Wablieft: An Easy-to-Read Newspaper Corpus for Dutch

Vincent Vandeghinste
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
`vincent.vandeghinste@ivdnt.org`

Bram Bulté
KU Leuven
Belgium
`bult@ccl.kuleuven.be`

Liesbeth Augustinus
KU Leuven
Belgium
`liesbeth@ccl.kuleuven.be`

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] `http://www.wablieft.be/`

Probably the most well-known initiative in this respect is Wikipedia Simple English,[2] 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[3] for Dutch-speaking children and Vikidia[4] 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 LäSBarT 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 Wablieft 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).

```
/wa(\d{3,4})(bi|ka|...)(\d).txt/                    (1)
```

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.[5]

3.2 Automated annotations

We used the LaMachine unified software distribution for Natural Language Processing[6] 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.[7]

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.vikidia.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

Dutch name	English name	number of sentences	number of words
Binnenland	Domestic	58,560	489,296
Blog	Blog	5,464	40,990
Buitenland	Foreign	40,953	337,536
Cijfer van de week	Number of the week	1,595	11,973
In de kijker	In the spotlight	49,438	398,366
Jaaroverzicht	Annual overview	301	2,484
Mening	Opinion	6,117	45,860
Samenleving	Society	23,660	189,847
Sport	Sports	24,747	196,697
Tip	Hint	12,869	100,513
Verhaal	Story	2,529	19,581
Voorpagina	Front page	5,121	42,527
Weetjes	Facts	19,927	156,222
Zomer	Summer	5,448	42,599
Total		256,729	2,074,491

Table 1: Categories of the Wablieft corpus

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

Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2017. GrETEL: A Tool for Example-based Treebank Mining. In *CLARIN in the Low Countries*, chapter 22, pages 269–280. London: Ubiquity Press.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: the autmoatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden.

Bram Bulté, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, 8:24–48, Dec.

Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93. Association for Computational Linguistics.

Catia Cucchiarini, Joris Driesen, Hugo Van hamme, and Eric Sanders. 2008. Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *LREC 2008*.

Helena de Medeiros Caseli, Tiago de Freitas Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline, Gasperin, and Sandra M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of CICLing*.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, 52(1):217–247, Mar.

Natalia Grabar and Rémi Cardon. 2018. Clear – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9. Association for Computational Linguistics.

Wolfgang Lezius, H. Biesinger, and Ciprian Gerstenberger, 2002. *Tiger-XML quick reference guide*.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Katarina Mühlenbock. 2009. Readable, legible or plain words - presentation of an easy-to-read Swedish corpus. In *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume Studia Linguistica Upsaliensia 8, pages 325–327.

Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In *CLARIN in the Low Countries*, chapter 23, pages 281–297. London: Ubiquity Press.

United Nations. 1984. *Handbook of Household Surveys, Revised Edition*, volume No. 31 of *Studies in Methods, Series F*. United Nations, New York.

Antal Van den Bosch, Gert Jan Busser, Walter Daelemans, and Sander Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium.

Maarten van Gompel and Martin Reynaert. 2013. Folia: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, Dec.

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

Victoria Yaneva. 2015. Easy-read documents as a gold standard or evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36. INCOMA Ltd. Shoumen, Bulgaria.