Abstract

This paper presents the Wablieft corpus, a two million words corpus of a Belgian easy-to-read newspaper, written in Dutch. The corpus was automatically annotated with CLARIN tools and is made available in several formats for download and online querying, through the CLARIN infrastructure. Annotations consist of part-of-speech tagging, chunking, dependency parsing, named entity recognition, morphological analysis and universal dependencies. By making this corpus available we want to stimulate research into text readability and automated text simplification.

1 Introduction

Easy-to-read texts are texts that are targeted at people with a limited functional literacy. According to the United Nations Handbook of Household Surveys, "a person is functionally illiterate who cannot engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing and calculation for his own and the community’s development" (United Nations, 1984). Limited functional literacy is not an infrequent phenomenon. One out of ten people in Flanders, Belgium are low literate, and have trouble reading text on paper and on web sites. Many texts are too difficult, for example because they contain too many difficult words and/or complex sentences.

The goal of the Wablieft organisation is to address this issue on both sides: people who like to read easy texts can read the Wablieft newspaper, and people and organisations that want to publish easily readable texts can take training sessions at Wablieft, or can ask Wablieft to rewrite their texts. The Wablieft newspaper, 1 established in 1989, is a weekly newspaper in so-called clear language (“duidelijke taal” in Dutch), with more than 10,000 readers. An online archive of volumes published since 2009 is available on the organisation’s website, and it is this online archive which is now made available as an automatically linguistically annotated corpus of about two million word tokens.

Our aim is to make this corpus available to the natural language processing community as a target corpus of clear, easy-to-read Dutch for automatic simplification tools, to the applied linguistics community as an example of texts that were written with the explicit intention of being accessible, even by readers with low literacy skills, and to the linguistics community in general as a set of texts showing the linguistic characteristics of what is understood as clear writing in Flanders. The corpus has already been used by Bulté et al. (2018) in a lexical simplification task for the identification of difficult words and the selection of potentially easier replacements.

2 Related work

There are a number of approaches towards writing text while addressing the issue of limited literacy. Some texts are written with the general aim of being easy-to-read, others have children as their target group, and yet other texts aim at people with cognitive disabilities, so there is a large spectrum of possible target users.

---

1 This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/finalpaper:en-ukversion(tolicense, alicence)

1http://www.wablieft.be/
Probably the most well-known initiative in this respect is Wikipedia Simple English, which is written in so-called simple English. Amongst other things, its authors are instructed to use only the 1000 most frequent words of English. Other Wiki-initiatives are Wikikids for Dutch-speaking children and Vikidia for speakers of French, Italian, Spanish, English, Basque, Catalan, German, Russian, Greek and Sicilian.

In an academic context, a number of available easy-to-read corpora have been described and used, such as the Swedish LiSBarT corpus (Mühlenbock, 2009), a corpus for Brazilian Portuguese (Aluísio et al., 2008), the French CLEAR medical corpus (Grabar and Cardon, 2018), and a very small (227 sentences) corpus for Basque (Gonzalez-Dios et al., 2018). There have been some efforts to compile monolingual comparable corpora, aligning normal text with its easy-to-read variant. Alignment can be at the text level, the paragraph level or the sentence level. A list of English comparable corpora with an easy-to-read side can be found in Yaneva (2015), and also for French (Cardon and Grabar, 2018) and Brazilian Portuguese (de Medeiros Caseli et al., 2009) there have been efforts to create such a comparable corpus.

We are not aware of any such efforts for Dutch. Concerning more simple forms of Dutch, the Dutch data in the CHILDES project might be worth mentioning (MacWhinney, 2000), as well as the JASMIN speech corpus, consisting of recordings of Dutch speech by young people, non-native speakers, and elderly people (Cucchiarini et al., 2008). These two projects record active speech, whereas the Wablief corpus contains texts focusing on the passive language knowledge of the target users.

3 Corpus processing and availability

3.1 Creation of the metadata file

We received the data from our data providers as a zip file containing a number of text files, with no further information. The names of the text files were structured according to the regular expression pattern in (1), with the first set of digits indicating the newspaper volume (between the first set of parentheses), the article category (between the second set of parentheses), and the article number inside this category (between the third set of parentheses).

\(/\text{wa}[^{3,4}]\{\text{bi|ka|...}\}[^{\text{d}}].\text{txt/}\)  

The dates from the text files we received mostly agree with the publication dates, so we took this information as the publication dates of the articles in the metadata file, which we provide as a tab-separated value file (tsv). Table 1 presents the different categories that are distinguished in the newspaper and in the corpus.

3.2 Automated annotations

We used the LaMachine unified software distribution for Natural Language Processing to perform processing with Frog (Van den Bosch et al., 2007) and with Alpino (van Noord, 2006). LaMachine was available through the CLARIN Switchboard. Unfortunately, this is no longer the case, due to the separate registration procedure for the servers of the Radboud University in Nijmegen.

Frog is an NLP suite based on memory-based learning and trained on large quantities of manually annotated data. It automatically annotates the word tokens in Dutch text files. Frog’s output is available in two formats, which contain the same information: the FoLiA format (van Gompel and Reynaert, 2013), and the tab-delimited column-formatted output, one line per token (also known as CoNLL format). The ten columns contain (1) the token number within the sentence; (2) the token itself; (3) the predicted lemma; (4) the predicted morphological segmentation; (5) the predicted part of speech (PoS) tag; (6) the confidence with which the PoS tag was predicted; (7) the predicted named entity type, distinguishing

\(^2\)https://simple.wikipedia.org/wiki/Main_Page
\(^3\)http://www.wikikids.nl
\(^4\)http://www.wikidia.org
\(^5\)The presented numbers are those of the treebank version and might slightly differ from the non-treebank versions, as parsing might have failed in a number of cases.
\(^6\)https://proycon.github.io/LaMachine/
\(^7\)https://webservices-lst.science.ru.nl/register
between *person*, *organization*, *location*, *product*, *event*, and *miscellaneous*, using a IOB encoding;\(^8\) (8) the predicted phrase chunk in BIO encoding; (9) the predicted token number of the head word in the dependency graph; and (10) the predicted type of dependency relation with head word.

Alpino is a hybrid dependency parser for Dutch, which uses rule-based constraints combined with corpus-based statistics. It provides its own XML tree format, which is isomorphous to the syntax tree, unlike XML tree representations like FoLiA and TigerXML (Lezius et al., 2002). This makes it suitable for XPath and XQuery searches and scripts, allowing easy inclusion in CLARIN treebank query tools like GrETEL (Augustinus et al., 2017) and PaQu (Odijk et al., 2017). We also provide a Universal Dependencies annotation\(^9\) in CoNLL-UD format, in which the Alpino parses are automatically converted into CoNLL-UD using the script from Bouma and van Noord (2017).\(^10\)

### 3.3 Availability

The corpus can be downloaded for non-commercial purposes at the Dutch Language Institute, the CLARIN-B centre for Flanders.\(^11\) It comes with several (automatic) annotations and is delivered in a variety of formats: one directory per newspaper article, with one file per sentence in alpino XML; one XML file per newspaper article; frequency lists; the *frogged* versions of the files with automatic sentence detection, both in FoLiA and CoNLL format; the newspaper articles, one sentence per line (automatic detection); the original texts with paragraph and header markup, fixed for UTF-8; and one file with the Universal Dependencies annotation in CoNLL-UD format. We also included Wablieft in the GrETEL treebank query tool (Augustinus et al., 2017),\(^12\) so it can easily be queried.

### 4 Conclusions and future work

We presented the Wablieft corpus and how it was automatically annotated using the CLARIN infrastructure. The corpus is also made available through the CLARIN infrastructure.

In future work, we want to create a comparable corpus through aligning articles from regular newspapers with the articles in the Wablieft corpus, not unlike Cardon and Grabar (2018), on which we can train a text simplifier for Dutch. We also intend to use this corpus for language modelling of easy-to-read language.

---

\(^8\)https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging)
\(^9\)https://universaldependencies.org/
\(^10\)https://github.com/gossebouma/lassy2ud
\(^11\)https://ivdnt.org/downloads/taalmaterialen/tstc-wablieft-corpus-1-1
\(^12\)http://gretel.ccl.kuleuven.be/
References


Gertjan van Noord. 2006. At last parsing is now operational. In TALN 2006, pages 20–42.