Computational Linguistics in the Netherlands Journal 12 (2022) 3-5 Submitted 12/2022; Published 12/2022

Preface

Eva Vanmassenhove* E.O.J.VANMASSENHOVE@TILBURGUNIVERSITY.EDU
Mirella De Sisto* M.DESISTO@TILBURGUNIVERSITY.EDU
Raquel G. Alhama* RGALHAMA@TILBURGUNIVERSITY.EDU
Tomas O. Lentz** T.O.LENTZ@TILBURGUNIVERSITY.EDU
Jan Engelen** J.A.A.ENGELEN@TILBURGUNIVERSITY.EDU
Dimitar Shterionov* D.SHTERIONOV@TILBURGUNIVERSITY.EDU

* *Department Cognitive Science and Artificial Intelligence, School of Humanities and Digital Sciences, Tilburg University, The Netherlands*

** *Department of Communication and Cognition, School of Humanities and Digital Sciences, Tilburg University, The Netherlands*

The 12^{th} special issue of the Computational Linguistics in the Netherlands (CLIN) Journal contains a selection of papers that were presented at the 32^{nd} edition of the annual CLIN conference (CLIN32). The CLIN conference takes place on a yearly basis and invites researchers working on a wide range of topics related to computational linguistics to share and present their work. This year's edition took place, in person, in the Koning Willem II stadium in Tilburg, after last year's virtual edition organized by Ghent University.

Aside from theoretical and applied work on all aspects of computational linguistics and natural language processing (NLP), there were two special tracks. One track was dedicated to work on language technology for Dutch Sign Language (NGT) and Flemish Sign Language (VGT).[1] The other track focused on language technology for Dutch and Flemish online (forum) discussions.[2] In addition to the special tracks, a panel discussion (panelists: Antal van den Bosch, Dimitar Shterionov and Simone Ashby) took place where the current state-of-the-art of language technology and related outstanding questions were debated, focusing particularly on the Low Countries. In total, 203 participants from 30 different institutions registered for CLIN32.

A total of 114 abstracts was received, of which 110 were accepted. Forty of the accepted abstracts were presented during one of the three presentation sessions. The remaining seventy were accepted as posters and presented during one of the two afternoon poster sessions. The topics covered during the conference reflect the current global trends in computational linguistics and NLP, with the most frequently reappearing topics being Machine Translation, BERTology (specifically for Dutch), corpora creation, speech (recognition, translation, analysis) and grounding and evaluation. Aside from that, the special tracks attracted many contributions, particularly the track on Sign Language (focused on NGT and VGT) with six oral presentations and two posters. The special track on Online Discussions featured submissions on current issues such as moderation and topic mining on platforms such as Reddit, and automatic analysis of reviews. The topics presented at CLIN32 also reflected some more recent research lines related to interpretability and explainability, inclusiveness and (gender and linguistic) bias, and two presentations related to COVID-19.

We were very happy to welcome two invited speakers: (i) Anna Rogers, assistant professor at the Center for Social Data Science of the University of Copenhagen and a visiting researcher with the RIKEN Center for Computational Science (Japan); and (ii) Jan-Willem van de Meent, associate professor at the University of Amsterdam who is also affiliated with the Northeastern University. Rogers discussed in her talk 'When does a machine "understand" what it "reads"' the complexities of attaching the label 'understanding' to any language model. Van de Meent delved into another

1. https://signon-project.eu/
2. https://better-mods.uvt.nl/

very timely topic 'What is Next for Large Language Models' where he addressed the future of large Language models.

From the 21 submissions we received for this year's volume, 16 were eventually selected for the 12^{th} special issue of the CLIN journal. The papers reflect the breadth of computational linguistics research conducted in and around the Low Countries as well as the special tracks presented at CLIN32. Two papers are by authors who participate in the European SignON project[3] which aims to develop a platform for communication among deaf, hard of hearing and hearing individuals in sign and spoken languages. The paper "Design principles of an Automatic Speech Recognition functionality in a user-centric signed and spoken language translation system" by Aditya Parikh, Louis ten Bosch, Henk van den Heuvel and Cristian Tejedor-García describes design choices of the Automatic Speech Recognition component of the SignON application. Another paper "BeCoS corpus: Belgian Covid-19 Sign language corpus. A corpus for training Sign Language Recognition and Translation" by Vincent Vandeghinste, Bob Van Dyck, Mathieu De Coster, Maud Goddefroy and Joni Dambre presents the Belgian Federal COVID-19 corpus (BeCos) which contains official press conferences from the Belgian Federal Government concerning the COVID-19 pandemic in Dutch, French or German accompanied in almost all cases by live interpreting. A third paper related to the track on sign language, "A Sign Similarity Approach to an Information Retrieval Inspired Visual Dictionary for Sign Language Learners" by Mark Wijkhuizen, Onno Crasborn and Martha Larson introduces a sign similarity approach for the evaluation of visual dictionaries.

Some of the contributions in the journal are closely related to this year's second special track, which focused on language technology for Dutch and Flemish online (forum) discussions. For instance, the paper "Linguistic Analysis of Toxic Language on Social Media" by Ine Gevers, Ilia Markov and Walter Daelemans uses a linguistic analysis in order to discover language patterns that can discriminate between toxic and non-toxic language on social media. In "All that Glitters is not Gold: Transfer-Learning for Offensive Language Detection in Dutch" by Dion Theodoridis and Tommaso Caselli transfer-leaning is proposed as a strategy to boost the creation of language-specific datasets and systems using offensive language in Dutch tweets directed at Dutch politicians as a case study. Another paper, "Responsibility Framing under the Magnifying Lens of NLP: The Case of Gender-Based Violence and Traffic Danger" by Gosse Minnema, Gaetana Ruggiero, Marion Bartl, Sara Gemelli, Tommaso Caselli, Chiara Zanchi, Viviana Patti, Marco te Brömmelstroet and Malvina Nissim focuses on a framework to automatically analyze responsibility framing when reporting femicides (Italian press) and traffic crashes (Dutch/Flemish news reports). This research supports the development (and testing) of tools for NLP practitioners but also automatic large-scale linguistic analysis for journalists or activists.

Two papers in this volume use NLP techniques in the context of cultural heritage. The paper "Automatically Interpreting Dutch Tombstone Inscriptions" by Johan Bos, Cristian A. Marocico, A. Emin Tatar and Yasmin Mzayek proposes a way to assist human annotators and data curators with digital preservation of tombstones by using a pipeline system to automatically read tombstone inscriptions. The paper "OpenBoek: A Corpus of Literary Coreference and Entities with an Exploration of Historical Spelling Normalization" by Andreas van Cranenburgh and Gertjan van Noord presents the OpenBoek corpus – a corpus consisting of classic Dutch novels annotated for coreference. This work shows that while NLP tools struggle with historic ways of spelling Dutch, the effects can be mitigated with simple rule-based methods. Outside of the cultural heritage scope, another paper included in this volume of the journal focuses on coreference resolution in the Dutch language. The paper "Towards Fine(r)-Grained Identification of Event Coreference Resolution Types" by Loic De Langhe, Orphée De Clercq and Veronique Hoste discusses event-event relationships and event coreference resolution and presents the first coreference-based study of event-subevent relations for Dutch.

3. `https://signon-project.eu/`

Another recurring theme is the use of BERT-based language models. For instance, "BERT-Based Transformer Fine-Tuning for Dutch Wikidata" by Niels de Jong and Gosse Bouma proposes a transformer-based Dutch-to-SPARQL QA system that incorporates a multilingual BERT model in the encoder and decoder components of a transformer architecture. In "Noun Phrase and Verb Phrase Ellipsis in Dutch: Identifying Verb-Subject Dependencies with BERTje" by Tessel Haagen, Loïs Dona, Sarah Bosscha, Beatriz Zamith, Richard Koetschruyter and Gijs Wijnholds ellipsis is used to quantify BERTje's linguistic capacity. The paper uses ellipsis to analyze BERTje's ability to capture syntactic information which gives insights into how ellipsis is processed by a computational model.

In terms of NLP technology in bi- or multilingual settings, the paper "Integrating Fuzzy Matches into Sentence-Level Quality Estimation for Neural Machine Translation" by Arda Tezcan investigates how fuzzy matches retrieved from Neural Machine Translation training data can be leveraged for the prediction of sentence-level quality of the translations obtained from said model. While "Improving Domain-specific Cross-lingual Embeddings with Automatically Generated Bilingual Dictionaries" by Pranaydeep Singh, Ayla Rigouts Terryn and Els Lefever proposes a method to create domain-specific dictionaries on-the-fly by leveraging Wikipedia and cross-lingual links.

There are also a number of submissions that investigate neural approaches on Dutch textual data, starting with a comparison between graph and sequential encoders ("Comparing Neural Meaning-to-Text Approaches for Dutch" by Chunliu Wang and Johan Bos) and going further to the application of unsupervised learning to Dutch and English news articles ("Unsupervised text classification with Neural Word Embeddings" by Andriy Kosar, Guy De Pauw and Walter Daelemans). In "Don't do your experiments double-blind: The importance of checking your data" by Nelleke Oostdijk and Hans van Halteren, the authors remind us of the importance of careful inspection of experimental data before running experiments.

Finally, we would like to express our gratitude to all the participants and authors of this year's edition – we hope you enjoyed it as much as we did. Furthermore, we would like to thank our sponsors: Instituut voor de Nederlandse taal, Textkernel, NWO, Tilburg University, SignON, CrossLang, de Nederlandse Organisatie voor Taal- en Spraaktechnologie, Telecats, de Taalunie, Zeta Alpha, Computers and Composition and Textgain for their generous contributions. CLIN32 would not have been possible without this year's Organisation Committee: Raquel G. Alhama, Nadine Braun, Anouck Braggaar, Jan Engelen, Martijn Goudbeek, Emiel Krahmer (Chair), Chris van der Lee, Tom Lentz, Lauraine de Lima, Liesje van der Linden, Emiel van Miltenburg, Dimitar Shterionov, Mirella De Sisto, and Eva Vanmassenhove — all affiliated with the Tilburg School of Humanities and Digital Sciences. This special issue was edited by a subset of the organizing committee. Last but not least, we would like to sincerely thank all the reviewers who provided feedback and insightful comments for the papers published in this special issue. We hope you enjoy reading this year's contributions to the journal and are looking forward to meet you next year at CLIN33 which will take place in Antwerp.