Enriching the ELEXIS-WSD corpus for Dutch

Martin Kroon
Dutch Language Institute
martin.kroon@ivdnt.org

Carole Tiberius
Dutch Language Institute
Leiden University
carole.tiberius@ivdnt.org

Sanne van der Wal
VU Amsterdam
s.f.vander.wal@student.vu.nl

Relevant UniDive working groups: WG1, WG2

1 Introduction

The ELEXIS-WSD-NL corpus is the Dutch part of the ELEXIS-WSD corpus, a parallel sense-annotated corpus originally developed within the H2020 ELEXIS project (see Martelli et al., 2021, 2023). The corpus consists of 2,024 parallel sentences taken from WikiMatrix (Schwenk et al., 2021) and is entirely manually curated, starting from the translations through to the 5 annotation layers that were added during the ELEXIS project: tokenization, subtokenization, lemmatization, part-of-speech tagging, and word sense disambiguation. The ELEXIS-WSD-NL corpus contains 34,923 tokens (6,488 unique lemmas), of which 13,551 have been semantically annotated[1] with Open Dutch WordNet (Postma et al., 2016). Within the UniDive COST action, the ELEXIS-WSD-NL corpus is being enriched with three more annotation layers: syntax, multiword expressions and named entities. In this paper, we report on the manual curation of the syntactic annotation layer, during which we often consult treebanks of other languages, seeking cross-linguistically consistent annotation. The manual curation of the corpus is carried out in INCEpTION[2].

2 Syntactic Annotation Layer

Within UniDive, syntactic annotations (according to the Universal Dependencies system; de Marneffe et al., 2021) have automatically been added to all subcorpora of the ELEXIS-WSD corpus using UDPipe (Straka, 2018). Although manual verification of the syntactic layer is optional, we are manually verifying the Dutch syntax to serve our broader goal of revising and refining UD syntax for Dutch corpora at the Dutch Language Institute more generally.

Among the most striking features of the syntactically parsed data for Dutch coming out of UDPipe, were the syntactic analyses of the subtokenized tokens in the corpus. In the original annotation process within the ELEXIS project, it was decided to split compounds that were not found in a general reference dictionary[3] of the language to avoid having a substantial amount of compounds in the corpus that would not be found in the sense inventory. In total, 620 compounds were subtokenized in the Dutch corpus and UDPipe provided a wide range of different analyses for them. To harmonize the analysis of the subtokenized compounds in the dataset, it was decided to automatically convert them to instances of the `compound` relation.

After that, an updated version of the corpus was uploaded in INCEpTION and we started manually verifying the syntactic UD annotations with a team of three (computational) linguists. As many errors were found in the trees, it was decided that, at least for the moment, all annotators would inspect the same sentences and discuss unresolved issues during meetings, after which a single master annotation would be maintained. In the remainder of this paper, we discuss some of the errors and problems encountered so far. For example, we found many errors in the syntactic analysis of the subtokenized compounds. Very often, dependents such as determiners and adjectives were attached to the first element of the compound, instead of its final element, i.e. its head (cf. Fig. 1).

An additional complexity related to these subto-

[1] Only content words have been assigned a sense.
[2] https://inception-project.github.io/
[3] For Dutch the Van Dale Dictionary was used as a guideline.

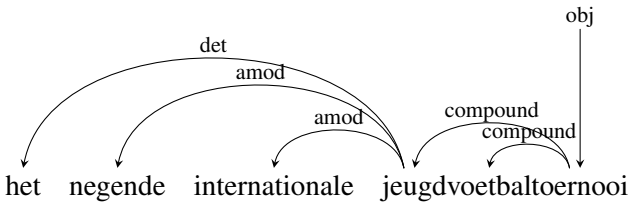

Figure 1: 'The ninth international youth football tournament': An example of an incorrect dependency analysis of a subtokenized compound. The determiner and adjective dependents should be attached to *toernooi* 'tournament', the head.

kenized compounds is that their internal hierarchy is not always immediately obvious (e.g. is *internetmarktwetgeving* 'internet market legislation' to be analyzed as *internet[marktwetgeving]* or *[internetmarkt]wetgeving*?), and that, in some instances, one of the components is a phrase instead of a single word (e.g. *Voetbal voor Vriendschapbeweging* 'Football for Friendship movement' which consists of the phrase *Voetbal voor Vriendschap* and *beweging*). For the latter, we decided to analyze the internal structure of the phrase *Voetbal voor Vriendschap*, but to keep the `compound` relation between the phrase and *beweging*.

We also noted that Dutch has a relatively high number of fixed relations in the corpus compared to some of the other languages. For instance, English and Estonian have, respectively, 129 and 23 fixed relations in the ELEXIS-WSD corpus, whereas Dutch has 442 (Tiberius et al., 2024). This seems to be caused by phrases such as *zo niet* 'if not', *voor het eerst* 'for the first time (lit. for the first)' and *meer dan (7.000 mensen)* 'more than (7.000 people)' (partially) receiving `fixed` relations, while in other languages they are fully analyzed. We deem the high number of `fixed` relations in the Dutch corpus to be undesirable, as it obscures (cross-)linguistically relevant features of the syntactic analysis, negatively impacting UD's universality in annotation and ability to capture diversity between languages. We therefore endeavour to replace as many `fixed` relations as possible.

The `flat` relation is also frequent in the Dutch corpus (933 instances; Tiberius et al., 2024). Although we do not oppose the `flat` relation in the cases of full names, borrowings, etc., it is often used in named entities that are syntactically analyzable. Notably, *Europese Unie* 'European Union' and *Verenigde Staten* 'United States' often receive a `flat` relation (and the tags PROPN), but are fully analyzable as a noun with a preceding adjective. Cases where the `flat` relation is fully syntactically analyzable are also replaced as much as possible, for the same reasons as `fixed`.

An outstanding challenge for Dutch concerns the subtokenization of R-pronominal adpositionals, such as *erop* 'on it'. These can be written as one word, but they can also be split with words ocurring between the R-pronoun and the adposition. Especially in cases where *er* functions as an expletive, not subtokenizing the one-word R-pronominal adpositional leads to a loss of information: either the adposition is not analyzed or the expletive is not analyzed, both of which are suboptimal. Figure 2 shows an example in which *erop* is subtokenized, making the presence of both the expletive *er* 'it' and the adposition *op* 'on' explicit.

A more cross-linguistically relevant challenge concerns whether *gaan* 'to be going to; to go' and *willen* 'to want' should be analyzed as an auxiliary. At the moment, they are analyzed as a verb having an `xcomp` child infinitive. This seems unfortunate because it often causes crossing relations, it obscures (cross-)linguistically relevant syntactic features such as concerning the formation of future tenses, the verbs do not select a particle *te* 'to', and they exhibit word orders that are characteristic of auxiliaries in Dutch. Additionally, *gaan* is synonymous with *zullen* 'shall', which is analyzed as an auxiliary. While French *aller* 'to be going to; to go' and *vouloir* 'to want' are analyzed the same as their Dutch translations, German *wollen* 'to want' is overwhelmingly analyzed as an auxiliary in German UD treebanks. No final decision has yet been made, but any decisions we do make on *gaan* and *willen* will be relevant to German or to French, and will further impact UD's universality in annotation and ability to capture diversity between languages.

Furthermore, it appears that UDPipe sometimes returns a different root than desired. One example is where the sentence-initial adverbial *zo niet* 'if not' is assigned the root rather than the main predicate. This is most likely a result of the ad-

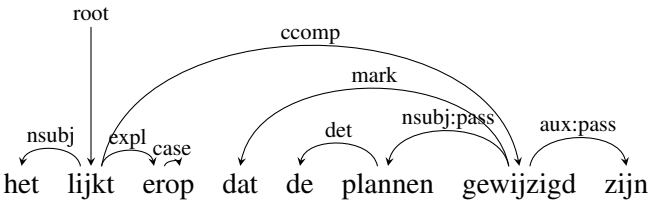

Figure 2: '... it seems plans have changed': An example in which the subtokenization of *erop* 'on it' makes the presence of both the expletive *er* 'it' and the adposition *op* 'on' explicit, whereas otherwise they would have been obscured.

verbial inaccurately being assigned the head of a `parataxis` relationship, since `parataxis` often assigns headedness to the root of a sentence. Additionally, in cases of reported speech, where the content of the reported speech should attach to the main verb's valency through a `ccomp` relation, UDPipe wrongfully connects the main verb of the reported content to the reporting verb through e.g. an `aux` relation. This places the `root` of the sentence in the clausal complement, rather than the main sentence.

Lastly, in copula constructions, it can be ambiguous which part of the sentence functions as the predicate (the `root`), and which part as the subject. Whereas some languages disambiguate the subject via morphosyntactic features (in certain cases), Dutch does not provide this information through its form. This remains a challenge for the systematic annotation of copula constructions.

3 Concluding remarks

This paper summarizes some of the challenges encountered during the manual verification of the syntactic annotation layer in the ELEXIS-WSD corpus for Dutch. In addition to syntax, the corpus is also enriched with manually annotated MWEs following the PARSEME guidelines[4] and Named Entities. Manual annotation and verification will be completed by August 2026.

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, and Tina Munda. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pages 377–395.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, and Sia Kolkovska. 2023. Parallel sense-annotated corpus ELEXIS-WSD 1.1. Slovenian language resource repository CLARIN.SI.

Marten Postma, Emiel Van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch Wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Carole Tiberius, Jaka Čibej, Jelena Kallas, Kertu Saul, Kadri Muischnek, and Simon Krek Krek. 2024. UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study. In *UniDive 2nd General Meeting (Naples, Italy)*, Naples, Italy.

[4] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/