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

Carole Tiberius[1], Jaka Čibej[2], Jelena Kallas[3], Kertu Saul[3], Kadri Muischnek[4], Simon Krek[5]

[1]Instituut voor de Nederlandse Taal , The Netherlands, [2]Faculty of Arts, University of Ljubljana, Slovenia, [3]Institute of the Estonian Language, Estonia, [4] University of Tartu, Estonia, [5]Jožef Stefan Institute, Slovenia

What is the ELEXIS corpus?

An entirely manually-curated lexical-semantic resource available in ten European languages combining corpora and sense inventories (dictionaries and WordNets)

- Origin: WikiMatrix
- Number of sentences per language: 2,024
- Identical sentences in 10 languages
- Manually checked translations
- Manually checked annotations

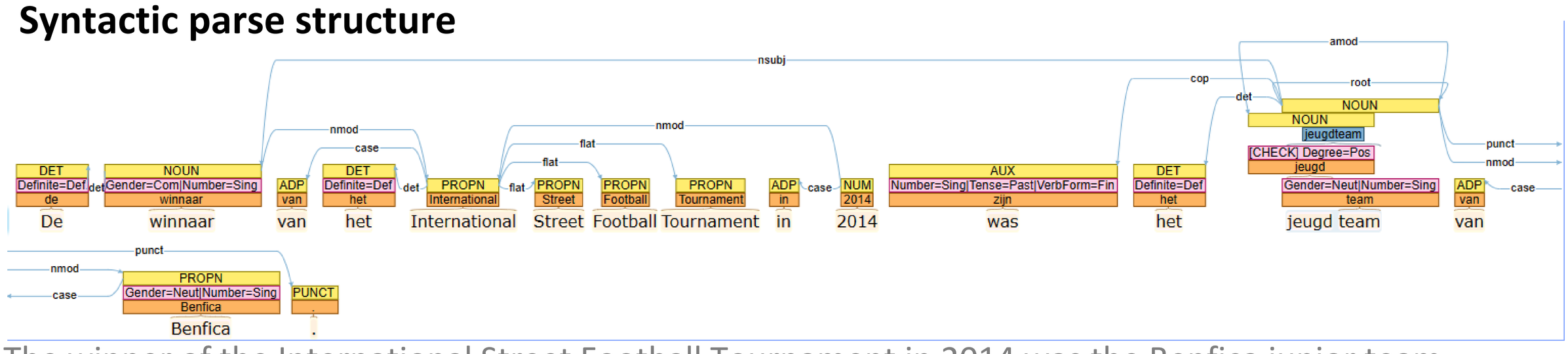
ELEXIS corpus and Unidive

Possible extension of the current dataset with additional languages and additional annotation layers

- annotation of **multiword expressions** following the PARSEME annotation guidelines
- annotation of **named entities**
- **syntactic parse structure** following Universal Dependencies

Annotation in INCEpTION

Syntactic parse structure

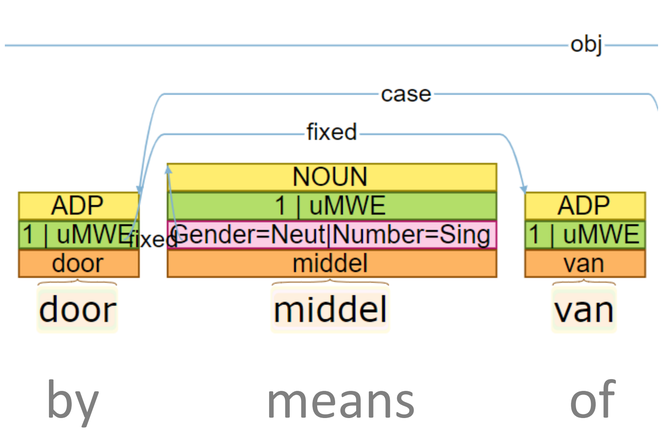


The winner of the International Street Football Tournament in 2014 was the Benfica junior team.

Availability (repository & online concordancer)

Martelli, Federico, et al. 2022**. Parallel sense-annotated corpus ELEXIS-WSD 1.0**, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042.

Multiword expressions



Design of the dataset

Martelli, Federico, et al. 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. 377-395.

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

Relevance: UniDive **WG1** and **WG2**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.