indices. *Language Testing*, 29(2):243–263. 29
- Crystal, D. (2011). Dictionary of linguistics and phonetics, volume 30. Wiley-Blackwell. 85
- Curto, S. and Mendes, A. (2012). Question generation based on lexico-syntactic patterns learned from the web. *Dialogue & Discourse*, 3(2):147–175. 21
- Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin.* 64
- Dale, R., Anisimoff, I., and Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62. Association for Computational Linguistics. 28
- Dale, R. and Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249. Association for Computational Linguistics. 28
- Danielsson, P. and Mühlenbock, K. (2000). Small but efficient: The misconception of high-frequency words in Scandinavian translation. *Envisioning Machine Translation in the Information Future*, pages 158–168. 88
- Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). DKPro TC: a Java-based framework for supervised learning experiments on textual data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland. Association for Computational Linguistics. 7, 111, 140
- De Belder, J. and Moens, M.-F. (2012). A dataset for the evaluation of lexical simplification. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 426–437. Springer Berlin Heidelberg. 167
- De Groot, A. M. B. and Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1):1–56. 75, 81, 82, 85
- de Melo, G. and Weikum, G. (2009). Towards a universal WordNet by learning from combined evidence. *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 513–522. 95

- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. 127
- DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In Long, M. H. and Doughty, C. J., editors, *The Handbook of Language Teaching*, chapter 8, pages 119–138. Wiley-Blackwell, Oxford, UK. 72
- Dela Rosa, K. and Eskenazi, M. (2011). Effect of word complexity on L2 vocabulary learning. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–80. Association for Computational Linguistics. 33, 106
- Dell'Orletta, F., Montemagni, S., and Venturi, G. (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics*, 165(2):163–193. 67
- Deorowicz, S. and Ciura, M. G. (2005). Correcting spelling errors by modelling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2):275. 105
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1). 29
- Dörnyei, Z. and Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, 9(2):187–206. 38, 44, 83
- Douglas, D. and Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27:115–132. 13
- DuBay, W. H. (2004). The principles of readability. *Impact Information*, pages 1–76. 61, 63
- Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., and Dang, H. T. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (\*SEM): Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274. Association for Computational Linguistics. 30
- Eckes, T. (2011). Item banking for C-tests: a polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, 53(4):414–439. 50, 52, 53
- Eckes, T. and Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education*, 28(2):85–98. 37, 54, 130, 179

- Eckes, T. and Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23(3):290–325. 36
- Ehara, Y., Sato, I., Oiwa, H., and Nakagawa, H. (2012). Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 799–814. 110
- Ellis, N. (1994). Consciousness in second language learning: psychological perspectives on the role of conscious processes in vocabulary acquisition. *AILA Review*, 11:37–56. 27, 72, 75
- Erbs, N., Santos, P. B., Zesch, T., and Gurevych, I. (2015). Counting what counts: Decompounding for keyphrase extraction. In *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*, pages 10–17, Beijing, China. Association for Computational Linguistics. 83
- Farhady, H. and Jamali, F. (2006). Varieties of C-test as measures of general language proficiency. Twenty-five years of living with applied linguistics: collection of articles, pages 287–302. 34
- Fellbaum, C. (1998). WordNet: An electronic database. MIT Press, Cambridge, MA. 22, 82, 95
- Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237, Morristown, NJ, USA. Association for Computational Linguistics. 19, 220
- Feng, L. and Huenerfauth, M. (2010). A Comparison of Features for Automatic Readability Assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 276–284. 63, 66, 67
- Ferreira de Souza, V. (2003). *The role of cognates in reading comprehension*. Phd thesis, Universidade Federal de Santa Catarina, Florianópolis, Brazil. 85
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32. 83
- Flor, M., Klebanov, B. B., and Sheehan, K. M. (2013). Lexical tightness and text complexity. In *Proceedings of the 2nd Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 29–38. 67

- Forman, G. and Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57. 59, 138
- François, T. and Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. 67, 72
- Frith, U. (1980). Unexpected spelling problems. Cognitive Processes in Spelling. 106
- Fulcher, G. and Davidson, F. (2007). *Language testing and assessment*. Routledge London and New York. 10
- Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611. 127
- Gamer, M., Lemon, J., Fellows, I., and Singh, P. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84. 137
- Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal*, 93:719–740. 12, 16
- Gartland, L. B. and Smolkin, L. B. (2015). The histories and mysteries of grammar instruction. *The Reading Teacher*, pages 1–9. 16
- Gates, D. M. (2011). How to generate cloze questions from definitions: A syntactic approach. In *AAAI Fall Symposium: Question Generation*, AAAI Technical Report. Association for the Advancement of Artificial Intelligence. 24
- Geertzen, J., Alexopoulou, D., and Korhonen, A. (2012). Automatic linguistic annotation of large scale L2 databases: the EF-Cambridge Open Language Database (EFCamDat). In Miller, R. T., editor, *Selected Proceedings of the 2012 Second Language Research Forum*. MA: Cascadilla Proceedings Project. 27, 108
- Gilhooly, K. J. and Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427. 82, 149
- Gilmore, A. (2011). "i prefer not text": Developing japanese learners' communicative competence with authentic materials. *Language Learning*, 61(3):786–819. 2, 18
- Glaboniat, M. (2010). Sprachprüfungen für Deutsch als Fremdsprache. In Krumm, H.- J., Fandrych, C., Hufeisen, B., and Riemer, C., editors, *Handbuch Deutsch als Fremd-und Zweitsprache*, volume 2, pages 1288–1299. Walter de Gruyter Berlin New York. 10, 17

- Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., and Freynik, S. (2014). Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning*, 27(1):70–105. 18
- Gomes, L. and Pereira Lopes, J. G. (2011). Measuring spelling similarity for cognate identification. *Progress in Artificial Intelligence*, pages 624–633. 89, 96
- Graesser, A. C. and McNamara, D. (2004). Coh-Metrix: analysis of text on cohesion and language. *Behavior Research Methods*, 36(2). 19, 63, 65, 67, 82
- Greenfield, J. (2004). Readability formulas for EFL. Journal of the Japan Association for Language Teaching, 26(1):5–24. 72
- Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225. 67
- Grotjahn, R. (2009). Prüfen Testen Bewerten. In Jung, U. O. H. and Jung, H., editors, Praktische Handreichung für Fremdsprachenlehrer, pages 221–230. Peter Lang, Bochum. 10, 17
- Grotjahn, R., Klein-Braley, C., and Raatz, U. (2002). C-tests: an overview. In Coleman, J. A., Grotjahn, R., and Raatz, U., editors, *University language testing and the C-Test*, pages 93–114. AKS-Verlag, Bochum. 34, 51
- Grotjahn, R. and Stemmer, B. (2002). C-tests and language processing. In Coleman, J. A., Grotjahn, R., and Raatz, U., editors, *University language testing and the C-Test*, pages 115–130. AKS-Verlag, Bochum. 5, 36
- Gunning, R. (1969). The Fog Index after twenty years. *Journal of Business Communication*, 6(2):3–13. 63
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). A large-scale unified lexical-semantic resource based on LMF. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. 7, 95, 168, 220, 222
- Haastrup, K. (1991). Lexical inferencing procedures, or, talking about words: Receptive procedures in foreign language learning with special reference to English. Gunter Narr Verlag. 83
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18. 111, 140, 144, 156

- Hambleton, R. K. and Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47. 53
- Hancke, J. and Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research 2013, Book of Abstracts*, pages 54–56. 29
- Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. *Proceedings of the 24th International Conference on Computational Linguistics*. 67
- Harsch, C. and Hartig, J. (2010). Empirische und inhaltliche Analyse lokaler Abhängigkeiten im C-Test. In Grotjahn, R., editor, *Der C-Test: Beiträge aus der aktuellen Forschung*, pages 193–204. Peter Lang. 37, 130, 131
- Harsch, C. and Hartig, J. (2015). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*. 37
- Hartrumpf, S. (2003). *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany. 65
- Heift, T. (2010). Developing an intelligent language tutor. *CALICO Journal*, 27(3):443–459. 18
- Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467. Association for Computational Linguistics. 19, 25, 63, 66, 72, 83
- Heilman, M. and Eskenazi, M. (2006). Language learning: Challenges for intelligent tutoring systems. In *Proceedings of the workshop of intelligent tutoring systems for ill-defined tutoring systems at the 8th international conference on intelligent tutoring systems*, pages 20–28. 19
- Heilman, M. and Eskenazi, M. (2007). Application of automatic thesaurus extraction for computer generation of vocabulary questions. In *Speech and Language Technology in Education (SLaTE)*, pages 65–68. 22, 23
- Heilman, M. and Smith, N. A. (2010). Good question! Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. Association for Computational Linguistics. 21, 24

- Hockly, N. (2015). Developments in online language learning. *English Language Teaching Journal*, pages 1–6. 12, 14
- Horbach, A., Palmer, A., and Pinkal, M. (2013). Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM)*, volume 1, pages 286–295. 30
- Horsmann, T. and Zesch, T. (2014). Towards automatic scoring of cloze items by selecting low-ambiguity contexts. *Northern European Journal of Language Technology Proceedings Series Vol. 22*, pages 33–42. 33
- Hoshino, A. and Nakagawa, H. (2007). Assisting cloze test making with a web application. In *Society for Information Technology & Teacher Education International Conference*, volume 2007, pages 2807–2814. 25
- Hoshino, A. and Nakagawa, H. (2008). A cloze test authoring system and its automation. In *Proceedings of the 6th International Conference on Advances in Web-based Learning*, pages 252–263. Springer. 39, 156
- Hoshino, N. and Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, 106(1):501–11. 75
- Hughes, A. (1989). Testing for Language Teachers. Cambridge University Press. 10
- Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257. 88, 96, 97, 100
- Jafarpur, A. (1995). Is C-testing superior to cloze? Language Testing, 12(2):194–216. 34
- Jakschik, G., Klemmert, H., and Klinck, D. (2010). Computergestützter Multiple Choice C-Test in der Bundesagentur für Arbeit: Bundesweite Erprobung und Einführung. In Grotjahn, R., editor, *Der C-Test: Beiträge aus der aktuellen Forschung / The C-Test: Contributions from Current Research*, pages 231–264. Peter Lang International Academic Publishers. 37, 39
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21(1):47–77. 72, 73, 74, 79
- Joseph, S. R. and Uther, M. (2009). Mobile devices for language learning: Multimedia approaches. *Research and Practice in Technology Enhanced Learning*, 4(01):7–32. 12
- Joshi, R. and Aaron, P. (2013). Handbook of Orthography and Literacy. Taylor & Francis. 107

- Kalady, S. (2010). Natural language question generation using syntax and keywords. *Proceedings of the Third Workshop on Question Generation*. 21
- Kamimoto, T. (1993). Tailoring the test to fit the students: Improvement of the C-Test through classical item analysis. *Language Laboratory*, 30:47–61. 53
- Karimi, N. (2011). C-test and vocabulary knowledge. *Language Testing in Asia*, 1(4):7. 38, 65
- Kaushanskaya, M. and Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. *Psychonomic Bulletin & Review*, 19(5):935–941. 81
- Khodadady, E. and Hashemi, M. (2011). Validity and C-Tests: The role of text authenticity. *Iranian Journal of Language Testing*, 1:30–41. 34, 35
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, DTIC Document. 19, 63
- Kinoshita, S. (2000). The left-to-right nature of the masked onset priming effect in naming. *Psychonomic Bulletin & Review*, 7(1):133–141. 34, 40
- Klein-Braley, C. (1984). Advance prediction of difficulty with C-Tests. In Culhane, T., Klein-Braley, C., and Stevenson, D. K., editors, *Practice and problems in language testing*, volume 7. 38, 44, 52, 62, 68, 69, 70, 143, 153
- Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing*, 2(1):76–104. 35
- Klein-Braley, C. (1996). Towards a theory of C-Test processing. In Grotjahn, R., editor, *Der C-Test. Theoretische Grundlagen und praktische Anwendungen 3*, pages 23–94. Brockmeyer, Bochum. 38, 44, 52, 83
- Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, 14(1):47–84. 32, 34, 160
- Klein-Braley, C. and Raatz, U. (1982). Der C-Test: ein neuer Ansatz zur Messung allgemeiner Sprachbeherrschung. *AKS-Rundbrief*, 4:23 37. 33
- Klein-Braley, C. and Raatz, U. (1984). A survey of research on the C-Test. *Language Testing*, 1(2):134–146. 33, 65
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1459—1473. 101

- Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, 86(4):571–586. 35, 38, 45, 75, 81, 83
- Köberl, J. and Sigott, G. (1994). Adjusting C-test difficulty in German. *Der C-Test. Theoretische Grundlagen und praktische Anwendungen 2.* 34, 39, 118, 151, 169
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Zens, R., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Dyer, C., Constantin, A., and Herbst, E. (2007). Moses:
  Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics*. 20, 89
- Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pages 288–295. 88
- Kondrak, G. and Dorr, B. (2004). Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 952–958. 88, 96
- Krampen, D. (2014). Zur Bedeutung des Testformats für die Testauswertung: Aufgabenstammund Antwortabhängigkeiten im C-Test. Peter Lang Frankfurt. 37, 129, 130
- Kroll, J. F. and Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33:149–174. 73, 74, 75
- Landerl, K., Wimmer, H., and Frith, U. (1997). The impact of orthographic consistency on dyslexia: a German-English comparison. *Cognition*, 63(3):315–34. 104
- Larson, J. W. and Madsen, H. S. (1985). Computerized adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal*, 2(3):32–36. 13
- Larsson, P. (2006). Classification into Readability Levels Implementation and Evaluation. Master's thesis, Uppsala University, Sweden. 67
- Laufer, B. and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22:15–30. 76, 79
- Lee, J. and Seneff, S. (2007). Automatic generation of cloze items for prepositions. In *Proceedings of INTERSPEECH*, pages 2173–2176, Antwerp, Belgium. 25

- Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44(2):325–343. 29
- Lemhöfer, K. and Dijkstra, T. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):12. 85
- Lervåg, A. and Aukrust, V. G. v. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, 51(5):612–20. 76, 79, 84
- Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics-Doklady*, volume 10, pages 707–710. 108
- Levy, M. (2009). Technologies in use for second language learning. *The Modern Language Journal*, 93(s1):769–782. 12
- Liang, F. M. (1983). Word Hy-phen-a-tion by Com-put-er. PhD thesis, Stanford University, California, US. 221
- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the* 17th international conference on on Computational linguistics, volume 2, pages 768–774. Association for Computational Linguistics. 84
- Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015). Character-based neural machine translation. *Computing Research Repository*, 1511.04586. 104, 178
- Liou, H.-C. (2013). Can grammatical CALL help EFL writing instruction? *CALICO journal*, 10(1):23–44. 16
- List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS\UNCLH*, pages 117–125. 88
- Liu, C.-L., Wang, C.-H., Gao, Z.-M., and Huang, S.-M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 1–8. Association for Computational Linguistics. 25, 26
- Liu, L., Mostow, J., and Aist, G. S. (2013). Generating example contexts to help children learn word meaning. *Natural Language Engineering*, 19:187–212. 24
- Ljubešic, N., Lucica, I., and Fišer, D. (2013). Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77. 100

- Lord, F. M., Novick, M. R., and Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley. 51
- Lotto, L. and de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, 48(1):31–69. 75
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208. 66
- Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. *Fortgeschrittene Lernervarietäten*, pages 119–140. 28
- Malmasi, S., Tetreault, J., and Dras, M. (2015). Oracle and human baselines for native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178. 27, 75
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. 161, 219
- McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–616. 84
- McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646. 63
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. 111
- Mead, A. D. and Meade, A. W. (2010). Test construction using CTT and IRT with unrepresentative samples. *Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA.* 56
- Meara, P. (1988). Learning words in an L1 and an L2. Polyglot, 9(3):1-11. 73, 76
- Medero, J. and Ostendorf, M. (2009). Analysis of vocabulary difficulty using Wiktionary. *Proceedings of the 2nd Workshop on Speech and Language Technology in Education*. 106
- Meißner-Stiffel, M. and Raatz, U. (1996). \_\_\_\_ oder \_ \_ \_ ? Zwei Grundlagenuntersuchungen zum C-Prinzip bei L1-Lernern. In Grotjahn, R., editor, *Der C-Test. Theoretische Grundlagen und praktische Anwendungen 3*, pages 173–181. Brockmeyer, Bochum. 40

- Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130. 88, 125
- Meunier, L. E. (2013). Computer adaptive language tests (CALT) offer a great potential for functional testing. Yet, why don't they? *CALICO journal*, 11(4):23–39. 13
- Meurers, D. (2012). Natural language processing and language learning. *The Encyclopedia of Applied Linguistics*. 18
- Meurers, D. (2015). Learner corpora and natural language processing. In Sylviane Granger, G. G. and Meunier, F., editors, *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press. 27
- Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., and Ott, N. (2010). Enhancing authentic web pages for language learners. In *The 5th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*, volume 2. 23, 25
- Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9. Association for Computational Linguistics. 30
- Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press. 101
- Mitkov, R., Ha, L. A., and Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194. 21, 25, 26
- Mitkov, R., Pekar, V., Blagoev, D., and Mulloni, A. (2008). Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53. 100
- Mitschian, H. (2010). *M-Learning-die neue Welle? Mobiles Lernen für Deutsch als Fremd-sprache.* kassel university press GmbH. 12
- Montalvo, S., Pardo, E. G., Martinez, R., and Fresno, V. (2012). Automatic cognate identification based on a fuzzy combination of string similarity measures. *IEEE International Conference on Fuzzy Systems*, pages 1–8. 88, 92, 96
- Moser, J. R., Gütl, C., and Liu, W. (2012). Refined distractor generation with LSA and stylometry for automated multiple choice question generation. In *AI 2012: Advances in Artificial Intelligence*, pages 95–106. Springer. 25

- Mulloni, A. (2007). Automatic prediction of cognate orthography using support vector machines. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, pages 25–30. 89
- Nagata, N. (2009). Robo-Sensei's NLP-based error detection and feedback generation. *Calico Journal*, 26(3):562–579. 18
- Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 10, pages 301–305. 92
- Nakov, S. (2009). Automatic identification of false friends in parallel corpora: statistical and semantic approach. *Serdica Journal of Computing*, 3:133–158. 100
- Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Teaching Methods. Heinle & Heinle. 80
- Nation, P. (2003). The role of the first language in foreign language learning. *Asian EFL Journal*, 5(2):1–8. 73
- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12. 28, 105
- Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12. 28, 105
- Nicolai, G., Hauer, B., Salameh, M., Yao, L., and Kondrak, G. (2013). Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145. 104
- Niraula, N. B. and Rus, V. (2015). Judging the quality of automatically generated gap-fill question using active learning. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications held in conjunction with NAACL 2015*, pages 196–206. 24
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51. 94
- Odlin, T. (1989). Language Transfer: Cross-Linguistic Influence in Language Learning. Cambridge University Press. 74

- Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning*, 23(1):105–118. 32, 40
- Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die neueren Sprachen*, 75(2):165–174. 32
- Olney, A., Graesser, A., and Person, N. (2012). Question generation from concept maps. *Dialogue & Discourse.* 21
- Ott, N. and Ziai, R. (2010). Evaluating dependency parsing performance on German learner language. *Proceedings of TLT-9*, 9:175–186. 27
- Papineni, K., Roukos, S., Ward, T., and Ibm, W.-j. Z. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association of Computational Linguistics*, pages 311–318. 91
- Pathan, M. M. (2012). Computer Assisted Language Testing [CALT]: Advantages, implications and limitations. *Research Vistas*, 1:30–45. 12
- Pauls, A. and Klein, D. (2011). Faster and smaller n-gram language models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 107, 222
- Paulston, C. B. (1970). Structural pattern drills: A classification. *Foreign Language Annals*, 4(2):187–193. 16
- Peereman, R., Lété, B., and Sprenger-Charolles, L. (2007). Manulex-infra: Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, 39(3):579–589. 106
- Perez-Beltrachini, L., Gardent, C., and Kruszewski, G. (2012). Generating grammar exercises. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 147–156. 23, 25
- Perlmann-Balme, M. (2010). Testen und Prüfen von Sprachkenntnissen. In Krumm, H.-J., Fandrych, C., Hufeisen, B., and Riemer, C., editors, *Deutsch als Fremdsprache. Ein internationales Handbuch.*, volume 2, pages 1272–1288. De Gruyter. 13, 16
- Petersen, S. E. and Ostendorf, M. (2007). Text simplification for language learners: A corpus analysis. In *Proceedings of Speech and Language Technology in Education (SLaTE2007)*, pages 69–72. 20
- Pilán, I., Volodina, E., and Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184. 20, 67, 178

- Pino, J. and Eskenazi, M. (2009). Semi-automatic generation of cloze question distractors effect of students' L1. In *Proceedings of the International Workshop on Speech and Language Technology in Education*, pages 65–68. 25
- Pino, J., Heilman, M., and Eskenazi, M. (2008). A selection strategy to improve cloze question quality. *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, pages 22–32. 24, 25
- Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics. 63, 67
- Potter, M. C., So, K.-F., Eckardt, B. V., and Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, 23(1):23–38. 74
- Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, volume 17991802. 126
- Raatz, U. and Klein-Braley, C. (2002). Introduction to language testing and to C-Tests. *University language testing and the C-test*, pages 75–91. 33, 105
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, volume 1 of *Studies in Mathematical Psychology / Danmarks Paedagogiske Institut*. Nielsen and Lydich, Kopenhagen. 52, 129
- Reichert, M., Keller, U., and Martin, R. (2010). The C-test, the TCF and the CEFR: a validation study. *Der C-Test: Beiträge aus der aktuellen Forschung. The C-Test: Contributions from Current Research*, pages 205–231. 34
- Revelle, W. (2014). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 1.4.5. 59, 143
- Ringbom, H. and Jarvis, S. (2009). The importance of cross-linguistic similarity in foreign language learning. In Long, M. H. and Doughty, C. J., editors, *The Handbook of Language Teaching*, chapter 7, pages 106–118. John Wiley & Sons. 74, 75, 103
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25. 53
- Rus, V., Cai, Z., and Graesser, A. C. (2007). Experiments on generating questions about facts. *Computational Linguistics and Intelligent Text Processing*, 4394:444–455. 21

- Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204. 22
- Russell, R. and Odell, M. (1918). Soundex. US Patent, 1. 88
- Sakaguchi, K., Arase, Y., and Komachi, M. (2013). Discriminative approach to fill-in-the-blank quiz generation for language learners. 25
- Salehi, M. and Sanjareh, H. B. (2013). On the comparability of C-test and cloze: A verbal protocol approach. *English for Specific Purposes World*, 14:1–18. 36
- Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008). Automatic assessment of Japanese text readability based on a textbook corpus. In *6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. 67
- Scherrer, Y. and Erjavec, T. (2015). Modernising historical Slovene words. *Natural Language Engineering*, FirstView:1–25. 104, 178
- Scherrer, Y. and Sagot, B. (2014). A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA). 104, 178
- Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer. 161, 219, 220
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2):129–158. 23
- Schmidt, R. W. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and Awareness in Foreign Language Learning*, pages 1–63. 73
- Scholten-Akoun, D., Mashkovskaya, A., and Tischmeyer, D. (2014). Language competencies of future teachers—design and results of an empirical study. *Applied Linguistics Review*, 5(2):401–423. 34, 50, 151
- Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377. 107, 223

- Schroeders, U., Robitzsch, A., and Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-Tests. *Journal of Educational Measurement*, 51(4):400–418. 37, 54, 130, 179
- Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, pages 523–530. 63, 65, 66, 84
- Sepúlveda Torres, L. and Aluisio, S. M. (2011). Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, pages 67–76. 88, 96
- Settles, B. (2013). Duolingo's data-driven approach to education. Post on blog Duolingo, 31 January. Available at http://duolingo.tumblr.com/post/41960192602/duolingos-data-driven-approach-to-education. last accessed: October 22, 2015. 14
- Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press. 31
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76. 30
- Siddharthan, A. (2014). A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298. 19, 20, 166
- Sigott, G. (1995). The C-test: Some factors of difficulty. AAA. Arbeiten aus Anglistik und Amerikanistik, 20(1):43-54. 36, 37, 38, 44, 83, 141, 220, 221
- Sigott, G. (2006). How fluid is the C-Test construct? In Grotjahn, R. and Sigott, G., editors, *Der C-Test: Theorie, Empirie, Anwendungen / The C-Test: Theory, Empirical Research, Applications*, pages 139–146. Peter Lang. 3, 5, 36, 44, 45, 129, 144
- Sigott, G. and Köberl, J. (1996). Deletion patterns and C-Test difficulty across languages. In Grotjahn, R., editor, *Der C-Test. Theoretische Grundlagen und praktische Anwendungen 3*, pages 159–172. Brockmeyer, Bochum. 39, 40
- Sigurd, B., Eeg-Olofsson, M., and van Weijer, J. (2004). Word length, sentence length and frequency Zipf revisited. *Studia Linguistica*, 58(1):37–52. 64
- Simard, M., Foster, G. F., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81. 88

- Singleton, D. and Little, D. (1991). The second language lexicon: some evidence from university-level learners of French and German. *Second Language Research*, 7:61–81. 36
- Sinha, R. and Mihalcea, R. (2009). Combining lexical resources for contextual synonym expansion. In *Proceedings of the 2009 International Conference Recent Advances in Natural Language Processing*, pages 404–410. Association for Computational Linguistics. 124
- Skory, A. and Eskenazi, M. (2010). Predicting cloze task quality for vocabulary training. In *The 5th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*. Association for Computational Linguistics. 3, 25, 33, 38, 44, 153
- Smith, E. A. and Senter, R. (1967). *Automated readability index*. Cincinnati University Ohio. 63
- Smith, S., Kilgarriff, A., Sommers, S., Wen-liang, G., and Guang-zhong, W. (2009). Automatic cloze generation for English proficiency testing. In *Proceeding of the LTTC International Conference on English Language Teaching and Testing*, Taipei, Taiwan. 24
- Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (\*SEM)*, pages 347–355. Association for Computational Linguistics. 167
- Spolsky, B. (1969). Reduced redundancy as a language testing tool. In Perren, G. and Trim, J., editors, *Applications of linguistics*, pages 383–390. Cambridge University Press, Cambridge. 31, 117
- Stolcke, A. (1994). *Bayesian learning of probabilistic language models*. PhD thesis, University of California, Berkeley. 126
- Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904. 90
- Sumita, E., Sugaya, F., and Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 61–68. Association for Computational Linguistics. 25
- Sun, K.-T., Huang, Y.-M., and Liu, M.-C. (2011). A WordNet-based near-synonyms and similar-looking word learning system. *Journal of Educational Technology & Society*, 14(1):121–134. 23
- Susanti, Y., Iida, R., and Tokunaga, T. (2015). Automatic generation of English vocabulary tests. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, pages 77–78. 22

- Svetashova, Y. (2015). C-Test item difficulty prediction: Exploring the linguistic characteristics of C-tests using machine learning. Master's thesis, Eberhard Karls Universität Tübingen. 39, 46, 56, 149
- Tavakoli, M., Ahmadi, A., and Bahrani, M. (2011). Cloze test and C-test revisited: The effect of genre familiarity on second language reading test performance. *Iranian Journal of Applied Linguistics (IJAL)*, 14(2):173–204. 34, 164
- Taylor, W. L. (1953). "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly*, 30:415–433. 32, 37, 44
- Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57. 27
- Tharp, J. B. (1939). The measurement of vocabulary difficulty. *The Modern Language Journal*, 24(3):169–187. 85
- Tonelli, S., Manh, K. T., and Pianta, E. (2012). Making readability indices readable. In *Proceedings of NAACL-HLT: Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48. 66, 67
- Toutanova, K. and Moore, R. C. (2002). Pronunciation modeling for improved spelling correction. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 105
- Turney, P., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics. 82
- Uitdenbogerd, S. (2005). Readability of French as a foreign language and its uses. In *Proceedings of the Australian Document Computing Symposium*, pages 19–25. 85
- Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173. 19, 65, 66, 72
- Vajjala, S. and Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297. 20

- Vajjala Balakrishna, S. (2015). Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications. PhD thesis, Universität Tübingen. 19, 20, 63
- Verspoor, M. and Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, 53(3):547–586. 81
- Viberg, O. and Grönlund, Å. (2012). Mobile assisted language learning: A literature review. In 11th World Conference on Mobile and Contextual Learning. 12
- Volodina, E. and Pijetlovic, D. (2013). Towards a gold standard for Swedish CEFR-based ICALL. In *2nd Workshop on NLP for Computer-Assisted Language Learning*, pages 48–65. Linköping University Electronic Press, Linköping Universitet. 77
- Vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep syntactic and semantic indicators. In *11th International Multiconference: Information Society*, pages 92–97. 65, 67
- Vossen, P. (1998). A multilingual database with lexical semantic networks. Springer. 100
- Šuster, S., Titov, I., and van Noord, G. (2016). Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of the The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. To appear. 102
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press. 1, 62
- Warschauer, M. and Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(02):57–71. 16
- Webb, S. A. and Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, 43(1):113–126. 76
- Wesche, M. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *Canadian Modern Language Review*, 52:155–78. 14, 15, 16, 17
- Wyse, B. and Piwek, P. (2009). Generating questions from OpenLearn study units. In *AIED* 2009 Workshop Proceedings Volume 1: The 2nd Workshop on Question Generation. 21
- Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* 28, 109

- Yao, X., Bouma, G., and Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2):11–42. 21
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145. 129, 130
- Yin, C., Hirokawa, S., Flanagan, B., Suzuki, T., and Tabata, Y. (2012). Mistake discovery and generation of exercises automaticity in context. In *International Conference on Advanced Applied Informatics*, pages 163–167. 25
- Zesch, T. and Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists—measuring the semantic relatedness of words. *Natural Language Engineering*, 16(1):25–59. 84
- Zesch, T. and Melamud, O. (2014). Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148. Association for Computational Linguistics. 25, 173, 174, 175
- Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications held in conjunction with NAACL 2015.* 30, 140
- Zhang, D. and Koda, K. (2011). Contribution of morphological awareness and lexical inferencing ability to L2 vocabulary knowledge and reading comprehension among advanced EFL learners: testing direct and indirect effects. *Reading and Writing*, 25(5):1195–1216. 83, 87
- Zobel, J. and Dart, P. (1996). Phonetic string matching: lessons from information retrival. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 166–172. 88
- Zobl, H. (1980). Developmental and transfer errors: their common bases and (possibly) differential effects on subsequent learning. *TESOL Quarterly*. 75
- Zweig, G. and Burges, C. J. (2011). The microsoft research sentence completion challenge. Technical report, Technical Report MSR-TR-2011-129 Microsoft. 170, 171
- Zweig, G. and Burges, C. J. (2012). A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 29–36. Association for Computational Linguistics. 7, 47

Zweig, G., Platt, J. C., Meek, C., Burges, C. J., Yessenalina, A., and Liu, Q. (2012). Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 601–610. Association for Computational Linguistics. 126, 127, 128

# **Appendix**

# A Readability Features

For our work on exercise difficulty, we focus on the readability features that have been found to be most predictive in previous work. Complex inter-sentential relations can be neglected because the texts in our data are all very short (three to five sentences). In addition, we apply the constraint that the features should be applicable for English, German and French. We provide a short summary of the implemented features which can be found at: https://github.com/dkpro/dkpro-tc/tree/master/dkpro-tc-features-readability

The input data is already tokenized. For POS-tagging, lemmatization and chunking, the *Treetagger* modules are used with corresponding models for English, German and French (Schmid, 1999). For English and German, named entity recognition is performed using the *Stanford Named Entity Recognizer* (Manning et al., 2014).

#### A.1 Surface

- AvgSentenceLength: Average sentence length in words.
- AvgWordLengthInCharacters: Average word length in characters.
- AvgWordLengthInSyllables: Average word length in syllables. The syllables are determined based on a method implemented by Torsten Zesch in https://github.com/dkpro/dkpro-core/tree/master/de.tudarmstadt.ukp.dkpro.core.readability-asl.

#### A.2 Lexical-Semantic

- TypeTokenRatio: Ratio of types and tokens.
- VerbVariation: Ratio of verb types and verb tokens.

### A.3 Syntactic

- UniqueEntitiesPerSentence: Average number of unique entities per sentence. As the texts contain only very few named entities, all noun chunks are also considered as entities in line with Feng et al. (2009).
- NounChunksPerSentence: Average number of noun chunks per sentence.
- VerbChunksPerSentence: Average number of verb chunks per sentence.
- **PrepositionalChunksPerSentence**: Average number of prepositional chunks per sentence.
- AdverbalChunksPerSentence: Average number of adverbal chunks per sentence.
- SBarsPerSentence: Average number of sbars per sentence.
- ChunksPerSentence: Average number of chunks per sentence.
- NounsPerSentence: Average number of nouns per sentence.
- AdjectivesPerSentence: Average number of adjectives per sentence.
- AdverbsPerSentence: Average number of adverbs per sentence.
- VerbsPerSentence: Average number of verbs per sentence.

#### A.4 Discourse

• PronounRatio: Ratio of pronouns and tokens.

# **B** Word Difficulty Features

For the task of exercise difficulty prediction in chapter 7, we implemented the following word difficulty features. The exercise data is already tokenized. For POS-tagging, lemmatization and chunking, the TreeTagger modules are used with corresponding models for English, German and French (Schmid, 1999). The probability of n-grams is calculated based on the *Web1T* corpus for all three languages (Brants and Franz, 2006). Other resources are specified in the description of the corresponding feature.

### **B.1** Word Familiarity

- WordLengthInCharacters: The length of the word in characters.
- WordLengthInSyllables: The length of the word in syllables. The syllables are determined based on a method implemented by Torsten Zesch in <a href="https://github.com/dkpro/dkpro-core/tree/master/de.tudarmstadt.ukp.dkpro.core.readability-asl">https://github.com/dkpro/dkpro-core/tree/master/de.tudarmstadt.ukp.dkpro.core.readability-asl</a>.
- Polysemy: Number of senses for the word in the resource *UBY* (Gurevych et al., 2012).
- UnigramProbability: The log-probability of the word. For the calculation of the Sigott-baseline in chapter 7, the absolute frequency is used as specified by Sigott (1995).

### **B.2** Morphology and Compounds

- **IsLemma**: The word is in lemma form.
- IsInflectedNoun: The word is a noun and not in lemma form.
- IsInflectedAdjective: The word is an adjective and not in lemma form.
- IsDerivedAdjective: The word is a derived adjective (e.g. understandable).
- IsInflectedVerb: The word is a verb and not in lemma form.
- **IsCompound**: The word is a compound. Compounds are determined using the jword-splitter: https://github.com/danielnaber/jwordsplitter
- IsCompoundBreak: The hint splits the word at a compound break.
- IsSyllableBreak: The hint splits the word within a syllable. Syllables are determined using an implementation of the latex hyphenation rules by Christian M. Meyer and an hyphenation dictionary for English. Christian M. Meyer provides the following references: "The implementation follows Frank M. Liangs hyphenation algorithm (Liang, 1983), which is also an integral part of the TeX typesetting system. Mathias Nater provides a good overview of the algorithm, see: <a href="http://tug.org/tex-hyphen/pdf/hyphenator.pdf">http://tug.org/tex-hyphen/pdf/hyphenator.pdf</a>". The hyphenation dictionary is available at <a href="http://hindson.com.au/info/free/free-english-language-hyphenation-dictionary">http://hindson.com.au/info/free/free-english-language-hyphenation-dictionary</a> (accessed: December 7, 2015).

### **B.3** Syntactic Behavior and Context Specificity

- **TrigramLogProbability**: The log-probability of the micro context. The micro context consists of the left and right neighbor of the word.
- LeftBigramLogProbability: The log-probability of the left bigram of the micro context.
- **RightBigramLogProbability**: The log-probability of the right bigram of the micro context.
- POSProbability: The probability of the POS sequence of the micro context.
- IsFunctionWord: The word is a function word.
- IsADJ, IsAdv, IsArt, IsConj, IsNN, IsNP, IsPP, IsPr, IsV: The word is an adjective / adverb / article / conjunction noun / proper noun/ preposition / pronoun / verb.
- IsADJC, IsADVC, IsNC, IsNN, IsPC, IsSBar, IsVC: The word occurs in an adjective / adverbal / noun / prepositional / sbar / verb chunk. These features are only used for the baseline calculation in chapter 7. Sigott (1995) specified the phrase type in which a word occurs as important feature for difficulty prediction. In our analyses of the training data, we did not find such a correlation and therefore do not use the feature for our own approach.

#### B.4 L1 influence

- **UbyCognate**: The maximum XDice score between the word and the translations of the word retrieved from *UBY* (Gurevych et al., 2012).
- COPCognate: A cognate of the solution exist in a list retrieved from the cognate production algorithm *COP* (Beinborn et al., 2013) and filtered as explained in section 5.2.4. This feature is not available for French.
- IsAcademicWord: The solution can be found in a list of academic words. This feature is only available for English. The word list is merged from two different academic word lists that are recommended for English: the COCA list of academic words (http://www.academicvocabulary.info/samples/general-core.pdf) and the Coxhead list (http://www.victoria.ac.nz/lals/resources/academicwordlist).
- HasLatinRoot: The solution can be found in a list of words with Latin roots (https://en.wikipedia.org/wiki/List\_of\_Latin\_words\_with\_English\_derivative, accessed: December 15, 2015). This feature is only available for English.

### **B.5** Spelling Difficulty

The models for the following features are only trained on the 800-1000 most frequents words of each language. This *Basic Vocabulary* is extracted from existing word lists. <sup>95</sup>.

- CharacterLMprobability: Probability of the character sequence of the word according to a character-based language model trained using the framework *berkeleylm* version 1.1.2 (Pauls and Klein, 2011).
- LMprobabilityOfPrefix: Probability of the character sequence of the prefix. This is only used for C-tests.
- LMprobabilityOfSolution: Probability of the character sequence that needs to be filled in by the participant.
- LMprobabilityOfPostfix: Probability of the character sequence of the postfix. This is only used for X-tests.
- OccursAsText: The word occurs previously in the text (not as a gap).
- PhoneticDifficulty: The phonetic score returned by the grapheme-phoneme alignment algorithm *Phonetisaurus* (http://code.google.com/p/phonetisaurus, accessed: March 12, 2015). The algorithm is trained on words from the *Basic Vocabulary*. The gold graphemes for the English words are obtained from the CMU pronunciation dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed: December 5, 2015)

<sup>&</sup>lt;sup>95</sup>en: http://ogden.basic-english.org

de: http://www.languagedaily.com/learn-german/vocabulary/common-german-words

fr: http://www.101languages.net/french/most-common-french-words

it: https://en.wiktionary.org/wiki/Wiktionary:Frequency\_lists/Italian1000

all accessed: September 17, 2015)

and the transcriptions for the German and Italian words have been queried from the text-to-speech synthesis system *MaryTTS* version 5.1.1 (Schröder and Trouvain, 2003).

• **PhoneticClarity**: The similarity between the predicted pronunciation by *Phonetisaurus* and the true pronunciation as found in the resources described above.

## **C** Candidate Ambiguity Features

The following features have been implemented to assess candidate ambiguity. The frequencies for the unigram, bigram, and trigram candidates are obtained from the *Web1T* corpus for all three languages Brants and Franz (2006). The thresholds are set to zero for our experiments, but the value could be tuned to the learner group.

#### C.1 Micro-level fitness

- NrOfCandidates: The number of candidates in the vocabulary. As vocabulary for the three languages, we use the word list package provided by Ubuntu for spell-checking (http://packages.ubuntu.com/de/precise/text/).
- NrOfUnigramCandidates: The number of candidates with a probability above a given threshold.
- NrOfBigramCandidates: The number of candidates for which the bigram probability of the candidate and the word left to the gap is above a threshold.
- NrOfTrigramCandidates: The number of candidates for which the trigram probability of the micro context (the candidate with the left and right neighbor) is above a threshold.
- **UnigramSolutionRank**: The rank of the solution within the unigram candidates that are ranked by descending probability.
- **BigramSolutionRank**: The rank of the solution within the bigram candidates ranked by descending probability.
- **TrigramSolutionRank**:The rank of the solution within the trigram candidates ranked by descending probability.
- MaxStringSimilarityWithCandidate: The maximum string similarity between the solution and each candidate based on the longest common subsequence ratio (LCSR).

#### C.2 Macro-level fitness

• LmRankOfSolution: The rank of the solution within the candidates that are ranked by descending log-probability of the language model for the full sentence in which the gap is replaced with the candidate. See the previous section for a more detailed description.

• EsaRankOfSolution: The rank of the solution within the candidates that are ranked by descending relatedness of the candidate with the sentence. Relatedness is calculated based on the ESA-index as described in the previous section. This feature is only applicable to the cloze dataset because it does not work for function words.

## **D** Item Dependency Features

The following features have been implemented to account for item dependency. The input data is already tokenized and the gaps are marked. The probability of n-grams is calculated based on the *Web1T* corpus for all three languages (Brants and Franz, 2006).

#### D.1 Position

- PositionOfGap: The position of the gap (indicated as the number of tokens).
- NrOfGapsInSentence: The number of gaps in the same sentence as the current gap.
- NrOfPrecedingGapsInSentence: The number of gaps that precede the current gap in the same sentence.
- NrOfPrecedingGaps: The number of gaps that precede the current gap in the document.

### D.2 Neighbor Effects

- LeftTrigramLogProbability: The log-probability of the trigram that spans over the gap and two words left to the gap. With a deletion rate of two as in our data, this trigram includes the previous gap.
- **RightTrigramLogProbability**: The log-probability of the trigram that spans over the gap and two words right to the gap. With a deletion rate of two as in our data, this trigram includes the following gap.
- OccursAsGap: The same words occurs as a gap somewhere else in the text.

### D.3 Referentiality

• **IsReferentialGap**: The prefix consists of *th*. This feature is only available for English.

# Wissenschaftlicher Werdegang der Verfasserin<sup>1</sup>

2005-2008	Bachelor of Science, Hauptfach: Computerlinguistik Nebenfach: Psychologie (an der UA Barcelona, Spanien) Robot Engagement Behaviors in the Opening Phase of a Conversation Gutachter: Prof. Dr. Manfred Pinkal, Dr. Geert-Jan Kruijff Universität des Saarlandes
2008-2010	Master of Science, Hauptfach: Language Science and Technology Nebenfach: Informatik (an der FU Bolzano, Italien) Post-Editing of Statistical Machine Translation – A crosslinguistic analysis of the temporal, technical and cognitive effort Gutachter: Prof. Dr. Hans Uszkoreit, Dr. Pirita Pyykkönen Universität des Saarlandes
2010-2011	Stipendiatin in der Forschungsgruppe: Cross-Modal Interactions in Natural and Artificial Cognitive Systems, Universität Hamburg
2012-2015	Promotionsstipendium der Klaus-Tschira-Stiftung am Ubiquitous Knowledge Processing Lab, TU Darmstadt
2014-2015	Fernstudium <i>Grundlagen und Konzepte des DaF-Unterrichts</i> vom Goethe-Institut in Kooperation mit der LMU München
seit September 2015	Wissenschaftliche Mitarbeiterin am Ubiquitous Knowledge Processing Lab, TU Darmstadt

# Ehrenwörtliche Erklärung<sup>‡</sup>

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Doktor der Naturwissenschaften" mit dem Titel "Predicting and Manipulating the Difficulty of Text-Completion Exercises for Language Learning" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 5. Juli 2016

Lisa Marina Beinborn, M.Sc.

 $<sup>\</sup>P_{\mbox{Gem\"{a}}\mbox{\sc }\S\mbox{\sc }20\mbox{\sc }\mbox{\sc }Abs.\mbox{\sc }3\mbox{\sc }\mbox{\sc }\mbox{$ 

<sup>‡</sup> Gemäß § 9 Abs. 1 der Promotionsordnung der TU Darmstadt

## Publikationsverzeichnis der Verfasserin

- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Predicting the Spelling Difficulty of Words for Language Learners, in: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications at NAACL 2016*: S. 73–83, San Diego, California, USA, Juni 2016.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Factors of Difficulty in German Language Proficiency Tests, in: *Book of Abstracts: Language, Learning, Technology Conference*, S. 24-25, Lüneburg, November 2015.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Candidate Evaluation Strategies for Improved Difficulty Prediction of Language Tests, in: *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications at NAACL 2015*: S. 1–11, Denver, Colorado, USA, Juni 2015.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Predicting the Difficulty of Language Proficiency Tests, in: *Transactions of the Association for Computational Linguistics (TACL)*, Vol. 2 (1), S. 517–529, November 2014.
- Lisa Beinborn: Multilingual Corpora and Multilingual Corpus Analyses (Book Review), in: *International Journal of Multilingualism*, Vol. 11 (2), S. 266-268, April 2014.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Readability for foreign language learning: The importance of cognates, in: *International Journal of Applied Linguistics*, Vol. 165 (2), S. 136–162, 2014.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Cognate Production using Character-based Machine Translation, in: *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, S. 883-891, Nagoya, Japan, Oktober 2013.
- Lisa Beinborn and Torsten Zesch and Iryna Gurevych: Towards fine-grained readability measures for self-directed language learning, in: *Proceedings of the Swedish Language Technology Conference: Workshop on NLP for CALL*, Vol. 80 (2), S. 11-19, Lund, Schweden, Oktober 2012.
- Lisa Beinborn: *Post-Editing of Statistical Machine Translation A crosslinguistic analysis of the temporal, technical and cognitive effort,* Masterarbeit, Universität des Saarlandes, September 2010.
- Lisa Beinborn: *Robot Engagement Behaviors in the Opening Phase of a Conversation*, Bachelorarbeit, Universität des Saarlandes, September 2008.