## Automating lexical simplification in Dutch

Bram Bulte, Leen Sevens and Vincent Vandeghinste

We discuss the design, development and evaluation of an automated lexical simplification tool for Dutch. Such text simplification tools are useful for a wide range of target populations (e.g. people with an intellectual disability, second language learners, aphasics and children). The ultimate aim of these tools is to adapt texts in such a way that they are easier to read and understand, while maintaining the original meaning as much as possible. To our knowledge, automated text simplification in Dutch has not been researched yet. We use a basic pipeline approach to tackle the problem of lexical simplification. Sentences are first pre-processed (i.e. tokenized, POS tagged and lemmatized) using TreeTagger (Schmid, 1994) and word sense disambiguation is performed using a tool based on support vector machines and trained on the data of the DutchSemCor project (Vossen et al., 2012). This tool links the identified word senses to lexical items in the Cornetto database (Vossen et al., 2013). The difficulty of each token in the input sentence is then estimated using two resources: (a) aggregated data coming from psycholinguistic studies into the average age of acquisition of Dutch words (Brysbaert et al., 2014), and (b) frequency information of Dutch tokens calculated on the basis of a large-scale corpus (of over 1000 million tokens) combining different sources, such as Subtitles2016 (Lison & 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). Cornetto is used to identify synonyms of words that have been identified as being difficult. Potential replacements are retrieved, ranked and selected, and a parsed version of the SONAR500 corpus is used to perform reverse lemmatization. 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. A neural language model is used to verify whether the selected replacement word fits the local context. A small development and test set, each consisting of 150 sentences taken from the Flemish newspaper De Standaard, are used to tune the system’s parametersand evaluate its output. In this study, a basic form of human evaluation is performed. Words in the test set are divided into two categories: (a) potentially difficult words, and (b) all other words. It is then verified whether the lexical simplification system changed the appropriate words and whether the proposed changes constitute an improvement in terms of decreasing the degree of difficulty, while maintaining the meaning and respecting grammar rules. We present quantitative results for different versions of the system that focus either on maximizing precision or recall, or on balancing both, and we show how the lexical simplification tool can be integrated in a full-fledged text simplification system for Dutch (Sevens et al., 2017). We expect that the presented research will lead to further developments in the field of augmentative and alternative communication.