The Future of Academic Lexicography

4 - 8 November 2019 @Oort

Academic or *evidence-based* lexicography has a long tradition of analysing large amounts of language data in a scientific way in order to compile concise, high-quality knowledge about words, and this for the benefit of the entire language community. However, lexicography faces challenges with respect to (a) its role in society, science and the knowledge economy, (b) the scalability of both the analysis and production process, and (c) the customizability and accessibility of its content for a diverse audience and for integration in new IT applications. The workshop explores how each of these challenges can become interesting case studies for research and development in neighbouring disciplines like Data Analytics, Artificial Intelligence, Citizen Science, Human-Computer Interaction and Sociology. As such the workshop aims to strengthen the position of academic lexicography as a locus for multidisciplinary scientific research with a direct relevance for, and impact on, society. Outcome of the Workshop - The direct tangible outcome of the workshop will be a White Paper on the future of academic lexicography, based on presentations and discussions. The white paper is intended as a strategic planning document that (1) spells out the specific issues with respect to scalability, content customization and societal and economic positioning (2) outlines and compares potential solutions and their feasibility, and (3) proposes formats for project-based collaboration between experts from the different scientific disciplines. A first version of the white paper is expected by February 2020 with a final version, incorporating feedback from scientific, societal and economic stakeholders to be published later in the spring. With the white paper as a strategic planning document, the organisers, together with other participants and stakeholders will then explore the formation of one or more project consortia in second half of 2020.

Developments and (beginning) scientific breakthrough - Already during the workshop, the first steps were undertaken to form the nucleus of a consortium between the Dutch Language Institute (INT - main organizer) and other scientific lexicographic institutes (a.o. Zentrum für digitale Lexikographie der deutschen Sprache, Danske Sprog- og Litteraturselskab, Dansk Sprognævn). Based on the recommendations expressed in the workshop discussions, representatives of the different lexicographic institutes have also committed themselves to strengthening the ties to the neighbouring disciplines by organizing workshops and training about Big Data and Artificial Intelligence both at lexicographic conference venues (e.g. ELEX, EuraLex, Globalex), in graduate training programmes (e.g. European Master in Lexicography) as well as within institutes: For example, INT will adjust its longterm policy plan to organise additional training and knowledge exchange between lexicographers and computer linguists and to become and associated member of the European Master in Lexicography. The workshop discussions also clearly indicated need for a permanent body on the European level to structurally support joint, cross-disciplinary R&D in e-lexicography for all European languages (and not only English). Both the organisers and participants have committed themselves to actively explore the options for such a permanent body as a follow-up of the ongoing (but time-limited) EU research infrastructure project ELEXIS. Furthermore, concrete co-operation opportunities within Leiden between the Dutch Language Institute and the Leiden Centre of Data Science on linguist knowledge induction became clear and will be pursued further in the near term. Finally, the specific challenges for Sign Language lexicography were an eye-opener and a concrete starting point for further interdisciplinary co-operation.

Notable new insights - The workshop also delivered a number of new insights into specific issues confronting interdisciplinary R&D in e-lexicography. These mainly relate to how researchers and developers with backgrounds in different disciplines should interface. For example, lexicographic output gives computer scientists very little insight into the knowledge discovery process of lexicographers and a more systematic approach to lexicographer-computer linguist-interactions is needed. The other way around, lexicographers need more training to optimize their understanding and successful use of computational tools. It became clear that the latest state-of-the-art in Computational Linguistics cannot straightforwardly be integrated in easily maintainable lexicographic software tools without compromising Quality of Service because of their complexity. In the same vein, lexicographic knowledge types must be boiled down to simple template-like knowledge structures before they are amenable to automation in AI and Data analytics approaches. Finally, a common interest in a Europe-wide co-operation in historic lexicography on, among other topics, a shared etymology database became so apparent that further co-operation between historic lexicographers form Oxford, Leiden and other universities and institutes will be initiated.

Feedback on Organization/Format - At the suggestion of the Lorentz-centre, the basic workshop included the novel format of "walking brainstorms" after each presentation by a specialist: during the coffee breaks/walking brainstorms, participants were invited to write their questions and proposals for discussion topics on post-its and post them on a blackboard. These topics and questions were then the input for working group discussion later in the day. Because these post-its were often already discussed 1-on-1 during the coffee breaks, they really jump-started the working group discussions. Additionally, (photographs of) the post-its were a tangible record of the idea generation process during the workshop which we rely on as input for compiling the reports and the White Paper afterwards. We really recommend this format for future workshop organizers.
In light of our positive experience with the posting of discussion topics in response to the workshop, we would advise the Lorentz Centre to invest in digital, secure infrastructure for brain storming and discussions that can enrich and optimize the live, face-to-face discussions. It is important that such digital infrastructure would also respect the informal and private (small-group) nature of the Lorentz workshops.