Computational Linguistics in the Netherlands Journal 8 (2018) 24–48 Submitted 06/2018; Published 12/2018

Automating lexical simplification in Dutch

Bram Bulté[*] BULT@CCL.KULEUVEN.BE
Leen Sevens[*] LEEN@CCL.KULEUVEN.BE
Vincent Vandeghinste[**] VINCENT.VANDEGHINSTE@IVDNT.ORG

[*] *Centre for Computational Linguistics, KU Leuven, Belgium*

[**] *Instituut voor de Nederlandse Taal*

Abstract

We discuss the design, development and evaluation of an automated lexical simplification tool for Dutch. A basic pipeline approach is used to perform both text adaptation and annotation. First, sentences are preprocessed and word sense disambiguation is performed. Then, the difficulty of each token is estimated by looking at their average age of acquisition and frequency in a corpus of simplified Dutch. We use Cornetto to find synonyms of words that have been identified as difficult and the SONAR500 corpus to perform reverse lemmatisation. Finally, we rely on a large-scale language model to verify whether the selected replacement word fits the local context. In addition, the text is augmented with information from Wikipedia (word definitions and links). We tune and evaluate the system with sentences taken from the Flemish newspaper De Standaard. The results show that the system's adaptation component has low coverage, since it only correctly simplifies around one in five 'difficult' words, but reasonable accuracy, with no grammatical errors being introduced in the text. The Wikipedia annotations have a broader coverage, but their potential for simplification needs to be further developed and more thoroughly evaluated.

1. Introduction

The current period in human history is often called the Digital Age or Information Age (Castells 1996). One of its characteristics is the continuous proliferation of information, often in written form. In the interest of democracy and also the economy, it is important to promote inclusion by making this information accessible to as many people as possible.[1] Accessibility in this context does not only refer to having physical or digital access to information sources, but also to being able to process and understand the provided information. Not everyone has the same level of proficiency in any given language, for a variety of reasons (e.g. intellectual disability, language pathology or autism, non-native speaker). At the same time not all texts have the same level of (linguistic) complexity, which influences the difficulty people have in understanding them.

In recent years, a considerable amount of effort has been put into exploiting one of the other characteristics of the current era, namely a steady increase in computerised automation, to try and tackle this problem. Different techniques for automated text simplification have been developed, focusing either on grammatical or lexical aspects of texts, or both (Saggion 2017, Siddharthan 2014). In essence, all of these techniques comprise two stages: (a) the identification of complex textual features, and (b) replacing these features with simpler ones. The ultimate aim is always to adapt texts in such a way that they become easier to read and understand, while maintaining the original meaning as much as possible. Most of these methods have focused on English, even though work has also been done on languages such as French (Brouwers et al. 2014), Spanish (Saggion et al. 2015), Italian (Barlacchi and Tonelli 2013), Portuguese (Candido Jr. et al. 2009) and Japanese (Inui et al. 2003). Alternatively, texts can be augmented with additional information, such as dictionary definitions of words, or links to other relevant sources of information (Brank et al. 2017, Kandula et al. 2010).

[1] See, e.g. United Nations (1994); http://nos.nl/artikel/2166447-laaggeletterdheid-kost-1-miljard-euro.html; http://www.wablieft.be/wablieft/over-wablieft/waarom-is-duidelijke-taal-belangrijk-

This paper explores a basic technique for lexical simplification in Dutch. To our knowledge, research on textual simplification in Dutch has so far only focused on syntactic compression for subtitles (Daelemans et al. 2004, Vandeghinste and Pan 2004). Recently, methods have also been developed that go one step further and transform text into a series of pictographs, a process which involves syntactic simplification as an intermediate step (Sevens et al. 2018). The technique discussed here involves text-to-text simplification. We develop a *general* lexical simplification system that is not targeted towards a particular audience, but whose parameters and settings can be tuned to the needs of a specific target population (e.g. people with low literacy or second language learners). Our main aim is to establish a baseline for lexical simplification in Dutch. The system we develop makes use of a pipeline approach, in the context of which words identified as difficult are replaced by simpler synonyms. In addition, it includes a component for text annotation (with word definitions and links to Wikipedia pages).

The paper is structured as follows: Section 2 provides background information on reasons for and approaches to text simplification. This is followed by an overview of the system setup (Section 3) and experimental design (Section 4). Sections 5 and 6 contain the results and discussion. Conclusions are drawn up in Section 7.

2. Background

In this section we provide a short overview of relevant literature on text simplification. We look at different reasons why simplifying texts can be useful, discuss which (linguistic) features make a text simple or complex, and present a number of automated methods for text simplification that have been discussed in the literature. Finally, we look at how text simplification systems can be evaluated. The issues covered here relate to text simplification in general, but where necessary we focus on more specific lexical aspects.

2.1 Why text simplification?

Learning how to read is a complex process that takes several years (Rayner et al. 2001). Even though in our society being able to read is often taken for granted, not everyone attains or possesses the same level of proficiency in this skill. This can have negative consequences for a person's social inclusion, employment status and even health (Davis et al. 2006). Just think of the problems people might have communicating through new media, following the news, or reading instructions on prescription labels, to provide some examples. Several (private[2] or public[3]) initiatives exist to try and keep written information as clear and simple as possible or to offer simplified alternatives (e.g. Wikipedia in simple English[4]), but many texts remain difficult to read and understand for certain audiences. For this reason it is useful to dispose of automated tools to simplify language. Such language simplification systems have been developed with different target audiences in mind, such as children (De Belder and Moens 2010), second language learners (Medero and Ostendorf 2011, Petersen and Ostendorf 2007), deaf (Chung et al. 2013, Inui et al. 2003) and blind people (Grefenstette 1998), aphasics (Carroll et al. 1999, Devlin and Unthank 2006), and people with autism (Barbu et al. 2015, Orăsan et al. 2018), dyslexia (Matausch and Peböck 2010, Rello et al. 2013) and low literacy (Aluísio et al. 2008, Specia 2010, Watanabe et al. 2009, Williams et al. 2003). The initiatives taken in the context of the European Commission funded Able to Include project[5], offering tools to promote the integration of persons with intellectual disabilities (such as services for text to pictograph translation and text simplification), are very relevant in this context as well.

[2] See e.g. the Plain English Campaign (http://www.plainenglish.co.uk).

[3] For example, guideline 1 of the Joint Practical Guide of the European Parliament, the Council and the Commission for persons involved in the drafting of European Union legislation reads: “Legal acts of the Union shall be drafted clearly, simply and precisely”.

[4] https://simple.wikipedia.org/wiki/Main_Page

[5] http://able-to-include.com/

Text simplification usually targets either syntactic or lexical elements, or both (Saggion 2017). The fact that lexical simplification is a useful tool for improving the comprehensibility of texts, is illustrated by psycholinguistic research on reading and vocabulary knowledge in a second language learning context (Hirsh and Nation 1992, Nation 2001). This research has shown that approximately 95% of the word types in a text need to be known by a reader to attain a basic level of understanding. This percentage is even higher if a more thorough comprehension, or a more pleasant reading experience, is required or preferred. By simplifying the vocabulary that is used in a text, the percentage of words that readers are familiar with can be increased. Moreover, it has been shown that lexical complexity has an effect on sentence processing (Cutler 1983), fixation times in reading (Rayner and Duffy 1986) and the comprehension of scientific texts (Arya et al. 2011).

2.2 What makes language simple or complex?

It is not easy to define what makes a text simple or complex. There is some general consensus that certain language varieties are ‘simple’ or ‘simplified’ (Siddharthan 2014), such as the language caregivers use when addressing infants or young children (*motherese*), the language variety used by (beginning) second language learners (*interlanguage*), the hybrid languages that developed in situations of language contact, when no common language was available (*pidgins*), or varieties of language that are produced with the specific intention to be clear and unambiguous, for example for use in technical manuals (*controlled language*). It is clear that these language varieties are all characterised by a reduced lexical and grammatical repertoire. However, defining what exactly constitutes language complexity and which linguistic features are more complex than others, is not such a straightforward task (DeKeyser 2005).

When it comes to complexity and language, it is important to distinguish between absolute, objective complexity on the one hand, and relative, subjective complexity, or difficulty on the other (Dahl 2004, Miestamo et al. 2008). Quite a lot of research in typological and diachronic linguistics has focused on the question which (formal, objective) linguistic features make one language more simple or complex than another, and how languages evolve in this respect (Hawkins 2004, Szmrecsanyi and Kortmann 2009, Lupyan and Dale 2010, McWhorter 2011, Trudgill 2011). In this context, complexity refers mainly to quantifiable aspects of language, in terms of numbers of features, elements, dependencies, etc. Objective complexity can also be applied to specific linguistic features, which are then analysed, for example, in terms of their compositionality or length, the mapping between form and meaning, or the number of hierarchical relationships they exhibit (Pallotti 2015). Difficulty is more of a psycholinguistic construct, which, broadly speaking, refers to the mental ease or difficulty with which language features are processed and/or acquired (Byrnes and Sinicrope 2008, Diessel 2004, Hulstijn and de Graaff 1994). It has been observed that some features of language are acquired later than others, both in first and second language acquisition, and that certain linguistic elements require more cognitive load to process (Pienemann 1998). For the purpose of automated text simplification, the ultimate interest is, most commonly at least, in language difficulty, since the aim is to make a text easier to process for readers. In this context, *complex* has to be understood as *difficult to understand*. This notion is closely related to research on text readability, which has a long-standing tradition in itself (Flesch 1948, Klare 1976). The assessment of how readable a text is has recently been automated for Dutch (Daelemans et al. 2017, De Clercq and Hoste 2016). The lexical and semantic features used in this readability assessment system are informative in the context of lexical simplification as well.

Attempts have been made to determine what makes certain lexical items more difficult than others (Rayner and Duffy 1986, Wilkens et al. 2014). Such investigations are not only interesting from a theoretical point of view, but the insights they offer can guide practical applications, such as automated text simplification, as well. For example, automatically or manually composed purpose-built word lists have been used to identify difficult words in a text (Biran et al. 2011, Deléger et al. 2013, Yatskar et al. 2010), but most typically information from some other resource is used as proxy for lexical difficulty. Frequency lists are a popular choice (Devlin and Tait 1998, Siddharthan 2014), and it has been shown that word frequency is indeed a good predictor of difficulty (Wilkens et al. 2014). An important aspect here is the choice of corpus for compiling the frequency list (Wrobel 2016). Next to word frequency (or *unigram probability*), a number of linguistic criteria have been used as well, such as length in characters, syllables and morphemes, the consistency of the relationship between script and speech sounds, and the (automatically tagged) part-of-speech of the word (Davoodi et al. 2017, Gala et al. 2013). Also the ambiguity of a word form, as measured by its polysemy, is sometimes used as an indication of word difficulty (Walker et al. 2011). Recently, these resources have been supplemented with psycholinguistic sources of information, such as the quantitative results of studies investigating the perceived concreteness and imagery level of words, as well as reported familiarity levels and mean age of acquisition (Jauhar and Specia 2012). It has also been shown that taking into account information about the surrounding words (in terms of their frequency, length etc.) can further improve the accuracy of difficult word identification (Davoodi et al. 2017).

2.3 Approaches to automated text simplification

Current text simplification tools can be classified into two broad categories: (a) holistic systems that approach the problem as a monolingual translation task and make use of machine translation tools, and (b) handcrafted systems that tackle specific grammatical phenomena and/or individual lexical items (Siddharthan 2014). The first type of tools requires a, preferably large and aligned, corpus of original and simplified sentences or texts, and applies machine learning techniques to automatically extract translation (i.e. simplification) rules. For English, articles from Wikipedia and their Simple English counterparts have been used as training corpus (Xu et al. 2015, Zhu et al. 2010). Many of these systems (Coster and Kauchak 2011, Specia 2010, Wubben et al. 2012) make use of the phrase-based machine translation tool Moses (Koehn et al. 2007). Other researchers worked with syntax-based machine translation systems (Zhu et al. 2010). More recently, neural machine translation models have been applied to text simplification as well (Nisioi et al. 2017, Wang et al. 2016). Since datasets of simplified language are not commonly available for languages other than English, the translation approach to text simplification has largely been limited to this one language.

Handcrafted systems typically treat syntactic and lexical simplification separately (Siddharthan 2014). Systems for lexical simplification tend to make use of a number of external resources to identify difficult words or larger lexical units in a text and replace them with simpler alternatives (i.e. *lexical substitution*). The identification of difficult lexical units (see Section 2.2) is an important first step. Polysemy or homonymy constitutes a considerable obstacle for automated lexical simplification. After having identified difficult lexical items, they have to be replaced by simpler ones. If a word has multiple senses, we want to replace it with another word that retains the original meaning as much as possible. One way of addressing this issue is by implementing a form of word sense disambiguation before attempting to apply lexical simplification (Bott et al. 2012). Since the resources described above to help identify difficult words are purely formal in nature, it is difficult to take word sense information into account already in the identification step. In the replacement step, however, this is possible, especially if structured databases such as WordNet (Fellbaum 1998, Miller 1995) are used to find potentially easier synonyms, near-synonyms or hypernyms of an identified difficult word (Devlin and Tait 1998, Carroll et al. 1998). Word embeddings or context word vectors obtained from larger, unstructured corpora can also be used to directly identify potential replacements, without the need for a separate, possibly error-inducing, word sense disambiguation step (Baeza-Yates et al. 2015, Paetzold and Specia 2016). Other researchers have attempted to combine information coming from WordNet with extracted context vectors to select the best synonyms for replacement (Saggion et al. 2015).

Depending on the approach taken, a final step in the lexical simplification process may be necessary, namely morphology generation and adaptation and possibly syntactic modification (Bott et al. 2012, Chen et al. 2017, Wang et al. 2016). Lexical databases such as WordNet typically contain lemmatized word forms, without inflections. After having selected a synonym from such a database, inflections have to be added to the base form of the word to ensure the grammatical correctness of the output sentence. Moreover, synonyms selected for replacement can differ from the original word in terms of grammatical categories such as gender and number, which could necessitate further changes to the output sentence (e.g. to determiners, other dependents or arguments, depending on the language). This issue is less problematic in approaches that apply lexical simplification without making use of resources such as word lists and lexical/semantic databases, but instead rely on machine learning techniques and large corpora (Chen et al. 2017, Wang et al. 2016).

Sometimes, instead of replacing difficult words, texts are augmented with automatically retrieved explanations or definitions (Elhadad 2006, Kaji et al. 2002, Kandula et al. 2010), which can help to improve text comprehension. When second language learners or non-native speakers are the target population, these definitions or links to dictionary entries can be provided in the native language of the readers. Related work on other types of *text augmentation* is relevant in the context of lexical simplification as well, such as tools to automatically enhance texts by adding links to relevant web content, most typically Wikipedia pages (Brank et al. 2017, Noraset et al. 2014). Even though these types of text augmentation would not fall within a strict interpretation of 'lexical simplification', it is clear that they have the potential to make texts easier to read and understand, which is the ultimate aim of text simplification.

Since no (large enough) parallel corpus of simplified Dutch exists, we tackle the lexical simplification problem with a handcrafted pipeline approach (see Section 3). First, two resources are used to estimate the difficulty of lexical items: a frequency list compiled on the basis of a 'simplified' corpus and a list containing the mean age of acquisition of Dutch word lemmas. A combination of these two types of resources should work (reasonably) well in approximating human judgments of lexical difficulty (Jauhar and Specia 2012, Wilkens et al. 2014), especially since the corpus used to compile the frequency list consists of texts written with the explicit intention of producing 'clear and simple' language (Wrobel 2016). Second, potential replacement words are retrieved from the Dutch version of WordNet, meaning we opt for a formal approach using a manually compiled resource rather than a corpus-driven approach based on, for example, word embeddings (Paetzold and Specia 2016). To deal with the issue of polysemy, our system includes a word sense disambiguation component, and a large-scale language model is used to ensure grammatical correctness and context-dependent appropriateness (Bott et al. 2012). Finally, the system also includes a text annotation and augmentation component, which is mainly intended to increase the system's coverage, as automated lexical substitution based on synonym retrieval from resources such as WordNet tends to be hampered by the limited number of available replacements (De Belder and Moens 2010).

2.4 Evaluating text simplification systems

The evaluation of generated output constitutes an important step in each experimental study on text simplification systems. However, datasets consisting of original and simplified sentences are few and far between. Sometimes sentences from Simple Wikipedia are used for this purpose (Horn et al. 2014), or the output is compared to text that has been manually simplified, for example using evaluation metrics that are also used in the context of machine translation, such as BLEU or NIST (Coster and Kauchak 2011). The generated output can also be compared to the output of other automated simplification systems (Siddharthan 2014). Ideally, systems are assessed by means of human evaluation. This can mean both evaluation by fluent readers and, ultimately, by the envisaged target population of the application. Since human evaluators of text simplification systems can focus on different aspects of the generated text (Štajner 2018), such ratings can target different aspects of the generated output, such as fluency, grammaticality, simplicity and meaning preservation (Siddharthan 2006, Woodsend and Lapata 2011, Wubben et al. 2012). Sometimes other evaluation metrics, such as fixation times obtained by means of eye tracking experiments, are used as well (Bott et al. 2012).

Disposing of a *gold standard* and being able to automatically compare generated output to this standard is not only useful for testing purposes, but also for the tuning of parameters used by the system (weights for the different features in the model, thresholds for frequency or age of acquisition, etc.). We return to this issue in the next section.

3. System setup

We use a pipeline approach to tackle the problem of lexical simplification. Figure 1 provides a schematic overview of the different steps in the simplification process. Each of these steps is described in more detail below. Sentences are first pre-processed and word sense disambiguation is performed (Section 3.1), then difficult words are identified (Section 3.2), potential replacements are retrieved, ranked and selected (Section 3.3), and noun and verb inflections are added and determiners changed if necessary (Section 3.4). Trigram probabilities are calculated and compared for sentence fragments with original and replacement words to determine which potential replacements to keep and which not (Section 3.5). In a final step, the text is annotated with definitions and synonyms and links to Wikipedia pages are added (Section 3.6).

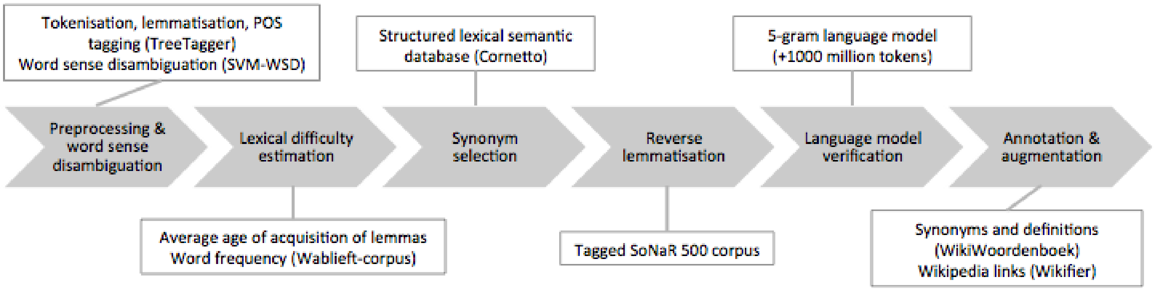

Tokenisation, lemmatisation, POS tagging (TreeTagger)
Word sense disambiguation (SVM-WSD)
Structured lexical semantic database (Cornetto)
5-gram language model (+1000 million tokens)
Preprocessing & word sense disambiguation
Lexical difficulty estimation
Synonym selection
Reverse lemmatisation
Language model verification
Annotation & augmentation
Average age of acquisition of lemmas
Word frequency (Wablieft-corpus)
Tagged SoNaR 500 corpus
Synonyms and definitions (WikiWoordenboek)
Wikipedia links (Wikifier)

Figure 1: Overview system architecture

3.1 Pre-processing and Word Sense Disambiguation

Pre-processing involves tokenization of the input sentences, part-of speech tagging, and lemmatization. All these steps are done using TreeTagger (Schmid 1994, Schmid 1995), a probabilistic tagger that uses a binary decision tree to estimate transition probabilities.[6] POS-tags are used to store and retrieve grammatical information. Lemmatization is needed to retrieve information from the lexical databases used for the identification of difficult words.

To tackle the problem of polysemy, a Dutch word sense disambiguation tool[7] is used that is based on support vector machines and trained on the data of the DutchSemCor project (Vossen et al. 2012). This tool uses a bag-of-words model for feature representation. The identified word senses are linked to the lexical items in the Cornetto database (Vossen et al. 2013), which is used further down the pipeline to identify potential synonyms for simplification. The tool estimates the probability of each sense of a word based on the other words in the sentence. We only use the word sense that is estimated to be the most likely one.

[6] We use TreeTagger since the word sense disambiguation tool relies on it.
[7] https://github.com/cltl/svm_wsd

3.2 Lexical difficulty estimation

Two resources are used to estimate the difficulty of each token in the input sentence: (a) aggregated data coming from psycholinguistic studies into the average age of acquisition (AoA) of Dutch words[8] (Brysbaert et al. 2014), and (b) frequency information of Dutch tokens calculated on the basis of the *Wablieft* corpus, consisting of the archive of articles written for the weekly newspaper Wablieft[9] up until December 2017. The first resource contains information on approximately 30,000 word lemmas, whereas the frequency list was compiled on the basis of more than 2 million tokens. Importantly, the Wablieft newspaper is written with the explicit aim of using simple and clear language, and its target audience consists of people who have difficulty reading and/or who are functionally illiterate. Two thresholds are set for identifying words to be potentially replaced: (a) maximum average AoA of word lemma, and (b) minimum frequency of lemma. These thresholds are determined on the basis of a hill-climb algorithm (Section 4.2). If word lemmas are found to be acquired already at a young age, and they occur frequently in the reference database, they are deemed easy enough and potential replacements are not considered. Lemmas that do not occur in the AoA or frequency list are also considered to be potentially difficult. No specific treatment of compound words is foreseen, nor does the system target multiword units (see Section 6).

3.3 Synonym and hypernym identification and selection

The structured lexical semantic database Cornetto (Vossen et al. 2013) is used to identify synonyms of words that have been identified as difficult. Cornetto groups lexical items together in synonym sets and indicates the relationships between the different *synsets* (e.g. "`is hypernym of`" or "`is antonym of`"). For the purpose of lexical simplification, words belonging to the same synset as the difficult word are, theoretically speaking, the most interesting. In addition, we also consider words in synsets that have been identified as near synonyms, as well as hypernyms. A clear hierarchy is respected when looking for simplified replacement words: first synonyms are considered, then hypernyms, then near synonyms.[10]

For those lexical items identified as difficult, the average AoA and frequency of retrieved synonyms is verified. If a synonym is found for which the AoA is (at least a certain percentage) lower than that of the original word, and the corresponding frequency is (a certain percentage) higher, it is labeled as potential replacement word.[11] The lemma with the lowest AoA is selected as best alternative. If no synonyms are found that satisfy the replacement conditions, hypernyms are considered, but only if the original word respects potentially stricter thresholds than those set for synonyms (i.e. a higher AoA and lower frequency for the original lemma). For hypernyms, the same procedure is followed as for synonyms, also with potentially more stringent conditions (i.e. in terms of percentage change in AoA and frequency). Finally, near synonyms are considered when no synonyms or hypernyms are found as potential replacements.

3.4 Reverse lemmatization and grammatical adaptation

Dutch is a morphologically not very rich language, but a certain number of grammatical categories are expressed by means of inflectional morphemes (e.g. plural for nouns, certain verb forms in present and past tense, adjective gender and definiteness). In order to preserve these inflections, the lemmas that have been selected for replacement have to be de-lemmatized. Most information on the inflectional form of the original word is stored in the POS tag by Treetagger during lemmatiza-

[8] http://crr.ugent.be/archives/1602

[9] http://www.wablieft.be/krant. Corpus: https://ivdnt.org/downloads/taalmaterialen/tstc-wablieft-corpus

[10] Hypernyms are considered before near synonyms since, for the sake of lexical simplification, we consider it preferable to lose specificity rather than to potentially shift meaning.

[11] These thresholds are tunable model parameters.

tion[12]. To perform reverse lemmatization, a parsed version of the 500-million-word SoNaR corpus is used (Oostdijk et al. 2013). For each lemma, different inflectional forms and their respective tags are stored. The correct inflectional form of the replacement word is selected by matching the Treetagger-tag of the lemma with the SoNaR-tag, and retrieving the corresponding form.

Apart from inflections, also other grammatical phenomena could interfere with the correctness of the output of the simplification system. For example, definite articles in Dutch are coded for gender (neuter vs. masculine/feminine), and so are possessive and demonstrative determiners. The choice of relative pronoun also depends on the gender of the noun it refers to. In the current version of the simplification tool, only definite articles are explicitly addressed. When a singular noun is replaced by an alternative word, the correct article for the alternative word is determined by looking up the bigram probabilities of the word preceded by the two possible articles (*de* and *het*) in a reference corpus (see following section). If the sentence in which the word is changed contains an article one or two tokens before the word in question, this article is matched with the retrieved one, and changed if necessary.

3.5 Trigram verification with language model

After having selected potential replacement words and having retrieved the correct inflectional forms, a language model is used to verify whether the replacements, ranked according to their AoA, are appropriate in the context of the original sentence. This language model was compiled on the basis of a large-scale corpus (of over 1000 million tokens) combining different sources, such as Subtitles2016 (Lison and Tiedemann 2016), EUBookshop (Skadiņš et al. 2014), DGT, Europarl and Wikipedia (Tiedemann 2012), CGN Flemish (Oostdijk et al. 2002) and SONAR500 (Oostdijk et al. 2013). Trigram probabilities are calculated both for the original word and for the replacement word, and these are subsequently compared. Where possible, three trigram probabilities are calculated per word (i.e. $n-2$, $n-1$, n; $n-1$, n, $n+1$; n, $n+1$, $n+2$), but at sentence boundaries this number is restricted to two or even one. Replacement words can be rejected if (a) the cumulative probability of their trigrams is lower than the cumulative probability of the trigrams of the corresponding original word (potentially multiplied with a certain percentage), or (b) the probability of their trigrams meets some other criterion. The exact rejection criteria are determined during parameter tuning (Section 4.2).

3.6 Text annotation and augmentation

In a final step, input sentences are annotated in two ways: (a) for words that were identified as difficult but that were not simplified by the system, 'easier' synonyms from Cornetto as well as definitions retrieved from WikiWoordenboek[13], the Dutch Wiktionary, are added to the system output as html-tags ('hover text'), and (b) for multi-word units, proper nouns and difficult words, hyperlinks to pages on Wikipedia are added using Wikifier [14] (Brank et al. 2017). Both are intended as optional functionalities to potentially increase the coverage of the system.

The definitions are retrieved from a cleaned *database backup dump* of nlwiktionary[15]. Words without definition are filtered out, and only the first definition is retained. Also lexical 'meta-information' (such as POS identification or linguistic origin) is filtered out. For Wikipedia links, we give precedence to multi-word units, and consider only the highest-ranked option for each word (sequence), according to their *pagerank score*. The definitions and links are allowed to overlap. Certain additional restrictions were added on the basis of an analysis of development set output (see section 4.2).

[12] Adjective inflection is not coded, and certain distinctions for verbs are also missing (e.g. 1st vs 2nd and 3rd person singular for present tense).

[13] http://nl.wiktoinary.org

[14] http://www.wikifier.org

[15] https://dumps.wikimedia.org, 1 January 2018

4. Experimental design

4.1 Dataset

A small dataset[16] is used to develop and evaluate the system: 120 sentences taken from the Flemish newspaper De Standaard[17], most of which were also used in the context of a syntactic simplification project (Sevens et al. 2018), are divided in a development/tuning set (70 sentences) and a test set (50 sentences). The tuning set contains 1362 words, the test set 958. Some sample sentences from the development set are provided below (Examples 1-3). Manually identified difficult words (see Section 4.2) are tagged with an asterisk.

(1) We hebben altijd duidelijk gecommuniceerd dat we versnippering (*) zouden tegengaan.
We have always clearly communicated that we fragmentation (*) would combat.
'We always communicated clearly that we would combat fragmentation.'

(2) Het bedrijf heeft de intentie (*) om het aantal ploegen te verminderen en beschikbare
The company has the intention (*) to the number teams to diminish and available
middelen efficiënter (*) te gebruiken.
resources more efficiently (*) to use.
'The company intends to reduce the number of teams and to use the available resources in a more efficient way.'

(3) Dod zag de omzet (*) de voorbije jaren fors dalen en in mei werd de gerechtelijke (*)
Dod saw the revenue (*) the previous years sharply drop and in May was the judicial (*)
reorganisatie (*) opgestart.
reorganisation (*) started up.
'Dod saw sales fall sharply in recent years and in May the judicial reorganisation was started.'

4.2 Parameter tuning and setting

The system described above contains a number of parameters that can be set to specific values (e.g. thresholds for AoA and frequency, settings for trigram probabilities). It is also possible to bypass or exclude certain features (e.g. not include hypernyms or near synonyms, only consider AoA, bypass trigram verification). To automate a part of the tuning process, we manually identified words in the development set that should be targeted by the simplification system. The three authors independently tagged each word as being difficult or not, based on their own subjective evaluation, and without a specific target audience in mind. If at least two authors identified a word as difficult, it was coded as such.[18] Out of 700 content words in the development set, 81 (i.e. 11.57%) were identified as difficult. We used a local hill-climb algorithm to tune the AoA and frequency thresholds (as well as the way these parameters are combined) to approximate our manual classification as closely as possible, using the F2-measure.[19] The other system parameters were set on the basis of a qualitative evaluation of the system's output on the development set.

Table 1 provides an overview of the quantitative parameter settings after tuning. We found that a maximum average age of acquisition of 9.34 and a minimum frequency of 33 in the Wablieft-corpus, combined using an AND-condition, best approximated our classification of words into difficult and non-difficult ones, with an F2 value of 0.758 on the development set.[20] Depending on the specific target audience, these values can be increased or decreased, resulting in fewer or more words being targeted by the system.

[16] The full dataset is available upon request.

[17] http://www.destandaard.be

[18] Even following this loosely designed procedure for identifying difficult words, the agreement between coders was relatively high (pairwise Cohen's kappas ranged between 0.78 and 0.89).

[19] More weight was given to recall than precision given the relatively low proportion of difficult words, and the fact that we considered true positives more important than true negatives.

[20] Recall: 0.951; precision: 0.418; accuracy: 0.841. Note that proper nouns and adjectives were not filtered out, which resulted in an increased number of false positives (39 in total) and consequently lower evaluation scores.

Max AoA	**Min freq**	**% diff. AoA**	**Min pagerank**
9.34	33	95%	0.05

Table 1: Quantitative parameter settings after tuning

When it comes to picking simplified alternatives for words identified as difficult, our qualitative analyses of development set output indicate that it is better to discard near-synonyms and hypernyms, since these result in too many lexical errors. For a synonym to be considered as replacement, its AoA has to be lower than the AoA of the original word multiplied by 0.95. No restrictions with regard to frequency are placed on the potential replacement word. With regard to trigram verification, the only requirement that is retained is that at least one of the trigrams containing the potential replacement word should occur at least once in the reference corpus. This seemingly not very restrictive condition appears to filter out most of the grammatically and also lexically inappropriate replacements, taking into account the local context of the word in the sentence.

Finally, the following restrictions are added to improve the selection of Wikipedia links: (a) pagerank scores lower than 0.05 are filtered out, (b) multi-word units cannot start with an article, particle or preposition, and cannot end in a preposition, (c) no single-word links are added to verbs, and (d) words can only be attributed one link.

4.3 Evaluation

We use a stepwise evaluation procedure. First, we verify whether the words changed by the system correspond to the ones we tagged as difficult. Then, the first author identified substitutions that were either grammatically or lexically/semantically incorrect. Next, 7 speakers of Dutch were asked to evaluate the simplifications (excluding the incorrect ones), by indicating whether they thought the original or changed sentence was the simplest (a 'no difference' option was also provided). The evaluation was blind, i.e. the participants did not know which of the two sentences was the original one. Finally, the added definitions and links were manually checked by the first author for adequacy.

5. Results

Out of the 474 content words in the 50 sentences constituting the test set, 78 were manually identified as being potentially difficult (i.e. 16.46%). The confusion matrix associated with the difficult word identification is provided in Table 2. The system identified 161 difficult words (amongst which 42 proper nouns/adjectives) using the parameter settings found during tuning. Of the difficult words, 96.2% (i.e. 75) were also tagged as such by the system, whereas 11.1% of the non-difficult words (i.e. 44) were earmarked as difficult. This resulted in an F2 score of 0.793.[21] Thirty-five words (or 7.4% of content words) were changed by the lexical simplification system.[22] Of these 35 words, 24 were labeled difficult (68.6%). This means that out of the 78 words that were originally labeled as difficult, 30.8% were changed. Looking at the quality of the replacements, none resulted in grammatical errors, and 9 (i.e. 25.7%) were judged to be lexically or semantically incorrect. The blind evaluation of the 26 remaining replacements (each occurring in a different test sentence) by 7 raters showed that a majority of the raters esteemed 18 of the sentences containing a modification (i.e. 69.2%) to be simpler than their original counterparts. Of these 18 successful simplifications, 16 affected words were manually identified as difficult, meaning that 20.5% of the words originally identified as difficult were successfully adapted by the simplification system. The quantitative results of the analyses are summarised in Table 3.

[21] Recall: 0.962; precision: 0.466; accuracy: 0.812.

[22] Also two definite articles were changed accordingly.

	Predicted difficult	**Not predicted difficult**	*Total*
Manually tagged difficult	TP: 75	FN: 3	*78*
Manually tagged not difficult	FP: 44 + 42 proper N/adj	TN: 310	*396*
Total	*161*	*313*	*474*

Table 2: Confusion matrix difficult word identification (TP=True positives, FP=False positives, FN=False negatives, TN=True negatives)

			Content words			**Named entities**	**Total**
			Difficult	Non-difficult	Total		
Total in text			78	396	474	61	-
Identified as difficult by system			75	44	119	42	161
Adaptation	No error	Judged simpler	16	2	18	-	18
		Not simpler	4	4	8	-	8
	Error	Grammatical	0	0	0	-	0
		Lexical/semantic	4	5	9	-	9
	Total		24	11	35	-	35
Definition	No error	Rated helpful	22	8	30	12	42
		Not helpful	17	13	30	3	33
	Error		3	3	6	7	13
	Total		42	24	66	22	88
Synonym	No error	Rated helpful	4	1	5	-	5
		Not helpful	0	0	0	-	0
	Error		1	2	3	-	3
	Total		5	3	8	-	8
Link	Correct		26	6	32	37	69
	Incorrect		5	10	15	3	18
	Total		31	16	47	40	87

Table 3: Overview results

Examples 4-7 show four successful simplifications, i.e. changes that were deemed necessary and that were rated as resulting in simpler sentences. A more extensive list of successful simplifications is provided in Appendix A.

(4) *De test* (< het experiment) gaat vijf jaar en twee maanden duren en begint op zijn vroegst
The test (< the experiment) will five years and two months last and starts at its earliest
eind 2019.
end 2019.

'*The test* (< the experiment) will take five years and two months and starts at the earliest at the end of 2019.'

(5) In zijn *preek* (< homilie) vroeg de paus een oplossing voor de situatie in Syrië.
In his *sermon* (< homily) asked the pope a solution for the situation in Syria.

'In his *sermon* (< homily), the Pope asked for a solution to the situation in Syria.'

(6) De anderen hopen dat de curator het faillissement zo snel mogelijk zal *afwerken*
The others hope that the bankruptcy trustee the bankruptcy as soon possible will *deal with*
(< afhandelen), zodat ze hun ontslagvergoeding kunnen krijgen.
(< conclude), so they their severance pay can receive.

'The others hope that the bankruptcy trustee will *deal with* (< conclude) the bankruptcy as quickly as possible, so that they can receive their severance pay.'

(7) Google heeft in dat land zijn Europese hoofdkwartier gevestigd en een *groot*
Google has in that country its European headquarters established and a *large*
(< aanzienlijk) deel van de winst die het haalt in Europa wordt ook in Ierland geboekt.
(< significant) part of the profit that it gets in Europe is also in Ireland booked.
'Google has established its European headquarters in that country and a *large* (< significant) part of the profit it makes in Europe is also booked in Ireland.'

The simplifications affect nouns (with and without article change), verbs (with different inflections) and adjectives. The example sentences also provide ample evidence of difficult words that were not changed by the system, such as *curator* (*bankruptcy trustee*) and *ontslagvergoeding* (*severance pay*). A qualitative analysis of the unchanged difficult words shows that for most of them no (simpler and appropriate) synonyms are available in Cornetto. Especially compound words pose a problem for the simplification system.[23] We also identified a number of cases where the WSD tool did not link a word to the correct entry in Cornetto (e.g. the verb *analyseren* - *analyse* was tagged as a noun).

Sentences 8-10 provide examples of changes made by the system that were, respectively, not necessary (since they affected words that we did not identify as difficult), did not lead to simpler sentences (as judged by the 7 raters) or resulted in lexical/semantic errors. A more extensive list of examples is given in Appendix B.

(8) SilentKeys beschermt je niet alleen tijdens je surftochten thuis, waar je wellicht al
SilentKeys protects you not only during your surfing trips home, where you perhaps already
een *sterk* (< krachtig) antivirusprogramma draait.
a *strong* (< powerful) antivirus program run.
'SilentKeys not only protects you during your surfing trips at home, where you might already run a *strong* (< powerful) antivirus program.'

(9) Commissaris Vandersmissen werd aan het einde van de *landelijke* (< nationale) betoging
Commissioner Vandersmissen was at the end of the *nationwide* (< national) demonstration
aangevallen door een relschopper in een rood T-shirt.
attacked by a rioter in a red T-shirt.
'Commissioner Vandersmissen was attacked by a rioter wearing a red T-shirt at the end of the *nationwide* (< national) demonstration.'

(10) Daarop beslisten bijna dertig landen om meer dan 140 Russische *politici* (< diplomaten)
Then decided almost thirty countries to more than 140 Russian *politicians* (< diplomats)
uit te zetten.
to expel.
'Then almost 30 countries decided to expel more than 140 Russian *politicians* (< diplomats).'

It should be noted here that different people can (and do) have opposing views on what constitutes simplification, on which changes sufficiently retain the original meaning of the sentence and on which lexical choices are appropriate in which context. To illustrate this, for only 3 out of 26 sentence pairs presented to the 7 raters, all raters agreed, and for 6 sentence pairs, 6 out of 7 raters picked the same option. In 9 cases, the answer with the highest frequency was picked only by 3 or 4 raters.[24]

Next, we look at the annotations that are added to the text. In total, 88 definitions are provided, 42 of which target difficult words. Twenty-two are added to named entities, and 24 to non-difficult words. We only did an informal analysis of the quality of the definitions, by loosely classifying them into three categories: correct and helpful (42), correct but not helpful for understanding (33), and incorrect (13). Out of the 42 definitions provided for difficult words, 22 were rated helpful, 17 not helpful and 3 incorrect. Examples 11-13 illustrate each of these categories. For more examples, we refer to Appendix C.

[23] Examples include *arbeidsvoorwaarden* (*terms of employment*), *antivirusprogramma* (*antivirus program*) and *sportkledingfabrikant* (*sportswear manufacturer*).

[24] Pairwise unweighted kappa coefficients between raters did not exceed 0.38.

(11) fauna → Het geheel aan dieren in een gebied.
fauna → The whole of animals in an area.
'All of the animals in an area.'

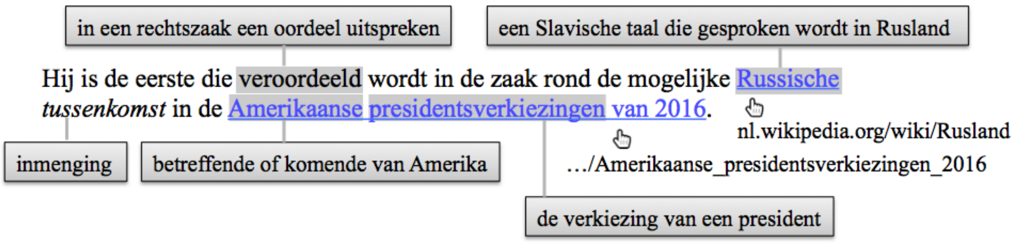
(12) wetsvoorstel → Door de regering vervaardigd ontwerp van een wet die aan de
draft law → By the government produced draft of a law that to the
volksvertegenwoordiging wordt voorgelegd.
people's representation is presented.
'Government draft of a law that is submitted to the parliament.'

(13) criminele *circuit* → Omloop (voor snelheidswedstrijden).
criminal *circuit* → Track (for speed races).
'Racetrack.'

The correct and helpful definitions explain difficult words using arguably simple vocabulary, which can help improve understanding (see example 11). In contrast, the correct definitions that do not help understanding either contain difficult vocabulary themselves (see example 12), or simply repeat (parts of) the defined word. The incorrect definitions can be attributed to the wrong word sense being identified (e.g. *criminele circuit* - *criminal circuit* interpreted as a racetrack), individual words belonging to expressions being literally interpreted (e.g. *in de kiem smoren* - *nip in the bud*), or words belonging to a multi-word unit being taken out of context (e.g. *Verenigde Staten* - *United States*). In addition to the definitions, on 8 occasions one or several synonyms from Cornetto are provided by the system. 5 of these appeared to be helpful, whereas 3 were incorrect.

Finally, we evaluate the Wikipedia links that were added to the test sentences (87 in total). Eighteen of these links spanned multiple words. Forty targeted named entities (of which 13 were multi-word units). We found 3 errors amongst these 40 links. In contrast, 21 named entities in the test sentences were not attributed a link to a Wikipedia page (for half of these no Dutch Wikipedia page existed, some others were not classified as difficult). Of the remaining 47 links, 32 referred to the correct Wikipedia page. Wikipedia links were added to 31 difficult words. Twenty-six of these links referred to the correct page. In four cases a Wikipedia link was added to a difficult word that was neither changed by the simplification system nor annotated with a definition. Many of the links were added to difficult words that received a correct definition that did not help in understanding the word.

Figure 2 shows an example of the system's output in html format. One word was changed (*inmenging* → *tussenkomst* - *interference* → *intervention*). The replacement word is shown in italics, and the replaced word appears as hover text. Definitions for four words are also available when hovering over them. This is indicated by gray shading. Finally, two Wikipedia links are added, one of which to a 4-word sequence. An example of the html code generated by the simplification system is provided in Appendix D.



in een rechtszaak een oordeel uitspreken

een Slavische taal die gesproken wordt in Rusland

Hij is de eerste die veroordeeld wordt in de zaak rond de mogelijke Russische *tussenkomst* in de Amerikaanse presidentsverkiezingen van 2016.

nl.wikipedia.org/wiki/Rusland

inmenging

betreffende of komende van Amerika

.../Amerikaanse_presidentsverkiezingen_2016

de verkiezing van een president

Figure 2: Example of annotated html output

6. Discussion

The lexical simplification system tested in this paper makes use of a pipeline in the context of which different resources are consulted and combined. This potentially creates problems, since errors that occur somewhere along the line are propagated downwards, and recovering from an error is nearly impossible. In the absence of a parallel corpus of original and simplified sentences in Dutch, however, it is difficult to imagine a lexical simplification tool that does not function in such a way. Otherwise it would be worthwhile to test how a holistic system, either based on phrase-based or neural machine translation, performs (Nisioi et al. 2017, Wang et al. 2016, Wubben et al. 2012). With a pipeline approach, the quality of the output of the system depends on the quality of each of the resources used. The psycholinguistic resource used to measure word difficulty (Brysbaert et al. 2014) appeared to be working relatively well in combination with the frequency list drawn from the corpus of simplified Dutch. A combination of these two resources managed to approximate our subjective binary classification of words (difficult or not) rather well (96.2% of difficult words were correctly identified as such; only 11.1% of the remaining content words were incorrectly classified as difficult). We believe that using a dedicated 'simple' corpus to compile the frequency list was beneficial for the estimation of lexical difficulty, but this should be tested in more detail (Wrobel 2016). Potential improvements here could focus on developing a more sophisticated classifier, for example taking into consideration features containing information about the local context of lexical items (Davoodi et al. 2017) or linguistic criteria such as polysemy and word length (Davoodi et al. 2017, Gala et al. 2013, Walker et al. 2011). Also the structured lexical database Cornetto (Vossen et al. 2013) often offered useful and appropriate synonyms. Some errors in tagging were detected, and there were also a number of problems with word sense disambiguation, but generally speaking the accuracy of the used tools was satisfactory. We could have opted to work with an n-best list throughout the pipeline (e.g. also allow for two or more word senses), but this would have complicated the selection procedure.

It can be argued that the adaptation component of the system tested here has low coverage, which is in line with previous studies using comparable resources (De Belder and Moens 2010). Only 20.5% of the original words that were manually labeled as being potentially difficult were adequately changed and simplified. This is not surprising given the fact that the modifications are restricted to single words being replaced by easier single-word synonyms, which are not always available in the language or in the used resource. The system tested here does not have the flexibility to offer alternative formulations or paraphrases. The coverage of the system could be improved by incorporating a corpus-driven component for the selection of potential synonyms, for example by using word embeddings or context word vectors (Baeza-Yates et al. 2015, Paetzold and Specia 2016, Saggion et al. 2015). Looking at precision and accuracy, 68.6% of the words changed by the system were labeled difficult, and 45.7% of the total changes were at the same time necessary, appropriate and correct. The trigram language model as final check acted as a rather strong filter. It made sure that the selected replacements were grammatically correct (no grammatical errors were introduced by the modifications) and adequate in the local context (only 25.7% of the changes were lexically/semantically incorrect). At the same time, however, potentially good candidates for simplification were sometimes rejected since the corresponding trigrams did not occur in the reference corpus.

The system's annotation component has higher coverage, but it is arguably less useful as a tool for simplification. Not only does the amount of text to read increase when adding definitions, our preliminary analyses showed that about half of the WikiWoordenboek definitions offered either use too difficult vocabulary themselves or are not particularly helpful in clarifying the meaning of difficult words. It is worth testing whether filtering definitions on the basis of their lexical difficulty (e.g. using the AoA and frequency lists) can reduce the number of unhelpful annotations. Another option is to try to simplify the definitions using the adaptation component of the system. The links to Wikipedia pages that were added using Wikifier were largely appropriate and correct, in part thanks to the imposed filters. However, in this paper we did not investigate how helpful the offered definitions and links are for improving the understanding of texts. In addition, the system's output could be enriched with pictographs or items from a picture database, depending on the target audience (Sevens et al. 2017).

A number of problematic issues were identified during the development and testing phase. For example, the system had difficulties with compound words, words forming part of larger word groups, expressions, and separable verbs. These problems require different solutions that could be tackled in future versions of the simplification tool. One possibility is to integrate a form of compound word splitting into the system (Vandeghinste 2002). Also adjectives were not dealt with in an entirely satisfactory way (simply adding an '-e' to the base form when necessary is, admittedly, a naive and too simplistic solution). Moreover, grammatical problems could arise when sentences contain relative pronouns or possessive determiners that are dependent on nouns that are changed. This issue should be tackled in future incarnations of the system.

It should be noted that the simplification system was automatically tuned on the basis of the authors' intuitions about which words are difficult and which ones are not. Difficulty is by definition a subjective notion, and it is highly likely that different target populations would benefit from different parameter settings. The system setup allows to easily adapt the potential coverage of the simplifications, by increasing or decreasing the required AoA and frequency. Our tuning showed that these two resources work well as proxies of perceived difficulty.

Finally, the design of the study itself was limited in a number of ways. First, only a small set of sentences taken from a single newspaper was used to test the system. It should be tested on a larger, more diverse dataset. Second, the final evaluation was to a considerable extent subjective in nature. Ideally, more raters should be involved in the process, and the evaluation should comprise different criteria (e.g. fluency, difficulty, grammaticality). Moreover, the final version of the system should be tested with members of one or more specific target populations. The system's annotation component was only evaluated in a preliminary way. Third, in relation to this, the tool was not developed with a specific target population in mind. Specialisation could help to improve performance. For example, if the tool is to be used by people with reading problems or with dyslexia, word length could also be taken into consideration as a factor. Moreover, in the case of people with dyslexia, attention should also be paid to issues such as proper font selection (Rello and Baeza-Yates 2013). Fourth, only the tuning of system parameters that govern the identification of difficult words was automated (on the basis of a manually annotated corpus). The tuning of replacement selection was done on the basis of a manual evaluation of development set output. Also this step could have been automated by developing a reference corpus. All of these shortcomings can be addressed in future studies.

7. Conclusion

In this paper we discussed the design, development and testing of an automated lexical simplification tool for Dutch. Such text simplification tools can be useful for a wide range of target populations, such as people with intellectual disabilities, second language learners, aphasics and children. The tool presented here makes use of a number of resources (e.g. a structured lexical database, a list with average AoA of words, and a reference corpus for determining frequencies, compiling a language model and performing reverse lemmatisation) and tools (e.g. tagger, word sense disambiguator, Wikifier) that are combined in a pipeline. Even though the results of the study indicate that only around 1 out of 5 potentially difficult words in the test set were adequately changed and simplified by the tool, in almost 70% of the cases the changes targeted words that had been labeled as being potentially difficult, and around 46% of all changes were both lexically and grammatically correct, next to leading to simplification. No grammatical errors were introduced by the modifications. The system's annotation component has a wider coverage than its adaptation component, but the usefulness of these annotations for lexical simplification should be tested more thoroughly. The first version of the lexical simplification system presented here should be further developed so that it can be used in practical applications, potentially in combination with a syntactic simplification component (Sevens et al. 2017).

References

Aluísio, Sandra M., Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes (2008), Towards Brazilian Portuguese automatic text simplification systems, *Proceedings of the Eighth ACM Symposium on Document Engineering*, pp. 240–248.

Arya, Diana J., Elfrieda H. Hiebert, and P. David Pearson (2011), The effects of syntactic and lexical complexity on the comprehension of elementary science texts, *International Electronic Journal of Elementary Education* **4** (1), pp. 107–125.

Baeza-Yates, Ricardo, Luz Rello, and Julia Dembowski (2015), Cassa: A context-aware synonym simplification algorithm, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp. 1380–1385.

Barbu, Eduard, M Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L Alfonso Ureña-López (2015), Language technologies applied to document simplification for helping autistic people, *Expert Systems with Applications* **42** (12), pp. 5076–5086.

Barlacchi, Gianni and Sara Tonelli (2013), Ernesta: A sentence simplification tool for children's stories in Italian, *in* Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing*, pp. 476–487.

Biran, Or, Samuel Brody, and Noémie Elhadad (2011), Putting it simply: a context-aware approach to lexical simplification, *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, Vol. 2, pp. 496–501.

Bott, Stefan, Horacio Saggion, and Simon Mille (2012), Text simplification tools for Spanish, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 1665–1671.

Brank, Janez, Gregor Leban, and Marko Grobelnik (2017), Annotating documents with relevant Wikipedia concepts, *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses.*

Brouwers, Laetitia, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François (2014), Syntactic sentence simplification for French, *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pp. 47–56.

Brysbaert, Marc, Michaël Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms (2014), Norms of age of acquisition and concreteness for 30,000 Dutch words, *Acta Psychologica* (150), pp. 80–84.

Byrnes, Heidi and Castle Sinicrope (2008), Advancedness and the development of relativization in l2 german: A curriculum-based longitudinal study, *in* Ortega, Lourdes and Heidi Byrnes, editors, *The longitudinal study of advanced L2 capacities*, Routledge, pp. 109–138.

Candido Jr., Arnaldo, Erick G. Maziero, Caroline Gasperin, Thiago A.S. Pardo, Lucia Specia, and Sandra M. Aluisio (2009), Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese, *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 34–42.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait (1999), Simplifying text for language-impaired readers, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 269–270.

Carroll, John, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait (1998), Practical simplification of English newspaper text to assist aphasic readers, *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology.*

Castells, Manuel (1996), *The Information Age: Economy, Society, and Culture. Volume I: The Rise of the Network Society*, Wiley Blackwell.

Chen, Ping, John Rochford, David N. Kennedy, Soussan Djamasbi, and Peter Fay (2017), Automatic text simplification for people with intellectual disabilities, *Proceedings of the International Conference on Artificial Intelligence Science and Technology*, pp. 725–731.

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim, and Jong C. Park (2013), Enhancing readability of web documents by text augmentation for deaf people, *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, p. 30.

Coster, William and David Kauchak (2011), Learning to simplify sentences using Wikipedia, *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pp. 1–9.

Cutler, Anne (1983), Lexical complexity and sentence processing, *in* Flores d'Arcais, G.B. and R.J. Jarvella, editors, *The Process of Language Understanding*, John Wiley and Sons, pp. 43–79.

Daelemans, Walter, Anja Hothker, and Erik Tjong Kim Sang (2004), Automatic sentence simplification for subtitling in Dutch and English, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1045–1048.

Daelemans, Walter, Orphee De Clercq, and Veronique Hoste (2017), Stylene: an environment for stylometry and readability research for Dutch, *in* Odijk, Jan and Arjan Van Hessen, editors, *CLARIN in the Low Countries*, Ubiquity Press, pp. 195–210.

Dahl, Östen (2004), *The Growth and Maintenance of Linguistic Complexity*, John Benjamins.

Davis, Terry C., Michael S. Wolf, Pat F. Bass, Mark Middlebrooks, Estela Kennen, David W. Baker, Charles L. Bennett, Ramon Durazo-Arvizu, Anna Bocchini, Stephanie Savory, and Ruth M. Parker (2006), Low literacy impairs comprehension of prescription drug warning labels, *Journal of General Internal Medicine* **21** (8), pp. 847–851.

Davoodi, Elnaz, Leila Kosseim, and Matthew Mongrain (2017), A context-aware approach for the identification of complex words in natural language texts, *IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 97–100.

De Belder, Jan and Marie-Francine Moens (2010), Text simplification for children, *Proceedings of the SIGIR Workshop on Accessible Search Systems.*

De Clercq, Orphée and Véronique Hoste (2016), All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch, *Computational Linguistics* **42** (3), pp. 457–490.

DeKeyser, Robert (2005), What makes learning second-language grammar difficult? A review of issues, *Language Learning* **55**, pp. 1–25.

Deléger, Louise, Bruno Cartoni, and Pierre Zweigenbaum (2013), Paraphrase detection in monolingual specialized/lay comparable corpora, *Building and Using Comparable Corpora* pp. 223–241.

Devlin, Siobhan and Gary Unthank (2006), Helping aphasic people process online information, *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 225–226.

Devlin, Siobhan and John Tait (1998), The use of a psycholinguistic database in the simplification of text for aphasic readers, *in* Nerbonne, John, editor, *Linguistic Databases*, CSLI Publications, pp. 161–173.

Diessel, Holger (2004), *The Acquisition of Complex Sentences*, Cambridge University Press.

Elhadad, Noémie (2006), Comprehending technical texts: Predicting and defining unfamiliar terms, *AMIA Annual Symposium Proceedings*, pp. 239–243.

Fellbaum, Christiane (1998), *WordNet: An Electronic Database*, MIT Press.

Flesch, Rudolf (1948), A new readability yardstick, *Journal of Applied Psychology* **32** (3), pp. 221–233.

Gala, Nuria, Thomas François, and Cédrick Fairon (2013), Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons, *Proceedings of Electronic Lexicography in the 21st Century: Thinking Outside the Paper*, pp. 132–151.

Grefenstette, Gregory (1998), Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind, *Intelligent Text Summarization, AAAI Spring Symposium Series*, pp. 111–117.

Hawkins, John A. (2004), *Efficiency and Complexity in Grammars*, Oxford University Press.

Hirsh, David and Paul Nation (1992), What vocabulary size is needed to read unsimplified texts for pleasure, *Reading in a Foreign Language* **8**, pp. 689–689.

Horn, Colby, Cathryn Manduca, and David Kauchak (2014), Learning a lexical simplifier using Wikipedia, *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics*, pp. 458–463.

Hulstijn, Jan and Rick de Graaff (1994), Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? a research proposal, *AILA Review* **11**, pp. 97–112.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura (2003), Text simplification for reading assistance: A project note, *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, pp. 9–16.

Jauhar, Sujay Kumar and Lucia Specia (2012), Uow-shef: Simplex – lexical simplicity ranking based on contextual and psycholinguistic features, *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pp. 477–481.

Kaji, Nobuhiro, Daisuke Kawahara, Sadao Kurohash, and Satoshi Sato (2002), Verb paraphrase based on case frame alignment, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 215–222.

Kandula, Sasikiran, Dorothy Curtis, and Qing Zeng-Treitler (2010), A semantic and syntactic text simplification tool for health content, *AMIA Annual Symposium Proceedings*, pp. 366–370.

Klare, George (1976), A second look at the validity of the readability formulas, *Journal of Reading Behavior* (8), pp. 129–152.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007), Moses: Open source toolkit for statistical machine translation, *Proceedings of the 45th Annual Meeting of the ACL*, pp. 177–180.

Lison, Pierre and Jörg Tiedemann (2016), Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles, *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pp. 923–929.

Lupyan, Gary and Rick Dale (2010), Language structure is partly determined by social structure, *PloS one* **5** (1), pp. e8559.

Matausch, Kerstin and Birgit Peböck (2010), Easyweb – a study how people with specific learning difficulties can be supported on using the internet, *12th International Conference ICCHP - Computers Helping People with Special Needs*, pp. 641–648.

McWhorter, John H. (2011), *Linguistic Simplicity and Complexity: Why Do Languages Undress?*, Walter De Gruyter.

Medero, Julie and Mari Ostendorf (2011), Identifying targets for syntactic simplification, *Proceedings of Speech and Language Technology in Education Workshop*, pp. 69–72.

Miestamo, Matti, Kaius Sinnemäki, and Fred Karlsson, editors (2008), *Language Complexity: Typology, Contact, Change*, John Benjamins.

Miller, George A. (1995), Wordnet: a lexical database for English, *Communications of the ACM* **38** (11), pp. 39–41.

Nation, Paul (2001), *Learning Vocabulary in Another Language*, Cambridge University Press.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu (2017), Exploring neural text simplification models, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 85–91.

Noraset, Thanapon, Chandra Bhagavatula, and Doug Downe (2014), Adding high-precision links to Wikipedia, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 651–656.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer, pp. 219–247.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat, and R. Harald Baayen (2002), Experiences from the spoken Dutch corpus project, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 340–347.

Orăsan, Constantin, Richard Evans, and Ruslan Mitkov (2018), Intelligent text processing to help readers with autism, *in* Shaalan, Khaled, Aboul Ella Hassanien, and Fahmy Tolba, editors, *Intelligent Natural Language Processing: Trends and Applications*, Springer, pp. 713–740.

Paetzold, Gustavo H. and Lucia Specia (2016), Unsupervised lexical simplification for non-native speakers, *Proceedings of The 30th AAAI*, pp. 3761–3767.

Pallotti, Gabriele (2015), A simple view of linguistic complexity, *Second Language Research* **31** (1), pp. 117–134.

Petersen, Sarah E. and Mari Ostendorf (2007), Text simplification for language learners: A corpus analysis, *Speech and Language Technology in Education*, pp. 69–72.

Pienemann, Manfred (1998), *Language Processing and Second Language Development: Processability Theory*, John Benjamins.

Rayner, Keith and Susan A. Duffy (1986), Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity, *Memory & Cognition* **14** (3), pp. 191–201.

Rayner, Keith, Barbara R. Foorman, Charles Perfetti, David Pesetsky, and Margaret Seidenberg (2001), How psychological science informs the teaching of reading, *Psychological Science in the Public Interest* **2** (2), pp. 31–74.

Rello, Luz and Ricardo Baeza-Yates (2013), Good fonts for dyslexia, *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility.*

Rello, Luz, Clara Bayarri, Azuki Gorriz, Ricardo Baeza-Yates, Saurabh Gupta, Gaurang Kanvinde, Horacio Saggion, Stefan Bott, Roberto Carlini, and Vasile Topac (2013), Dyswebxia 2.0!: more accessible text for people with dyslexia, *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pp. 25:1–25:2.

Saggion, Horacio (2017), Automatic text simplification, *Synthesis Lectures on Human Language Technologies* **10** (1), pp. 1–137.

Saggion, Horacio, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic (2015), Making it simplext: Implementation and evaluation of a text simplification system for Spanish, *ACM Transactions on Accessible Computing (TACCESS)*, Vol. 6, pp. 14:1–14:36.

Schmid, Helmut (1994), Probabilistic part-of-speech tagging using decision trees, *Proceedings of the International Conference on New Methods in Language Processing*, pp. 154–164.

Schmid, Helmut (1995), Improvements in part-of-speech tagging with an application to German, *Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Sevens, Leen, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde (2017), Simplified text-to-pictograph translation for people with intellectual disabilities, *Proceedings of the 22nd International Conference on Natural Language and Information Systems (NLDB)*, pp. 185–196.

Sevens, Leen, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde (2018), Less is more: A rule-based syntactic simplification module for improved text-to-pictograph translation, *Data and Knowledge Engineering.*

Siddharthan, Advaith (2006), Syntactic simplification and text cohesion, *Research on Language and Computation* **4** (1), pp. 77–109.

Siddharthan, Advaith (2014), A survey of research on text simplification, *International Journal of Applied Linguistics* **165** (2), pp. 259–298.

Skadiņš, Raivis, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne (2014), Billions of parallel words for free: Building and using the EU Bookshop corpus, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 1850–1855.

Specia, Lucia (2010), Translating from complex to simplified sentences, *Proceedings of the Conference on Computational Processing of the Portuguese Language*, pp. 30–39.

Štajner, Sanja (2018), How to make troubleshooting simpler? Assessing differences in perceived sentence simplicity by native and non-native speakers, *Proceedings of the Second LREC Workshop on Improving Social Inclusion: Tools, Methods and Resources (ISI-NLP 2).*

Szmrecsanyi, Benedikt and Bernd Kortmann (2009), The morphosyntax of varieties of English worldwide: A quantitative perspective, *Lingua* **119** (11), pp. 1643–1663.

Tiedemann, Jörg (2012), Parallel data, tools and interfaces in OPUS, *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 2214–2218.

Trudgill, Peter (2011), *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*, Oxford University Press.

Vandeghinste, Vincent (2002), Lexicon optimization: Maximizing lexical coverage in speech recognition through automated compounding, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 1270–1276.

Vandeghinste, Vincent and Yi Pan (2004), Sentence compression for automated subtitling: A hybrid approach, *Proceedings of the ACL-workshop on Text Summarization*, pp. 89–95.

Vossen, Piek T. J. M., Attila Görög, Rubén Izquierdo, and Antal van den Bosch (2012), DutchSemCor: Targeting the ideal sense-tagged corpus, *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 584–589.

Vossen, Piek T. J. M., Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke (2013), Cornetto: a combinatorial lexical semantic database for Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, Springer, pp. 165–184.

Walker, Andrew, Advaith Siddharthan, and Andrew Starkey (2011), Investigation into human preference between common and unambiguous lexical substitutions, *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 176–180.

Wang, Tong, Ping Chen, John Rochford, and Jipeng Qiang (2016), Text simplification using neural machine translation, *AAAI Conference on Artificial Intelligence*, pp. 4270–4271.

Watanabe, Willian Massami, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio (2009), Facilita: reading assistance for low-literacy readers, *Proceedings of the 27th ACM International Conference on Design of Communication*, pp. 29–36.

Wilkens, Rodrigo, Alessandro Dalla Vecchia, Marcely Zanon Boito, Muntsa Padró, and Aline Villavicencio (2014), Size does not matter. Frequency does. A study of features for measuring lexical complexity, *Proceedings of the Ibero-American Conference on Artificial Intelligence*, pp. 129–140.

Williams, Sandra, Ehud Reiter, and Liesl Osman (2003), Experiments with discourse-level choices and readability, *Proceedings of the 9th European Workshop on Natural Language Generation*, pp. 127–134.

Woodsend, Kristian and Mirella Lapata (2011), Learning to simplify sentences with quasisynchronous grammar and integer programming, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 409–420.

Wrobel, Krzysztof (2016), PLUJAGH at SemEval-2016 task 11: Simple system for complex word identification, *Proceedings of the 10th SemEval*, pp. 953–957.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer (2012), Sentence simplification by monolingual machine translation, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pp. 1015–1024.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles (2015), Problems in current text simplification research: New data can help, *Transactions of the Association for Computational Linguistics* **3**, pp. 283–297.

Yatskar, Mark, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee (2010), For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia, *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 365–368.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych (2010), A monolingual tree-based translation model for sentence simplification, *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1353–1361.

Appendix A. Examples of successful simplifications

Original sentence	**Simplification**
De anderen hopen dat de curator het faillissement zo snel mogelijk zal *afhandelen*, zodat ze hun ontslagvergoeding kunnen krijgen.	De anderen hopen dat de curator het faillissement zo snel mogelijk zal *afwerken*, zodat ze hun ontslagvergoeding kunnen krijgen.
Google heeft in dat land zijn Europese hoofdkwartier gevestigd en een *aanzienlijk* deel van de winst die het haalt in Europa wordt ook in Ierland geboekt	Google heeft in dat land zijn Europese hoofdkwartier gevestigd en een *groot* deel van de winst die het haalt in Europa wordt ook in Ierland geboekt.
Het Belgische modelabel, dat in 1983 werd overgenomen door Edouard Vermeulen, heeft voor zijn eerste buitenlandse winkel een geschikte *locatie* gevonden in de Hooftstraat.	Het Belgische modelabel, dat in 1983 werd overgenomen door Edouard Vermeulen, heeft voor zijn eerste buitenlandse winkel een geschikte *plaats* gevonden in de Hooftstraat.
Ik ga ervan uit dat die 519 miljoen euro technisch is *gecorrigeerd*, maar het is niet zo dat de sociale zekerheid nu met een gat van 748 miljoen euro wordt geconfronteerd.	Ik ga ervan uit dat die 519 miljoen euro technisch is *verbeterd*, maar het is niet zo dat de sociale zekerheid nu met een gat van 748 miljoen euro wordt geconfronteerd
Het experiment gaat vijf jaar en twee maanden duren en begint op zijn vroegst eind 2019	*De test* gaat vijf jaar en twee maanden duren en begint op zijn vroegst eind 2019
Het gaat om een kostbaar product dat onder geen *beding* mag weglekken naar het criminele circuit, citeert de Volkskrant uit het wetsvoorstel	Het gaat om een kostbaar product dat onder geen *voorwaarde* mag weglekken naar het criminele circuit, citeert de Volkskrant uit het wetsvoorstel
De regering van president Donald Trump draait zo nogmaals belangrijke milieuregulering terug uit *het tijdperk* van zijn voorganger Barack Obama.	De regering van president Donald Trump draait zo nogmaals belangrijke milieuregulering terug uit *de tijd* van zijn voorganger Barack Obama.
In zijn *homilie* vroeg de paus een oplossing voor de situatie in Syrië	In zijn *preek* vroeg de paus een oplossing voor de situatie in Syrië
De Verenigde Staten - een grote *bondgenoot* van Israël - ...	De Verenigde Staten - een grote *vriend* van Israël - ...
Volgens de krant is het niet zeker dat de gesprekken tot een *deal* zullen leiden.	Volgens de krant is het niet zeker dat de gesprekken tot een *akkoord* zullen leiden.

Appendix B. Examples of unsuccessful simplifications

Original sentence	**Simplification**
Unnecessary simplification	
SilentKeys beschermt je niet alleen tijdens je surftochten thuis, waar je wellicht al een *krachtig* antivirusprogramma draait	SilentKeys beschermt je niet alleen tijdens je surftochten thuis, waar je wellicht al een *sterk* antivirusprogramma draait
Modification but no simplification	
In de tekst wordt opgeroepen tot *terughoudendheid* en het voorkomen van verdere escalatie ...	In de tekst wordt opgeroepen tot *reserve* en het voorkomen van verdere escalatie ...
Daarnaast wordt gehamerd op het recht om *vreedzaam* te betogen en wordt het verdriet van de lidstaten geuit met contact tot het verlies van onschuldige Palestijnse levens.	Daarnaast wordt gehamerd op het recht om *vredig* te betogen en wordt het verdriet van de lidstaten geuit met contact tot het verlies van onschuldige Palestijnse levens.
Eric Domb, *oprichter* van het dierenpark Pairi Daiza, stelde het project gisteren voor aan het grote publiek.	Eric Domb, *vader* van het dierenpark Pairi Daiza, stelde het project gisteren voor aan het grote publiek.
Er heerst bijgevolg *onzekerheid* over de grootte van de afzetmarkt voor auto's met lagere uitstoot in de VS .	Er heerst bijgevolg *twijfel* over de grootte van de afzetmarkt voor auto's met lagere uitstoot in de VS .
Commissaris Vandersmissen werd aan het einde van de *nationale* betoging aangevallen door een relschopper in een rood T-shirt .	Commissaris Vandersmissen werd aan het einde van de *landelijke* betoging aangevallen door een relschopper in een rood T-shirt .
Lexical/semantic errors	
Ik ben *intens* blij dat ik mag terugkeren naar het plekje waar het voor mij op de radio allemaal begon .	Ik ben *diep* blij dat ik mag terugkeren naar het plekje waar het voor mij op de radio allemaal begon .
Daarop beslisten bijna dertig landen om meer dan 140 Russische *diplomaten* uit te zetten .	Daarop beslisten bijna dertig landen om meer dan 140 Russische *politici* uit te zetten .
Van der Zwaan was *betrokken* bij de redactie van een rapport waarin de vervolging en veroordeling van een politieke rivaal van Janoekovitsj werd verdedigd .	Van der Zwaan was *gemengd* bij de redactie van een rapport waarin de vervolging en veroordeling van een politieke rivaal van Janoekovitsj werd verdedigd .

Appendix C. Examples of provided definitions and synonyms

Word	Annotation
Correct and clear	
fauna	Het geheel aan dieren in een gebied
biodiversiteit	Soortenrijkdom, het aantal verschillende diersoorten
teelt	Het kweken
bijgevolg	Daardoor
confrontatie	Aanvaring met andere persoon of zware problemen tussen groepen
Correct but not helping understanding	
curator	Iemand die door de rechter is aangewezen om het beheer te voeren over de bezittingen van een natuurlijk persoon of van een rechtspersoon (curatele, faillissement)
ontslagvergoeding	Ter compensatie van inkomensverlies uitgekeerde geldsom bij ontslag
tewerkgesteld	Een arbeidsbetrekking verlenen aan iemand
redactie	Het redigeren, de werkzaamheden voor het opstellen en rangschikken van artikelen
wetsvoorstel	Door de regering vervaardigd ontwerp van een wet die aan de volksvertegenwoordiging wordt voorgelegd
afzetmarkt	Markt waar een product of dienst kan worden afgezet
betogers	Iemand die meedoet met een protestmars
Wrong definition	
fiches	Geld vervangend (plastic) schijfje dat geld vervangt bij spelen
criminele *circuit*	Omloop (voor snelheidswedstrijden)
in de kiem *gesmoord*	[sudderen] Iemand of iets het ademen beletten
Verenigde *Staten*	Een gewestelijke standenvergadering (volksvertegenwoordiging), oorspronkelijk bestaande uit de drie standen: adel, geestelijkheid en burgers

Appendix D. Example of html output

```
Hij is de eerste die <span title="in een rechtszaak een oordeel uitspreken">
<span style="background-color: #D3D3D3">veroordeeld</span></span> wordt in de zaak
rond de mogelijke <a href="http://nl.wikipedia.org/wiki/Rusland" target="_blank">
<span title="een Slavische taal die gesproken wordt in Rusland">
<span style="background-color: #D3D3D3">Russische</span></span></a>
<span title="inmenging"><i>tussenkomst</i></span> in de
<a href="http://nl.wikipedia.org/wiki/Amerikaanse_presidentsverkiezingen_2016"
target="_blank"><span title="betreffende of komende van Amerika">
<span style="background-color: #D3D3D3">Amerikaanse</span></span>
<span title="de verkiezing van een president"><span style="background-color:
#D3D3D3">presidentsverkiezingen</span></span> van  2016</a> .
```