UD Syntax for the ELEXIS-WSD Parallel Sense-Annotated Corpus: A Pilot Study

Carole Tiberius
Dutch Language Institute
The Netherlands
carole.tiberius@ivdnt.org

Jaka Čibej
Faculty of Arts,
University of Ljubljana,
Slovenia
jaka.cibej@ff.uni-lj.si

Jelena Kallas
Kertu Saul
Institute of the Estonian Language,
Estonia
jelena.kallas@eki.ee
kertu.saul@eki.ee

Kadri Muischnek
University of Tartu,
Estonia
kadri.muischnek@ut.ee

Simon Krek
Jožef Stefan Institute,
Slovenia
simon.krek@ijs.si

Relevant UniDive working groups: WG2, WG1

1 The ELEXIS parallel sense-annotated corpus

Within the H2020 ELEXIS project[1], a parallel sense-annotated corpus has been produced[2], an entirely manually-curated lexical-semantic resource covering 10 European languages - Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish - and featuring 5 annotation layers, i.e. tokenization, subtokenization, lemmatization, part-of-speech tagging and word sense disambiguation (see Martelli et al. (2021)). The parallel corpus consists of 2,024 sentences for each of these 10 languages (approximately 35,000 tokens per language). Only content words (i.e. between 14,000 and 18,000 words for each language) have been assigned a sense from the selected open source sense inventory (i.e. a dictionary) for the respective languages. + In the context of UniDive WG2 T2.2, an extension of this dataset is envisaged at two levels: (1) new languages, and (2) new annotation layers, i.e.

- Syntactic parse structure following UD[3].
- Annotation of multi-word expressions (MWEs) including verbal multi-word expressions (VMWEs) following the PARSEME annotation guidelines.[4] While MWE-annotation was undertaken by some of the 10 participating language teams, it has not yet been performed systematically across all 10 languages, and only spans of MWEs have been annotated (without assigning categories to the identified MWEs).
- Annotation of named entities. During the ELEXIS project, named entities were annotated in the English dataset. Similar to MWE-annotation, the subsequent annotation of named entities was left to the discretion of the individual language teams. In any case, only the spans of the named entities were annotated, while categories of named entities remain unassigned. Existing annotations from the English dataset could thus be used as a reference for NE-annotation in the remaining languages, possibly relying on the guidelines provided by Universal Named Entity Recognition project[5].

More details on the envisaged extensions can be found in the protocol for Task 2.2.[6]

2 Adding new annotation layers

In this presentation, we will focus on our plans and expectations for the second type of extension, i.e. with additional annotation layers. Starting with Dutch and Estonian, the automatically assigned syntactic annotation layer has been added using the UDPipe pipeline[7], and the resulting data has

[1] https://elex.is/
[2] hdl.handle.net/11356/1842
[3] https://universaldependencies.org
[4] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/index.php?page=home
[5] http://www.universalner.org/
[6] https://docs.google.com/document/d/1294ALlOb9Y5OVWF_M7BfFkX_t7wELA1l2xwM6Gsw8aQ/edit?usp=sharing
[7] https://lindat.mff.cuni.cz/services/udpipe/

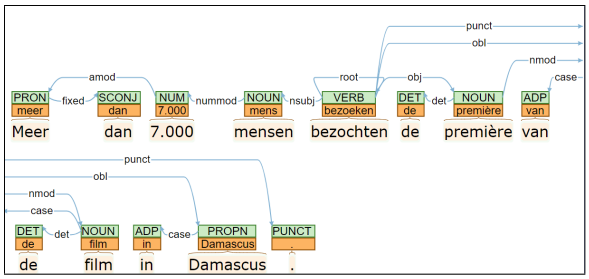

Figure 1: Annotation of a Dutch sentence in INCEpTION.

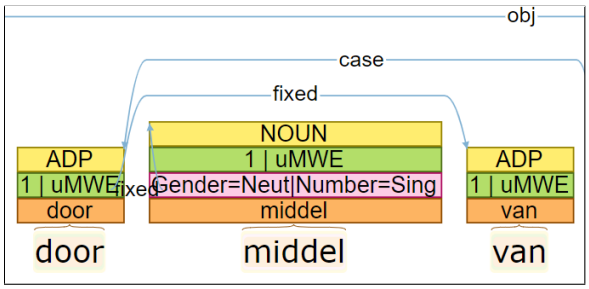

Figure 2: Annotation layers in INCEpTION: lemma (orange), morphosyntactic features (pink), POS-tags (yellow), MWEs (green), and UD-syntax (blue links).

been uploaded to INCEpTION[8], a platform for manual annotation (see Figures 1 and 2). Manual verification has just started and is expected to provide further insight in language-specific 'issues' with Universal Dependencies UD for these two languages. In the first phase, we have checked approximately 50 sentences per language. Some issues detected during this initial assessment necessitated the inclusion of morphological information in the corpus (which was absent in the published version). A separate layer for morphological features was added.

In addition, we can take advantage of the UD-parsing for the annotation of MWEs as MWE candidates can be automatically identified in the data using the three dependency relations `flat` (mostly names and complex numerals), `fixed` (functional MWEs, such as *op basis van* 'based on'), and `compound` (for identifying potential candidates of verb-particle constructions (VPC), light-verb constructions (LVC), and multi-verb constructions (MVC)). For Dutch, candidates of the inherently reflexive verb (IRV) class can also be identified from the UD annotation using the dependency relation `expl:pv`. Table 1 gives an overview of the number of times these particular relations are found in the corpus prior to manual verification.[9]

Dependency	English	Dutch	Estonian
FIXED	129	442	23
FLAT	282	933	482
COMPOUND	1333	291	202

Table 1: Number of occurrences of a particular dependency relation in the ELEXIS-WSD dataset

The subsequent annotation of MWEs can also take advantage of the subtokenization which was included in the original annotation process within ELEXIS for those cases in which the token was composed of two or more distinct lemmas. This was particularly useful and challenging for the Germanic languages in the dataset, where compounds are quite common and relatively dynamically generated, and more importantly: they are mostly written as a single word. The decision taken across all 10 languages in the corpus was that conventionalized compounds found in the dictionary of the language would be kept as such, while compounds not found in the dictionary would be split into lemmas included in the dictionary, so as to enable them to be semantically tagged. For instance, the Danish Dictionary (DDO)[10] was used as a guiding source for Danish and for Dutch compounds were subtokenized if they did not occur in the Van Dale dictionary[11]. Later, this criterion was slightly relaxed for Dutch and some other transparent compounds were also subtokenized, as otherwise a substantial number of compounds would not be found in the sense inventory. Overall, 620 compounds were subtokenized in the Dutch dataset. A close inspection and analysis of these cases can potentially provide useful examples for the extension of the PARSEME guidelines for noun compounds.

One of the applications of the resulting corpus will be the possibility for interlinking MWE lexicon entries with their occurrences in corpora, including the development of a lexicon-corpus interface.

3 Acknowledgments

The presented work was partly supported by the Estonian Research Council grant (PRG 1978) and

[8] https://inception-project.github.io/

[9] Note that this is a simple count and that multiple occurrences of a particular relation may actually belong to one MWE.

[10] https://ordnet.dk/ddo_en

[11] https://www.vandale.nl/

the research programme *Language Resources and Technologies for Slovene (P6-0411)* funded by the Slovenian Research and Innovation Agency.

References

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.