Extending the Treebank Query Application GrETEL and the Treebank Editor TrEd

Sheean Spoel	Gerson Foks	Jan Odijk
Utrecht University / Utrecht	Utrecht University / Utrecht	Utrecht University / Utrecht
s.j.j.spoel@uu.nl	g.foks@uu.nl	j.odijk@uu.nl

In the demonstration we will show the extensions we made to the existing treebank query application GrETEL (Augustinus, Vandeghinste & Van Eynde, 2012). This application allows researchers to search for syntactic structures in Dutch treebanks using user-specified XPATH queries. It is also possible to generate a query based on an example sentence and specify the properties to search through a graphical interface, making it possible to search effectively with no or a minimal knowledge of a formal query language such as XPATH. Regardless of search method (example or query-based) the matched tree structures are returned, showing their direct context and with a visualization of the syntactic structure. It is also possible to download the results for further analysis. The texts used for constructing the treebanks were parsed using the Alpino parser (Bouma, Van Noord & Malouf, 2001), manually verified and corrected (such as LASSY, Van Noord et al., 2013) and are then indexed using a BaseX server (Grün, 2010).

This work builds on extensions made earlier (Odijk, Klis & Spoel, 2018) to the GrETEL application, as part of the CLARIAH WP3 project. These included the ability to upload an existing corpus (plain text or CHAT format) or an already parsed Alpino treebank through a web interface. The update also added the ability to analyse the matched treebank structures on the available metadata (such as age of the speaker or their role) and the nodes' lemma or POS-tag using an interactive pivot table. Furthermore the XPATH-query entry was improved, allowing real-time validation, suggestions and entry of PaQu (Odijk, Noord, Kleiweg & Tjong Kim Sang, 2017) macros.

The extensions which will be demonstrated here, further improve the usability of the application and add additional functionality. They include (1) the ability to upload one's own existing FoLiA or TEI XML based corpus and search it for existing metadata and syntactical information (2) the ability to analyse specific nodes in a query using a graphical representation of the query tree, also allowing the specification of additional node properties for analysis (such as case, verb root and number) and (3) a more decoupled software architecture and improved user interface following current industry standards and best practices. These extensions have been made available as an update to the latest version (Version 4) of GrETEL.

As part of the AnnCor project the Dutch CHILDES transcripts of Van Kampen (2009) are being enriched with syntactic information and made searchable using the updated version of GrETEL. Because child speech is not always parsed correctly using the Alpino parser, the parses are manually checked and corrected by annotators. To facilitate their work, extensions have been developed for the tree editor TrEd (Pajas & Fabian, 2013) to make it easier to correct existing annotations and to reparse on request. These extensions are also described as part of this demonstration.

Acknowledgements

This work was financed by the Utrecht University internal AnnCor research infrastructure project and by the CLARIAH-CORE project funded by the Dutch National Science Foundation (NWO).

References

Augustinus, L., Vandeghinste, V. & Van Eynde, F., (2012) Example-based treebank querying. In *Proceedings of the 8th international conference on language resources and evaluation (LREC 2012)* (pp. 3161–3167), ELRA.

Bouma, G., Van Noord, G. & Malouf, R., (2001) Alpino: Wide-coverage computational analysis of Dutch, *Language and Computers*, *37*, 45–59.

Grün, C., (2010) *Storing and querying large XML instances* (Doctoral dissertation, University of Konstanz, Konstanz, Germany), retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-127142

Odijk, J., Klis, M. & Spoel, S., (2018) Extensions to the GrETEL treebank query application. In *Proceedings of the 16th international workshop on treebanks and linguistic theories* (pp. 46–55), retrieved from http://www.aclweb.org/anthology/W17-7608

Odijk, J., Noord, G. v., Kleiweg, P. & Tjong Kim Sang, E., (2017) The parse and query (PaQu) application, retrieved from https://doi.org/10.5334/bbi

Pajas, P. & Fabian, P., (2013) Tree editor TrEd, Prague Dependency Treebank, Charles University, Prague, retrieved from http://ufal.mff.cuni.cz/tred

Van Kampen, J., (2009) The non-biological evolution of grammar: Wh-question formation in Germanic, *Biolinguistics*, *3*(2-3), 154–185.

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., … Vandeghinste, V., (2013) Large scale syntactic annotation of written Dutch: Lassy. In *Essential speech and language technology for Dutch* (pp. 147–164), Berlin: Springer, retrieved from https://doi.org/10.1007/978-3-642-30910-6_9