An Empirical Analysis of Linguistic, Typographic, and Structural Features in Simplified German Texts

Alessia Battisti **Sarah Ebling** **Martin Volk**
Institute of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
alessia.battisti@uzh.ch, {ebling|volk}@cl.uzh.ch

Abstract

English. We investigate a newly compiled corpus of simplified German texts for evidence of multiple complexity levels using unsupervised machine learning techniques. We apply linguistic features used in previous supervised machine learning research and additionally exploit structural and typographic characteristics of simplified texts. The results show a difference in complexity among the texts investigated, with optimal partitioning solutions ranging between two and four clusters. They demonstrate that both linguistic and structural/typographic features are constitutive of the clusters.

Italiano. *Esaminiamo un nuovo corpus di testi in tedesco semplificato per cercare delle evidenze relative a molteplici livelli di complessità utilizzando tecniche di apprendimento automatico non supervisionato. Applichiamo variabili linguistiche utilizzate in precedenti ricerche con apprendimento automatico supervisionato e sfruttiamo inoltre le caratteristiche strutturali e tipografiche dei testi semplificati. I risultati mostrano una differenza di complessità tra i testi analizzati, con suddivisioni ottimali variabili da due a quattro cluster. Ciò dimostra che sia le caratteristiche linguistiche sia quelle strutturali/tipografiche sono costitutive dei cluster.*

1 Introduction

Simplified language aims at providing comprehensible information to persons with reduced reading abilities. This group includes persons with cognitive impairment and learning disabilities, prelingually deaf persons, functionally illiterate persons, and foreign language learners (Bredel and Maaß, 2016). Simplified language is characterised by reduced lexical and syntactic complexity and includes images, structured layout, and explanations of difficult words. For simplified German, several guidelines exist that define which structures need to be avoided, which need to be paraphrased, and which are comprehensible (Bundesministerium für Arbeit und Soziales, 2011; Inclusion Europe, 2009; Maaß, 2015; Netzwerk Leichte Sprache, 2013).

Various countries have acknowledged simplified language as a means of inclusion that enables the target populations mentioned above to inform themselves of their legal rights and participate in society. German-speaking countries have been promoting simplified language only in the last years, in particular since the ratification of the United Nations Convention on the Rights of Persons with Disabilities (United Nations, 2006) in Austria (2008), Germany (2009), and Switzerland (2014). As a result, large amounts of texts in simplified German have become available.

More recently, simplified German has been conceptualised as a construct with multiple complexity levels (Bock, 2014; Bredel and Maaß, 2016; Kellermann, 2014). However, these proposals are merely theoretical: They are not yet operationalised, i.e., no sets of guidelines exist that distinguish the proposed levels with reference to linguistic or other features. The social franchise network *capito*,[1] a provider of simplification services as well as training courses for simplified language translators, recognises three levels of simplified German corresponding to the Common European Framework of Reference for Language (CEFR)

[1] https://www.capito.eu/ (last accessed: June 27, 2019)

(Council of Europe, 2001) levels A1, A2, and B1. Being commercially orientated, *capito* does not make its CEFR adaptation publicly available.

In this paper, we present an unsupervised machine learning (clustering) approach to analysing texts in simplified German with the aim of investigating evidence of multiple complexity levels. To the best of our knowledge, this is the first study of its kind. We apply linguistic features used in previous supervised machine learning research (classification) and additionally exploit structural and typographic characteristics of simplified texts that have been described in the literature but not incorporated into clustering and/or classification approaches in the context of simplified language.

The remainder of this paper is structured as follows: Section 2 presents the research background. Section 3 describes our approach, introducing a novel dataset (Section 3.1), the feature design and engineering (Section 3.2), the clustering experiments (Section 3.3), and a discussion thereof (Section 3.4). Section 4 offers a conclusion and an outlook on future research questions.

2 Research Background

Two natural language processing tasks deal with the concept of simplified language: automatic readability assessment and automatic text simplification. Readability assessment refers to the process of determining the level of difficulty of a text. Traditionally, this has involved taking into account readability measures based on surface features such as the number of syllables in a word or number of words in a sentence, e.g., via the Flesch Reading Ease Score (Flesch, 1948). Recently, more sophisticated models employing deeper linguistic features such as lexical, semantic, morphological, morphosyntactic, syntactic, pragmatic, discourse, psycholinguistic, and language model features have been proposed (Collins-Thompson, 2014; Dell'Orletta et al., 2014; Heimann Mühlenbock, 2013; Schwarm and Ostendorf, 2005).

Readability assessment implies the existence of multiple complexity levels. Complexity levels are identified, e.g., along school grades or levels of the CEFR (Hancke, 2013; Pilan and Volodina, 2018; Reynolds, 2016; Vajjala and Lõo, 2014).

The work presented in this paper represents a preliminary stage of the readability assessment task for simplified German in that it investigates empirically whether different complexity levels exist in previous German simplification practice in the first place.

3 Clustering Simplified German texts

3.1 Dataset

Battisti and Ebling (2019) compiled a corpus of German/simplified German texts for use in automatic readability assessment and automatic text simplification. The corpus represents an enhancement of a parallel (German/simplified German) corpus created by Klaper et al. (2013). Compared to its predecessor, the corpus of Battisti and Ebling (2019) contains additional parallel data and newly contains monolingual-only data as well as structural and typographic information.

The authors collected PDFs and web pages from 92 different domains of public offices, translation agencies, and organisations publishing content in German and simplified German. Overall, the corpus consists of 6,217 documents (378 parallel and 5,461 monolingual). Metadata was recorded in the Open Language Archives Community (OLAC) Standard[2] and converted into the metadata standard CMDI of CLARIN, a European research infrastructure for language resources and technology.[3] If available, information on the language level of a simplified German text (typically A1, A2, or B1) was stored in the metadata. 52 websites and 233 PDFs (amounting to approximately 26,000 sentences) have an explicit language level label.

Linguistic annotation was added automatically using ParZu (Sennrich et al., 2009) (for tokens and dependency parses), NLTK (Bird et al., 2009) (for sentence segmentation), TreeTagger (Schmid, 1995) (for part-of-speech tags and lemmas), and Zmorge (Sennrich and Kunz, 2014) (for morphological units). In addition, information on text structure (e.g., paragraphs, lines), typography (e.g., boldface, italics), and images (content, position, and dimensions) was added. The annotations were stored in the Text Corpus Format by WebLicht (TCF) developed as part of CLARIN.[4]

For the experiments reported in this paper, we

[2] http://www.language-archives.org/OLAC/olacms.html (last accessed: June 27, 2019)

[3] https://www.clarin.eu/ (last accessed: June 27, 2019)

[4] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/TheTCFFormat (last accessed: June 27, 2019)

considered the monolingual documents of the corpus, i.e., the monolingual-only documents as well as the simplified German side of the parallel data. This amounted to 5,839 texts (193,845 sentences).

3.2 Features

In addition to constituting the first approach to investigating simplified German texts using unsupervised machine learning, the unique contribution of this paper consists of leveraging information that has been shown to be characteristic of simplified language (Arfé et al., 2018; Bock, 2018; Bredel and Maaß, 2016) but has not been incorporated into machine learning approaches involving simplified language. Specifically, we considered features derived from text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and image (content, position, and dimensions) information.

In a simplified text, typographical information, such as boldface and italics, serves as a discourse marker signalling words and phrases that require particular attention and convey different purposes (Arfé et al., 2018). Leveraging the concepts of multi-modality and multi-codality in the psychology of perception (Schnotz, 2014), images[5] are supposed to support the text by activating previous knowledge and exemplifying the objects in the text (Bredel and Maaß, 2016).

Subset	Features	Number
1	All	115
2	Surface	26
3	Deeper	89
4	Lexical + semantic	17
5	Morphological + syntactic	72

Table 1: Subsets of feature combinations.

Altogether, the feature set comprised 115 features arranged into five feature groups, as shown in Table 1. Subset 3 ("Deeper") consisted of lexical, semantic, morphological, and syntactic features. "Surface" is short for surface, structural, and typographic features.

Surface, structural, and typographic features: We took advantage of the structural and typographic information included in the corpus (cf. Section 3.1) and introduced as features the number of images, paragraphs, lines, words of a specific font type and style, and adherence to a one-sentence-per-line rule. We additionally included the number of digits and numbers in words (Saggion, 2017), number of abbreviations and initial letters, and the number of individual punctuation marks and special characters. Among the special characters was the *Mediopunkt* ('centred dot'), a typographic device proposed by Maaß (2015) for visually segmenting compound words. We also computed the *Läsbarhetsindex* ('readability index', LIX) (Björnsson, 1968).[6]

Lexical and semantic features: This group included features for lexical richness, lexical variation (e.g., nominal ratio, noun/pronoun ratio, bilogarithmic TTR (Vajjala and Meurers, 2012)), word frequency based on the German reference corpus DeReKo (Lüngen, 2017), and lists of words classified at different perceptive levels (Glaboniat et al., 2005). We also included question words and named entities, which may strain the reading comprehension process if the target reader does not have the appropriate knowledge.

Morphological, morphosyntactic, and syntactic features: In this group, we included particles, prepositions, demonstrative and personal pronouns, and (separately) first-, second-, and third-person pronouns. We additionally counted adverbs, modal verbs, subjunctions, and conjunctions. We added genitive attributes in relation to *von*+dative constructions.[7] We additionally included the number of negative forms, the presence of pre- and post-modifiers, and impersonal constructions. We took advantage of the verbal morphology and included verbal mood- and tense-based features (Dell'Orletta et al., 2011). We also considered direct vs. indirect speech constructions, the types of subordinate clauses as well as features based on word and sentence order.

[5] For the sake of simplicity, the term "images" here subsumes pictures, pictograms, photographs, graphics, and maps.

[6] LIX = N_w / N_s + (W x 100)/N_w, where N_w is the number of words, N_s is the number of sentences, and W is the percentage of tokens longer than six characters.

[7] In German, the genitive attribute can be substituted by a *von*+dative construction. Importantly, this is a case of simplified German conflicting with the grammar of Standard German, which encourages the use of the former construction.

3.3 Experiments and Results

3.3.1 Method

We applied agglomerative hierarchical clustering. We used the `scipy`[8] toolkit alongside with models recursively created with the `scikit-learn`[9] library. The data matrix was created using the cosine similarity metric and the average linkage function. Because of the significant variation in length of the documents, we normalised the features by dividing the values by the length of each document expressed in tokens. We then performed principal component analysis (PCA) to diminish the sparseness of the data matrix and avoid the curse-of-dimensionality trap. In a second experiment, we applied feature agglomeration instead of PCA prior to clustering. Feature agglomeration allows for a straightforward interpretation of the results.

Given the lack of a ground truth for our data, we evaluated the experiments using the following metrics: silhouette score, Calinski-Harabasz index, and Elbow method. These metrics were also used to choose the optimal number of clusters.

3.3.2 Results

Table 2 shows the results of the first three iterations of our clustering approach after the feature agglomeration step. We observed that a value between 2 and 4 (inclusive) represented a good clustering solution for the whole corpus according to the metrics. A dendrogram corroborated these results (cf. Figure 1).

Upon inspection of the clusters, we found the main differences to be due to the following features: number of nouns, number of verbs, number of paragraphs, adherence to one-sentence-per-line rule, number of interrogative clauses, number of different fonts, and number of words in bold. Considering the mean ratio of the features in a two-cluster solution, Cluster 1 displayed a higher frequency of nouns (0.31 vs. 0.24) and adjectives (0.9 vs. 0.6) and a lower frequency of verbs (0.13 vs. 0.17) than Cluster 2, which in turn included a slightly higher rate of images (0.008 vs. 0.004).

3.4 Discussion

The inverse proportion of the mean ratios concerning nouns and verbs (cf. Section 3.3.2) suggested that Cluster 1 included texts focusing on objects or concepts, since verbs (events, actions, etc.) had been turned into nouns (concepts, things, etc.) following the linguistic process of nominalisation, while the linguistic structure of texts in Cluster 2 was simpler.

Figure 2 visualises the box plots of six of the surface features of Subset 2 (number of full stops, number of commas, adherence to one-sentence-per-line rule, number of paragraphs, number of different fonts, number of images) based on the three-cluster solution suggested by the agglomerative hierarchical approach. The first cluster consisted of texts that followed the one-sentence-per-line rule, featured a low frequency of commas, and a high number of paragraphs. These characteristics are crucial properties of simplified texts. Our findings further emphasise the importance of distinguishing among different types of punctuation marks in the context of simplified language: while for commas, a low frequency is indicative of textual simplicity, the reverse is true for full stops. Texts included in Cluster 1 did not contain images. This outcome relates to the results of a more recent study by Bock (2018), according to which images should be used with caution even in simplified German texts to avoid the potential of distraction and cognitive overload.

4 Conclusion and Outlook

In this paper, we have presented the first approach to investigating simplified German texts by means of unsupervised machine learning techniques as a basis for future readability assessment studies on this language variety. In addition, we have introduced novel features that have been described in the literature but not incorporated into machine learning (clustering and/or classification) approaches in the context of simplified language, notably: number of images, number of paragraphs, number of lines, number of words of a specific font type, and adherence to a one-sentence-per-line rule. Our findings provide evidence that existing texts are not simplified at a unique complexity level of German. We have demonstrated that features based on structural information are capable of accounting for the different complexity levels found.

As a next step, we will use the results of the experiments presented in this paper to establish a framework of inductively generated complexity

[8] `https://www.scipy.org/` (last accessed: June 27, 2019)

[9] `https://scikit-learn.org/stable/` (last accessed: June 27, 2019)

	Subset 1		Subset 2		Subset 3		Subset 4		Subset 5	
	Sil	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**	**Sil**	**CH**
2	**0.601**	**3867.1**	0.373	1135.2	**0.675**	**5214.2**	**0.693**	**3593.9**	**0.695**	**5463.2**
3	0.532	2476.2	0.372	1266.3	0.617	3329.5	0.55	1824.8	0.572	3273.9
4	0.456	1698.3	**0.493**	**1417.6**	0.592	2572.7	0.505	1248.9	0.51	2517.8

Table 2: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) after feature agglomeration on all data samples.

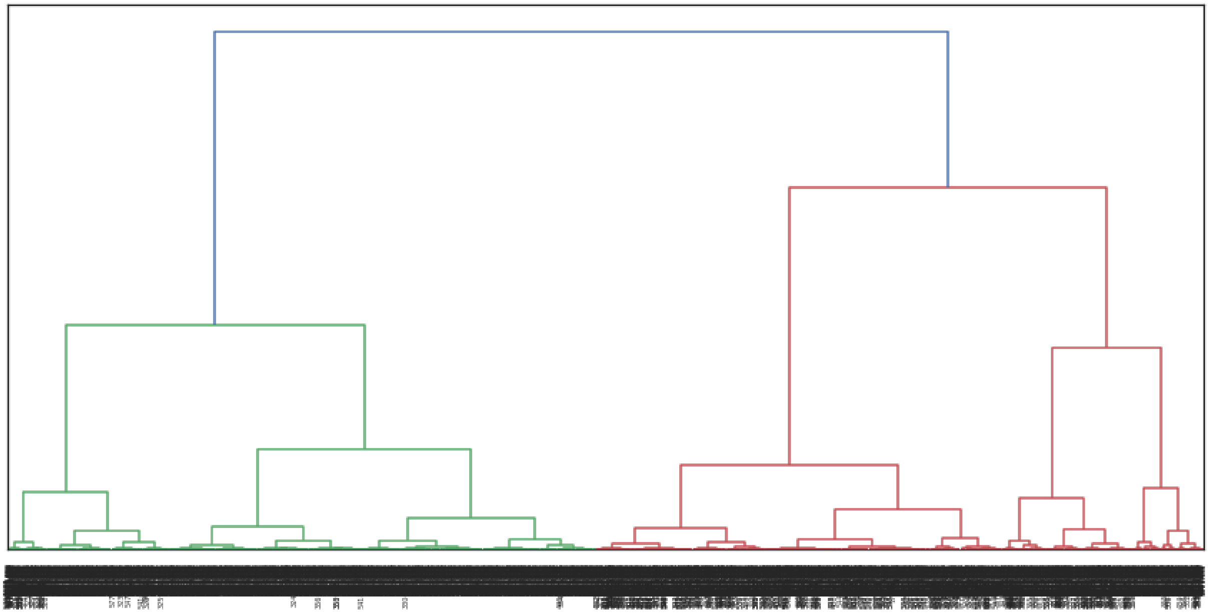

Figure 1: Dendrogram of the texts considering agglomerated features of Subset 1.

levels. This framework will serve as the basis for readability assessment in the context of simplified German. Knowledge derived from our study can also inform automatic and manual approaches to simplification of German.

References

Barbara Arfé, Lucia Mason, and Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210.

Alessia Battisti and Sarah Ebling. 2019. A corpus for automatic readability assessment and text simplification of german. arXiv:1909.09067.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

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

Bettina M. Bock. 2014. "Leichte Sprache": Abgrenzung, Beschreibung und Problemstellungen aus Sicht der Linguistik. *Sprache barrierefrei gestalten*, pages 17–51.

Bettina M. Bock. 2018. "Leichte Sprache" - Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt. Technical report, Universität Leipzig.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Duden, Berlin.

Bundesministerium für Arbeit und Soziales. 2011. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung-BITV 2.0). Technical Report Teil 1.

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

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

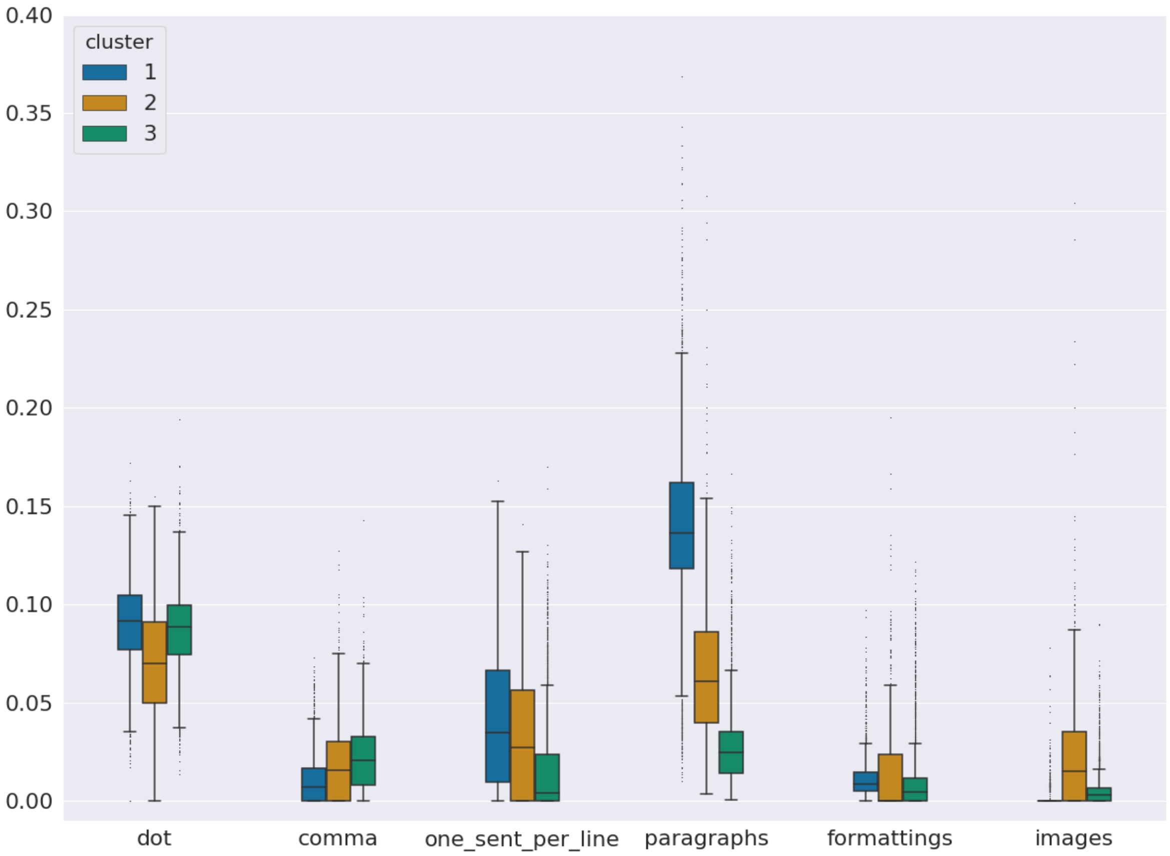

Figure 2: Six features of Subset 2.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Baltimore, Maryland, June. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Manuela Glaboniat, Martin Müller, Paul Rusch, Helen Schmitz, and Lukas Wertenschlag. 2005. *Profile Deutsch*. Klett Langenscheidt, Berlin/Munich, Germany.

Julia Hancke. 2013. Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master's thesis, University of Tübingen, Germany.

Katarina Heimann Mühlenbock. 2013. *I see what you mean: Assessing readability for specific target groups*. Ph.D. thesis, University of Gothenburg.

Inclusion Europe. 2009. Information für alle: Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht. Technical report, Inclusion Europe.

Gudrun Kellermann. 2014. Leichte und Einfache Sprache Versuch einer Definition. In *Aus Politik und Zeitgeschichte*, volume 64, pages 9–11.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *ACL Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Harald Lüngen. 2017. DEREKO - Das Deutsche Referenzkorpus. *Zeitschrift fur Germanistische Linguistik*.

C. Maaß. 2015. *Leichte Sprache: Das Regelbuch*. Barrierefreie Kommunikation. Lit Verlag.

Netzwerk Leichte Sprache. 2013. Die Regeln für Leichte Sprache. Technical report.

Ildiko Pilan and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings ofthe Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, California.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL'95 SIGDAT Workshop*, pages 47–50, Dublin, Ireland.

Wolfgang Schnotz, 2014. *An Integrated Model of Text and Picture Comprehension*, pages 72–103. Cambridge University Press, second edition.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, pages 523–530.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1063–1067, Reykjavik, Iceland. European Language Resources Association.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In *Proceedings of the Biennal GSCL Conference*, pages 115–124, Potsdam.

United Nations. 2006. Convention on the Rights of Persons with Disabilities and Optional Protocol.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, volume 107, pages 113–127, Uppsala, Sweden.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th workshop on building educational applications using NLP*, pages 163–173, Montral, Canada.