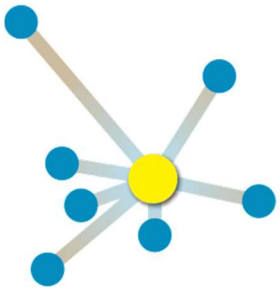
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination

Connecting Europe Facility

Deliverable D3.2.13
Task 3

ELRC Workshop Report for BELGIUM

Author(s): Ayla Rigouts Terryn, Véronique Hoste

Dissemination Level: Public

Version No.: <V1.0>

Date: 2021-11-10

Contents

1 Executive Summary

The third edition of the Belgian ELRC workshop took place online on the 8th of July. It was organised on the same day as the ELG workshop (ELRC in the morning, ELG in the afternoon) and the day before the CLIN (Computational Linguistics in the Netherlands) conference, so that people might be more likely to combine events. Despite a few challenges (our public NAP resigned a few months before the workshop, and organizing the workshop as an online meeting was challenging) and after a lot of work, the event was a success. There were over 60 registrations, and at most times during the workshop, between 40 and 50 people attended. Others, who could not make it or could only attend part of the workshop, expressed their interest in seeing the slides afterwards.

Past the introduction, the workshop opened with a very high-level presentation on NLP and AI and became increasingly specialised and specific, with presentations on neural MT, CEF AT eTranslation, and testimonies from collaborations between industry and the public sector. Despite the restrictions due to the online setting, the final panel session was interactive, with a fruitful discussion between the panel members and the audience.

Compared to a few years back, language resources are now increasingly acknowledged as important, but few concrete changes in the general situation could be identified since the first workshop. Legal issues are still often brought up, with GDPR now being used as an excuse to be overly cautious and not share any data. Often, the power to share data still lies with people who have no direct incentive to risk anything (even if they are aware of the value of language resources, there is no direct benefit for them in sharing their own data), so that, e.g., translators, have few arguments when advocating data sharing to their superiors. Another important remark was that there are three big European infrastructure projects for language resources, all targeted at different audiences (ELRC for the public sector, ELG for industry, and CLARIN for researchers). If you know where to look, there is a lot of data, but technical difficulties remain an issue for smaller users (e.g., huge downloads, complicated structure of the data, etc.).

2 Workshop Agenda

09:30 – 09:40: Welcome and introduction
By Prof. Dr. Véronique Hoste, LT3 Language and Translation Technology Team, Ghent University

09:40 – 10:00: The potential of Language Technology and AI – where we are, where we should be heading
By Prof. Dr. Véronique Hoste, LT3 Language and Translation Technology Team, Ghent University

10:00 – 10:30 Neural machine translation in Belgium
By Prof. Dr. Lieve Macken, LT3 Language and Translation Technology Team, Ghent University

10:30 – 10:45 Coffee Break

10:45 – 11:15 The CEF AT Platform
By François Thunus, computational linguist at the European Commission

11:15 – 11:45 Language technologies by/for the public sector
By Dr. Tom Vanallemeersch, CrossLang

11.45 – 12:15 Language data creation, management and sharing: existing practices and challenges - Panel session
Dr. Ayla Rigouts Terryn (Moderator)
Dr. Vincent Vandeghinste (INT & CLARIN)
Dr. Lore Vandevoorde (Council of the European Union – General Secretariat)
Nathalie De Sutter (Untranslate)

12:15 – 12:30 Conclusions
By Prof. Dr. Véronique Hoste, LT3 Language and Translation Technology Team, Ghent University

3 Summary of Content of Sessions

3.1 Welcome and introduction

Véronique Hoste welcomes the participants and gives a brief introduction of the ELRC and CEF programme. Apart from giving an update on the ELRC progress in Belgium, the focus of the meeting is discussed, i.e., broadening the scope of ELRC from automatic translation to other language technologies. A brief overview of the programme is also given.

3.2 The potential of Language Technology and AI – where we are, where we should be heading

Véronique Hoste delivers a broad presentation on the potential of artificial intelligence technologies then narrowing it down to natural language processing tools. A brief history of the domain of NLP is given, covering a wide range of methodologies from rule-based NLP to machine learning and deep learning. Finally, the main goals of the Flanders AI programme are discussed, also focusing on educating the broad population to the advantages and limitations of AI application and to the ethics involved in building AI applications.

3.3 Neural Machine Translation in Belgium

Lieve Macken gives a very insightful presentation on the methodologies involved in machine translation. She shows the analogy between human and machine translation methodologies in explaining MT core principles such as encoding and decoding. Starting with a short presentation of the rule-based and statistical MT frameworks, she concludes with an in-depth and very intuitive presentation of the neural methodologies which are currently state-of-the-art in machine translation. She shows different examples from the e-Translation platform, showcasing the importance of domain-tailored corpora for translation quality.

3.4 The CEF AT platform

The presentation of François Thunus, computational linguist at the European Commission, focuses on the CEF AT platform and nicely follows the presentation of Lieve Macken in which the main principles of neural machine translation were already discussed. Different examples of the e-translation tool are showcased and the importance of the use of the adequate training materials is also emphasized. Also other tools are introduced, such as anonymisation, named entity recognition, automatic speech recognition. Different links are provided for the audience to check out these tools.

3.5 Language technologies by/for the public sector

Tom Vanallemeersch from CrossLang gives a presentation about the integration of machine translation in the public sector. He briefly mentions two projects: the integration of MT and TM support in the Department of the Prime Minister in Belgium and a large integration project for the Belgian construction federation (MICE). In the latter project, a translation support environment was built based on the open source MateCAT tool used by the translator to validate the output of the machine translation and translation memory system. Several interesting examples are shown, and the presentation ends with a demo of this environment.

3.6 Language data creation, management and sharing: existing practices and challenges (Panel session)

Ayla Rigouts Terryn moderates a panel discussion with Vincent Vandeghinste, Lore Vandevoorde, and Nathalie De Sutter. After a brief introduction by the moderator, the panel members introduce

themselves in more detail and talk about their experiences with language resources. Vincent Vandeghinste comes from the academic world and is affiliated with the Dutch language institute (INT – Instituut voor de Nederlandse Taal) and CLARIN (academic language resource infrastructure project), and, through CLARIN, also the Dutch Language Union (Nederlandse Taalunie) and Leuven University (KU Leuven). Moreover, he is involved with the Dutch chapter of the ELG project as well. Lore Vandevoorde also holds a PhD in translation studies and now works as a translator for the European Council. She used language resources like the Dutch OPUS corpus during her research career and now often relies on the resources available to her through the EU. Finally, Nathalie De Sutter provides a perspective from the industry as the founder of her own company Untranslate. She has regular collaborations with public services as well, for instance through the SNOMED project on medical terminology.

All panel members are asked about the typical challenges they encounter regarding language resources. Interestingly, Vincent Vandeghinste notes that there are three large infrastructure projects: ELRC for public services, ELG for industry, and CLARIN for research. A lot of data is available, but it is not always easy to find. Regarding the quantity and availability of data, Nathalie De Sutter underlines that the data is not always user-friendly. As a user of the DGT translation memory for instance, she points out that it is too large for most translators to download and edit easily. She also recalls the SNOMED project, where users could consult a database of medical terminology, but the way data was formatted prevented most users from getting all (and only) the relevant data from that dataset because of its complicated structure (e.g., synonyms, preferred terms, contexts, etc.). A couple of questions are raised by the audience, for instance on the inclusion of audio-visual translation (e.g., of Flemish Sign Language), considering the European Language Equality project. A final question raised by the audience is on many translators' mind: will translation be completely replaced by post-editing in the near future? Vincent Vandeghinste reassured the audience: people have been saying for years that MT will be near-perfect within 5 years, and we're still not there yet. The response by Lore Vandevoorde provides a nice closing remark for the workshop: MT will become a standard part of the translation process (and is so already sometimes), but only alongside other technologies like translation memories and terminology support. We cannot look at post-editing versus translating "from scratch", but instead look at all these technologies as extra tools in a translator's kit, for a type of "multimodal" translation.

3.7 Take-home message and conclusions

Véronique Hoste closes the workshop. Firstly, she thanks all speakers, the two interpreters and all participants for their contribution. Then, she gives a short recap of the programme and reminds the participants about the various tools mentioned by the different speakers which can also be used freely. She also reminds the participants that they can share their data through the ELRC-Share platform available at https://elrc-share.eu/ and that they can contact the two NAPs if they need any help to share their resources or in case they have any questions with respect to the legal requirements when sharing data.

3.8 Demos session

No demos were shown in a separate session, though the speakers (e.g., Francois Thunus and Tom Vanallemeersch) did provide demos during their presentations.

4 Synthesis of Workshop Discussions

The main issues are not much changed since the last report, but the following is a list of some of the main concerns raised during the workshop:

- There are a lot of **initiatives for language resource infrastructure**, but they **are split** (e.g., ELRC for public services, ELG for industry, and CLARIN for research) so that it is easy to miss relevant data when looking at just one of these repositories.
- The **available data is not always easy to use** unless you have the infrastructure (for very large datasets) or technical knowledge (to extract relevant data from a structured dataset).
- While the participants of the workshop are convinced of the importance of data sharing, they are not always the ones in charge. Often, they are hindered by a strict hierarchy and the people who decide on the data sharing are not necessarily the ones who know the importance, and they don't have a direct incentive to do so. So, **along with the creation of more awareness around the importance of sharing data, it is important to focus on concrete incentives**. For instance, a visible ranking of active data contributors on a platform like ELRC, potentially even with concrete rewards, like better/cheaper access to the tools that are developed based on the data. That way, the language professionals have better arguments when they try to convince the higher-ups to share data.
- **Legal issues** remain an important obstacle, and the people responsible for big datasets tend to err on the side of caution and not release any data, rather than risk e.g., GDPR issues. Being able to refer to ELRC for legal assistance in this regard is important, because GDPR and privacy are too often used as an excuse to not share anything.
- Belgium has always had a dense bureaucratic system, with a complicated government system leading to **many different institutions and little contact between institutions**. This makes it difficult to find and address the right people. This problem was exacerbated by the withdrawal of our public NAP a few months before the workshop. Related to this issue, it is not always easy for people from the Flemish side to have the necessary contacts in the Walloon region.

5 Country Profile: Language data creation, management and sharing

There don't appear to be any significant changes in this sense since the last workshop, at least not in so far as we were able to tell from the discussions and presentations. We have summed up a few issues in the previous section. Of course, half a day online is not much time to get to the bottom of such issues, and this year's edition attracted more vocal responses from LT providers from industry and academia than from the Belgian public sector (people from the public sector were present but contributed a bit less this time). This may of course also be due to our lack of a public NAP.

The country survey was only filled out by four people, so not very much could be derived from that. Three of the respondents said eTranslation is used in their organisation, one even uses it exclusively (no other LTs). Other LTs that are used include Wordbee, DeepL, IATE, EuroVoc, and EurLex. Three of the respondents also work for organisations where language resources are available and this mostly concerns translations (in the three national languages + English). Three out of four respondents do have data management plans in their organisation, but not much additional information is provided. All four respondents cite legal issues as the main obstacle towards sharing more resources, once along with "other" (specified as: "available time and resources to properly manage the data"). One respondent mentions medical and personal data in most of the resources, which makes them difficult to share.