CLARIN Flanders: new prospects

Vincent Vandeghinste
Instituut voor Nederlandse Taal, Netherlands
vincent.vandeghinste@ivdnt.org

Els Lefever
Ghent University, Belgium
els.lefever@ugent.be

Walter Daelemans
University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

Tim Van de Cruys
University of Leuven, Belgium
tim.vandecruys@kuleuven.be

Sally Chambers
Ghent University, Belgium
sally.chambers@ugent.be

Abstract

We describe the creation of CLARIN Belgium (CLARIN-BE) and, associated with that, the plans of the CLARIN-VL consortium within the CLARIAH-VL infrastructure for which funding was secured for the period 2021-2025.

1 Introduction

We describe the efforts that have been undertaken to ensure the re-entry of Flanders, the Dutch-speaking community in Belgium, into the world of the CLARIN ERIC, in section 2. We also describe the new linguistic tools and services that are planned to be developed within the second phase of the CLARIAH-VL project, which has recently started, in section 3.

2 CLARIN-VL and CLARIN-BE

Given that, in Belgium, most funding of scientific research happens at the level of the communities, of which there are three: the *Vlaamse Gemeenschap* (the Flemish Community – Dutch speaking), the *Fédération Wallonie-Bruxelles (FWB)* (the Federation Wallonia-Brussels – French speaking) and the very small *Deutschsprachige Gemeinschaft* (the German-speaking Community), and given that members or observers of CLARIN ERIC have to be countries or intergovernmental organizations,[1] it follows that Flanders cannot be a member of CLARIN directly.

In the past, Flanders participated in CLARIN through the international organization *Nederlandse Taalunie* (Dutch Language Union), but such a construction was no longer possible after 2018, resulting in a *Flexit* from CLARIN.

The only possible way to become a member of CLARIN ERIC was to apply for political support from Flanders (without funding) for the formal founding of CLARIN Belgium and payment of the CLARIN ERIC membership fees by the Belgian Science Organization BELSPO,[2] in a similar construction as the DARIAH infrastructure.[3] Such political support was granted and membership of Belgium should become a fact in 2021.

The Flemish CLARIN consortium consists of several research groups from three Flemish universities, and the *Instituut voor de Nederlandse Taal* (INT – Dutch Language Institute)[4] as third party, and is open for more research groups. INT is located in the Netherlands but is partly funded by Flanders, and is the *de facto* CLARIN-B centre for Flanders, serving as a data depositing centre. Currently different data sets and tools developed in Flanders have been integrated into the CLARIN infrastructure already, and more will follow, see section 3. CLARIN-VL also focuses on user involvement, through the organisation of CLARIN information sessions and lectures during different courses.

[1] cf. https://www.clarin.eu/content/participating-consortia
[2] https://www.belspo.be
[3] https://be.dariah.eu/
[4] http://www.ivdnt.org

In the meantime, efforts are undertaken to involve users from the French-speaking community in Belgium to take part actively in CLARIN Belgium, and we expect contributions to CLARIN from several of these research groups once Belgian membership has been formally established.

3 CLARIAH-VL

CLARIN-VL decided to join forces with DARIAH Flanders in the FWO International Research Infrastructure project *CLARIAH-VL*, and has secured funding until early 2025, for the development of several infrastructural services. This section presents our plans and ongoing work, so that CLARIN users know what to expect, and other CLARIN members know what we are working on, in order to promote cooperation and avoid parallel development of similar tools and resources.

We aim to develop a *Digital Text Analysis Dashboard and Pipeline* for processing both Dutch texts and parallel texts. There are already several pipeline approaches available (Bel, 2010; Hinrichs et al., 2010; Zinn, 2018; van der Sloot et al., 2018), but these are often limited to linguistic analysis — tokenization, pos-tagging, lemmatization, named entity labeling, dependency parsing. We aim at an approach which re-uses existing (CLARIN) tools and pipelines, but is extended with a variation of models and several natural language understanding analysis layers.

The user-friendliness of the design will be ensured through a user-centred approach involving also non NLP-users. To this end, a list of dedicated use cases will be defined from both the NLP and digital humanities research communities in Flanders and will be worked out in detail. Users will also be consulted for personalization of the dashboard and adapting it to their own needs.

An example of such a use case could be the Spoken Academic Belgian Dutch corpus, which is currently under development and which needs to be speech recognized, manually corrected and (automatically) linguistically annotated. Another use case is the processing of parallel data, with sentence and word alignment tools and extended corpus search functions to allow searching in parallel data with Blacklab (de Does et al., 2017), a cooperation with CLARIAH-NL, and with an example-based query engine similar to PolyGrETEL (Augustinus et al., 2016). A final digital humanities use case could be pipelines for the text and data mining of sub-corpora of digitised newspapers from KBR, the Royal Library of Belgium's BelgicaPress.[5]

For a number of NLP tasks, different alternatives are available in different forms and programming languages. We will benchmark existing tools and new models. The results allow users to curate which alternatives to integrate in the pipeline. This includes (re)training tools on existing resources, such as creating state-of-the art methods for language modelling of historical Dutch or specialized text corpora (e.g., medical text, legal text, etc).

A first set of tools that will be benchmarked are **linguistic processing tools**, which enrich corpora in plain text format with linguistic information, such as part-of-speech tags, lemmas (basic form as found in a dictionary), named entity information and chunk information. Existing and new implementations for English, French, Dutch and German will be tested. Amongst these we will test LeTs Preprocess: the Multilingual LT3 Linguistic Preprocessing Toolkit (Van de Kauter et al., 2013), which provides linguistic annotation for 4 languages (English, Dutch, French, German): tokenisation, part-of-speech tagging, lemmatisation, chunking and named entity recognition. Other tools that will be evaluated are the memory-based linguistic analysis tool Frog (van der Sloot et al., 2018), and its successor in the deep learning paradigm DeepFrog[6], which was developed in CLARIAH-NL, and which provides neural network models for Dutch NLP, part-of-speech and named entity tagging.

The previous two pipelines will be compared to other state-of-the-art open source libraries, such as for instance the Spacy libraries for text processing[7], or the Stanza Python NLP package[8].

Secondly, we focus on the benchmarking of NLP tools for **natural language understanding**. Recent machine learning methods based on neural transformer architectures have greatly improved the state of

[5] https://www.belgicapress.be
[6] https://github.com/proycon/deepfrog
[7] https://spacy.io/usage/linguistic-features
[8] https://stanfordnlp.github.io/stanza/

the art for a wide range of natural language understanding (NLU) tasks. In order to provide an extensive testbed for language-specific NLU models, we will develop a suite of NLU evaluation tasks, similar to the well-known GLUE (Wang et al., 2018) and SuperGlue (Wang et al., 2019) evaluation frameworks for English. Specifically, we will focus on an evaluation suite for Dutch. Due to the laborious nature of manual labeling, we will mainly focus on the semi-automatic construction of evaluation tasks (construction of datasets from web forums and resources, prediction of discourse markers, ...), as well as the compilation of existing evaluation sets within one overarching suite. In a second stage, the construction may be complemented by manually labeled evaluation instances, gathered by means of a voluntary crowdsourcing setup.

Additionally, we will explore the application of neural network models for the **search and extraction of linguistic structures**. There is corroborating evidence that self-supervised transformer architectures implicitly encode a wide range of linguistic knowledge, from part of speech information over syntactic structure to co-reference information (Peters et al., 2018; Hewitt and Manning, 2019; Clark et al., 2019). We will investigate to what extent such implicit linguistic representations might be exploited as a tool for linguistic analysis. More specifically we will investigate whether the linguistic information present in the models might be distilled for the purpose of similarity computations. Such a process would allow to automatically harvest a corpus of linguistically similar structures, in order to support linguistic analysis. Moreover, as transformer architectures simultaneously encode syntactic and semantic information in their contextualized representations, this would allow to automatically harvest syntactically disparate realization of similar semantic content, providing an adequate means for a linguistic analysis of the syntax-semantics interface.

The second batch of tools that will be integrated in the pipeline are tools aiming at solving specific natural language processing and understanding tasks, amongst which:

- **Sentiment analysis**: a pipeline annotating text strings of varying length (e.g., words, chunks, sentences, reviews, documents, etc.) with polarity information (positive, negative, neutral) and unsupervised learning techniques for adapting dictionary-based sentiment analysis tools to new domains.
- **Emotion detection**: a pipeline annotating text strings with more fine-grained emotion information. Two types of emotion detection approaches will be evaluated: (1) classification approaches providing categorical labels (e.g. anger, disgust, fear, joy, sadness, surprise), and (2) dimensional models representing emotions as vectors in a multidimensional space, defined by three axes: valence (unhappiness/happiness), arousal (calmness/excitement) and dominance (submission/dominance). Every emotional state is then described by the combination of the values on these three axes.
- **Document similarity clustering**: a pipeline which allows uploading a set of documents and provides document clusters based on their similarity according to different models.
- **Topic modelling**: a pipeline extracting topics from text corpora using traditional (LDA, NMF) and more recent (top2vec) embedding based approaches. Visualisation in terms of topic maps and timelines.
- **Stylometry**: a pipeline for unsupervised and supervised machine learning based stylometry (authorship attribution, age, gender personality profiling) allowing the combination of various linguistic and stylometric information sources and adding new information sources, e.g., figurative language detection.

The third batch of tools that will be integrated are tools aiming at processing multilingual data:

- **Sentence alignment**: integration of a tool that 'aligns' (makes relations explicit) between the sentences of two texts that are literal translations.
- **Word alignment**: integration of a tool that identifies relationships among the words in a bitext, 'aligning' words that are translations of one another. Word alignment typically starts from pairs of sentences that have been sentence-aligned before.

Public datasets that have been processed with certain tools will be made available from within the dashboard, allowing users to search within these datasets.

Finally, a Help Desk will be developed to help users by advising users and tailoring tools to specific use cases or domains, as well as deal with feedback on annotation and analysis errors, leading to improved models. This help desk can be contacted through `servicedesk@ivdnt.org`. This Help Desk will provide information similarly to K-Dutch, the CLARIN Knowledge Centre for Dutch, which has been recognized by CLARIN-ERIC this summer.[9]

4 Conclusions

We are very pleased to announce the re-entry of Flanders in CLARIN through Belgian membership, and have presented our work plan for the next funding period.

References

L. Augustinus, V. Vandeghinste, and T. Vanallemeersch. 2016. Poly-GrETEL: Cross-lingual example-based querying of syntactic constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3549–3554, Portorož, Slovenia, May. European Language Resources Association (ELRA).

N. Bel. 2010. Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies. Panacea. *Procesamiento del Lenguaje Natural*, 45:327–328.

K. Clark, U. Khandelwal, O. Levy, and C.D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

J. de Does, J. Niestadt, and K. Depuydt. 2017. Creating Research Environments with BlackLab. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 20. Ubiquity Press, London, Dec.

J. Hewitt and C.D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

E.W. Hinrichs, M. Hinrichs, and T. Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

M. Peters, M. Neumann, L. Zettlemoyer, and W. Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

M. Van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste. 2013. LeTs Preprocess: the Multilingual LT3 Linguistic Preprocessing Toolkit. *Computational Linguistics in the Netherlands Journal*, pages 103–120.

K. van der Sloot, I. Hendrickx, M. van Gompel, A. van den Bosch, and W. Daelemans. 2018. Frog, a natural language processing suite for dutch. Reference Guide. Technical Report Language and Speech Technology Technical Report Series 18-02, Radboud University.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

C. Zinn. 2018. Squib: The language resource switchboard. *Computational Linguistics*, 44(4):631–639, December.

[9] `https://kdutch.ivdnt.org/`