Improving HDT-UD

Nina Böbel
Heinrich Heine University Düsseldorf, Germany
nina.boebel@hhu.de

About the Project

HDT-UD (Borges Völker et al. 2019), a conversion of the Hamburg Dependency Treebank (Foth et al. 2014), is the largest UD annotated treebank and consists of 3,8 million tokens. Even though the conversion has a high precision of over 97% (Borges Völker et al. 2019)[1], there are some aspects that can lead to difficulties when working with HDT-UD.

For example, HDT-UD is the only German treebank in which only indicative and imperative modes are annotated. There is no annotation of the subjunctive; instead, subjunctive verbs are normally annotated as indicative. Another problem is the lack of relatedness in the annotation of complex connectors such as the complex conditional connectors *angenommen dass* (‘assuming that’) or *für den Fall dass* (‘for the case that’). While comparable expressions are annotated as ‘fixed’ in Swedish, for example, they are not given connector status in HDT-UD and are not displayed as belonging together (Weissweiler et al. 2024). Other problems concern, for example, the presence of more than one subject in a sentence, the correct annotation of the semantically empty subject *es* (‘it’) or the assignment of the correct subordinate clause type in *dass* (‘that’) sentences.

Problems such as those mentioned here can lead to HDT-UD not being used for linguistic analyses because the search queries cannot be made precisely enough. Although such a large treebank is available, many researchers tend to use smaller corpora. For instance, the query for irreal conditional sentences is difficult due to the lack of subjunctive annotation.

Project Phase

Currently the team that is working on this project contains four people, including experts in the field of creating/revising treebanks as well as in the field of German linguistics (and thus the research-led use of the treebank). The project is in its initial phase; so far, selected problems have been systematically recorded by means of search queries in GrewMatch (Guillaume, 2021). The adjustments that can be made to the annotations to correct the problems have also been recorded.

Two aspects stand out as the next necessary steps: firstly, a workflow must be developed that allows the systematic and – if possible – complete detection of annotation errors in HDT-UD. Secondly, an effective procedure for recording the necessary changes must be created.

The project is therefore at a stage where the fundamental goal (increasing the

[1] See also https://github.com/UniversalDependencies/UD_German-HDT/blob/master/README.md

correctness and thus also the attractiveness of HDT-UD) has been established. First annotation errors/problems have already been identified and solutions worked out. The specific continuation still raises questions for various reasons.

Open Questions and Problems

The extent to which manual work on error correction can be simplified by automatic or semi-automatic processes has not yet been fully determined. In general, the development of a fixed, effective workflow is probably important in order to make the work at HDT-UD productive on the one hand and to make it easier for potential new people to join and participate in the project on the other.

One problem that probably exists is the danger of being guided by one's own linguistic focus. The open question here is also whether a semi-automatic approach can support the independent detection of annotation errors/problems.

At the beginning, we concentrated primarily on systematic errors, often relating to syntax. Another question concerns the handling of fixed multiword expressions. It is not yet clear how sensitive HDT-UD is to errors in this respect.

The open questions therefore mainly concern the workflow; the work on HDT-UD should be as effective as possible.

Benefits of the Training School

The Summer School offers several benefits for this project. In addition to specific tips and hints relating to the practical implementation of our project, general advice is also needed to ensure an effective workflow and thus a time- and resource-saving approach. The contact and exchange with other researchers who are also improving or even setting up new treebanks promises valuable opportunities to exchange ideas and information.

The aim is also to expand knowledge about the structure of corpora. Since an existing treebank is used for this project, some decisions regarding the structure and annotation do not necessarily appear to make sense; extended background knowledge can facilitate the understanding and assessment of such decisions.

Keywords

Universal Dependencies, HDT-UD, treebank

References

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. *HDT-UD: A very large Universal Dependencies Treebank for German.* In Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019), pages 46–57, Paris, France. Association for Computational Linguistics.

Bruno Guillaume. 2021. *Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion.* In Proceedings of the 16th Conference of the European Chapter of the

Association for Computational Linguistics: System Demonstrations, pages 168–175, Online. Association for Computational Linguistics.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. *Because Size Does Matter: The Hamburg Dependency Treebank.* In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Language Resources and Evaluation Conference 2014, pages 2326–2333, Reykjavik, Iceland, may. LREC, European Language Resources Association (ELRA).

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. *UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies.* arXiv preprint arXiv:2403.17748.