European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.

Deliverable D3.2.5
Task 8

ELRC Workshop Report for Belgium

Author(s):	Stijn De Smeytere, Véronique Hoste, Ayla Rigouts Terryn
Dissemination Level:	Public
Version No.:	1.0
Date:	2018-08-30

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Belgium, which took place in Brussels on the 29th of May 2018 at the Résidence Palace. It includes the agenda of the event (section 2) and briefly provides details on the content of each individual, interactive and panel workshop session (sections 3 & 4). The event was attended by 40 participants spanning a wide range of government departments and public organisations.

The dedicated event webpage can be found at http://www.lr-coordination.eu/l2belgium

2 Workshop Agenda

08:00 – 09:00	**Registration**
09:00 – 09:10	**Welcome and introduction** *Stijn De Smeytere, FPS Chancellery of the Prime Minister*
09:10 – 09:15	**Welcome by the EC** *Jeroen Aspeslagh, DG Translation, European Commission*
	Session 1. Connecting a multilingual Europe: European context and local needs
09:15 – 09:35	**Connecting public services across Europe: ambition and results so far** *Aleksandra Wesolowska, DG CONNECT (Video presentation)*
09:35 – 09:55	**National initiatives for digital public services and (open) data** *Luc Van Tilborgh, Program Manager eGovernment, Digital Transformation Office*
09:55 – 10:40	**CEF in Belgium: an outlook into current and future challenges – Panel session** Moderator: *Stijn De Smeytere, FPS Chancellery of the Prime Minister* Panelists: ● *Roel Arys, Federal e-Procurement service* ● *Wim Bonneux, Business Architect, Federal Public Service Finance* ● *Giselda Curvers, Programme Manager mypension.be* ● *Bart Hanssens, DG Digital Transformation*
10:40 – 11:00	**The CEF eTranslation platform @ work** *Jeroen Aspeslagh, DG Translation, European Commission*
11.00 – 11:30	*Coffee Break*
	Session 2. Engage: hands-on data
11:30 – 11:50	**The European Language Resource Coordination (ELRC) action** *Khalid Choukri, ELDA, ELRC*
11:50 – 12:05	**ELRC in Belgium** *Prof. Dr. Veronique Hoste, Ghent University*
12:05 – 13:05	*Lunch Break*
13:05 – 13:35	**Preparing and sharing data with the ELRC repository and New services provided by the ELRC consortium** *Khalid Choukri, ELDA*
13:35 – 14:05	**Can language data be shared and how?** *Joris Deene, Legal Expert, Everest Law*
14:05 – 15:00	**Identifying and managing your data: Questions & Answers** **+ Discussion and Conclusion** *Khalid Choukri, ELDA, ELRC* *Stijn De Smeytere, FPS Chancellery of the Prime Minister* *Prof. Dr. Veronique Hoste, Ghent University* *Ayla Rigouts Terryn, LT3, Ghent University*
15:00 – 16:00	*Reception and networking*

3 Summary of Content of Sessions

3.1 Welcome and introduction

Stijn De Smeytere opened the second Belgian ELRC workshop by thanking everyone for their attendance and informing the audience that the entire workshop would be interpreted into Dutch and French (depending on the speaker) by student interpreters from Ghent University. It had been decided to keep the slides in English, so they would be clear to the entire audience.

Stijn De Smeytere started his introduction by referring to a quote by Rocío Txabarriaga: “To stay competitive in a global market, European businesses must communicate multilingually. Likewise, governments are more effective when they can receive and transmit information to their linguistically diverse populations.” He stressed how this is even more true today, than at the time of the quote (2009). The need for instant translations is growing and we are looking into anything that can help the translation process and our lives in general easier. That, Stijn reminds us, is why we are here today.

He proceeds by explaining the goal of ELRC: collecting language resources, and explains what we mean by that and why they are so important. The connection was made to IBM’s Deep Blue and the very current question: “Will computers replace human translators?”. He answered this questioning by saying that, although the translator’s profession will undoubtedly change over the next few years, we should look at computers as things that will make our jobs easier, rather than replace us.

Stijn De Smeytere concluded by going over the workshop objectives and the agenda. He kindly reminded everyone to fill out the feedback and engagement forms at the end of the workshop, before giving the floor to Jeroen Aspeslagh.

3.2 Welcome by the EC

Jeroen Aspeslagh, the European Commission representative in Belgium, welcomed everyone on behalf of the EC and the DG Translation in particular. He talked about his personal experience with machine translation (MT) and how he followed the evolution from rule-based systems, to statistical MT, to current neural MT. He explained how the EU’s MT engine followed this evolution and that the eTranslation platform, freely available to all public administrations, is currently making the transition to neural MT. This has been a substantial improvement, especially for languages such as Finnish and German and soon also Dutch.

He concluded by stressing the importance of the ELRC project and our language resources to continue this positive evolution.

3.3 Connecting public services across Europe: ambition and results so far

A video presentation from Aleksandra Wesolowska (DG CONNECT) was played to the audience with live interpretation into Dutch and French.

3.4 National initiatives for digital public services and (open) data

After the video, Luc Van Tilborgh, who works on Federal Public Service Policy and Support at the Digital Transformation Office, gave a riveting presentation about Belgian National Initiatives regarding digital public services, open data and language resources. He started by addressing open data in the Belgian context, beginning with the implementation of the PSI directive. Belgium adopted a federal open data strategy with an ambitious roadmap for 2015-2020, where the norm is “open by default” for data in the public services. The only exceptions to this rule are documents that contain private information, or information that may harm public security.

Luc Van Tilborgh also introduced the Open Data Portal at https://data.gov.be, where over 7000 datasets are already available. He explained the multitudinous benefits of an open data policy, ranging from improvement of data quality to a better transparency towards citizens.

Next, he proudly stated how Belgium ranked 6th in the 2017 DESI rankings and how they are aiming for an even better ranking this year. Public services are, understandably, Belgium's weakness, because of the complicated structure of the country. He elaborated further on the specific eGovernment goals and our participation in several EU pilot projects.

He finished by stressing the importance of multilingualism in a trilingual country like Belgium and provided some useful links to the digital agenda and Babelfed.

3.5 CEF in Belgium: an outlook into current and future challenges – Panel session

Stijn De Smeytere led an interesting panel with Roel Arys (eProcurement), Giselda Curvers (ePension), Frank Baelus (eFinance) and Bart Hanssens (Open Data Portal). After a short round of introductions, Stijn De Smeytere asked the panellists about the importance of public digital services in their respective domains. Roel Arys answered that it was mainly a question of visibility for eProcurement: easily addressing a large audience for public tenders. Additionally, he named minimizing mistakes and maximizing transparency as well. Giselda Curvers gave a similar answer and added that digital services regarding ePension are especially useful for addressing a large audience and giving personalized information to anyone who wants it. Frank Baelus mentioned several eGov applications, most importantly Tax on Web, the largest eGov application in Belgium. In his department, they now focus on "straight through processing": personalized services on demand. Finally, Bart Hanssens gave an answer in three parts. First, sharing data improves transparency. Second, there is a treasure of information available within public services, ranging from weather patterns to house prices, which are useful outside of the government as well. Third and final: it is good publicity for the government to show the available expertise.

The next question regarded the CEF building blocks. Arys admitted they were rarely used, but that they were looking with great interest at the development of eSignature as an alternative for the small percentage of cases where an eID wasn't available. Their own eProcurement is an adapted version of the CEF eProcurement. Giselda Curvers also expressed an interest in alternatives for cases where an eID isn't an option, but admitted that, currently, the building blocks aren't used. Frank Baelus added further development of eInvoicing to the list of desired improvements. Bart Hanssens said that, per definition, identification and payment isn't required with open data, and, to date, there hadn't even been any requests for translations. However, he would be very interested in more sophisticated technology such as automatic classification in predefined categories and automatic summarization.

When asked about cross-border interaction, all parties agreed that, in a country as small and multicultural as Belgium, cross-border interaction is a necessity. ePension is already a member of a European consortium, looking to make an international version of ePension.

Next, Stijn De Smeytere inquired about multilinguality. Of course, in a trilingual country as Belgium, all services deal with multilinguality and sometimes the language laws are very specific regarding the language of certain documents. None of the panellists use MT for this purpose yet though and few of them currently feel the need to introduce MT. English translations are mentioned by some of the panellists, especially regarding increased visibility outside the country borders and in multilingual contexts, such as border-control.

The final question addressed the needs for improvement regarding this multilinguality. Roel Arys, Bart Hanssen and Frank Baelus state that they never receive translation requests, so they don't feel any pressure to work on this, even though they are aware of potential benefits. Giselda Curvers does see an opportunity, also for MT, especially for German, which is a small language in Belgium, but still they sometimes get large requests, so it can be difficult to handle with a small team.

Overall, the panellists all seemed very interested in any developments and aware of the services that are already available, even if they don't always feel the need to use them.

3.6 The CEF eTranslation platform @ work

Jeroen Aspeslagh, from the DG Translation for Dutch at the European Commission, gave an interesting presentation, with many informative examples about eTranslation. He started out by stressing the importance of domain adaptation, showing that an MT engine based on an EU legal corpus yields very good results on this legalese, but that it fails in other areas. He introduced eTranslation and clarified its relation to MT@EC and CEF.AT. He showed the availability of the tool and the recent improvements, such as support for more data formats and batch processing. He also spent time explaining how, in contrast to private MT engines, you can choose to have your data deleted within 24 hours, so it is safe to use for confidential documents.

With very clear examples, he showed the last improvements with neural MT versus the statistical engines and proudly announced that, by the end of the year, neural MT will have been installed for all official EU languages and that there are definite plans to add non-EU languages as well. He made another interesting point by stating that neural MT doesn't need as much data as statistical MT (though domain-specific data is still an issue), but that it is more important than ever to have high-quality data.

3.7 The European Language Resource Coordination (ELRC) action

Khalid Choukri, from ELDA, introduced the session by describing ELRC as a coordination founded in 2015, and headed by 4 organisations: Tilde, ELDA, DFKI and ILSP.

It is also supported by 60 National Anchor Points (NAPs): in Belgium's case the technical NAP is Prof. Dr. Véronique Hoste, Head of the Department of Translation, Interpreting and Communication at Ghent University and member of the LT3 language and Translation Technology Team. She is supported by a PhD student from the LT3 group: Ayla Rigouts Terryn. The public administration NAP is Stijn De Smeytere, who works at the Chancellery of the Prime Minister.

"What does the ELRC do?" Khalid Choukri explained that the aim of the ELRC is to try to set up a pipeline between EC services across EU member states, as well as Norway and Iceland. To achieve this, the ELRC collects datasets suitable for developing MT systems: parallel corpora, terminology databases - any digital text expressed in words by human experts.

He also described the need to identify the various requirements across Member states, saying that it is a critical issue, and that it is necessary to engage with each Member State to locate and collect existing language resources in a suitable manner. When the ELRC was first set up, they came across some issues, mainly technical and legal which are addressed through a helpdesk set up to deal with all related queries.

To the next question “Why ELRC?”, Khalid Choukri answered: to facilitate cross-border interaction. We can’t ask translators to do absolutely everything, there is simply too much to be done, especially in a trilingual country like Belgium. Translators need support.

And how to make it (MT) work? In-domain text that has been translated by experts is the key and Belgium ought to have an advantage here, with the “open by default” strategy of the public services. Khalid Choukri concluded his presentation by repeating that help is available for any data holder who needs it, and can be accessed via the online helpdesk.

3.8 ELRC in Belgium

This talk was given by Prof. Dr. Véronique Hoste, Head of the Department of Translation, Interpreting and Communication at Ghent University and member of the LT3 language and Translation Technology Team. She started by looking back to the first ELRC workshop, now 2 years ago, in Ghent. She gave an overview of all data collected since then, both for Belgium specifically and for our three national languages in general. While much has already been done, there is still definite room for improvement, especially for domain-specific data, where you will see an immediate return-on-investment in the improved MT quality.

She then dedicated time to explaining the lessons learnt from the previous two years. First, we saw how difficult it can be to find the right people to give permission to share the data. Often you have to go high-up in the administrative structure to find the right person. Next, there is the issue of anonymization. This can be more or less easily fixed with structured data, but is almost impossible for unstructured data. Then, there is the importance of obtaining the correct rights on your translations, especially when they are outsourced. Some of the technical issues can be easily fixed, such as adding metadata to translation memories to separate the sensitive data from the data that can be freely shared. Finally, Véronique Hoste added that she felt the lack of awareness and scepticism towards technology issues were already largely solved.

She concluded her presentation by mentioning how challenging it can be to identify the key players and asked the audience for support in this regard, by filling out the engagement forms.

3.9 Can language data be shared and how?

Joris Deene, Legal Expert at Everest Law gave a very clear presentation about the legal considerations concerning data sharing. He provided accurate and specific information, while remaining understandable to a laymen audience. He started by explaining copyright law in Belgium, with specific attention paid to how it applies to translations. Originality and creativity are key legal concepts in copyright law even though their assessment can encompass a certain measure of subjectivity. When asked about specific cases from the floor, he even illustrated this by stating that different courts in Flanders have different opinions, so that the same issue might be judged differently in Antwerp and in Ghent. He also mentioned the importance of obtaining the rights to translations, since, by default, in Belgium, the rights go to the “maker”, ergo the translator; not the client.

His next few slides were dedicated to the Public Sector Information (PSI) directive, which he summarized and explained clearly. In Belgium, this has been implemented separately in the federal and the regional governments and it has been changed in 2015.

Joris Deene considerately answered some questions from the floor regarding the definition of “originality” (as explained above). Some specific cases were mentioned such as ownership of MT output or terminological records. The final question regarded citations, which he assured us, are still allowed, provided you don't cite too much and properly mention the source.

3.10 Preparing and sharing data with the ELRC repository – and what happens next

The last presentation of the day was, again, by Khalid Choukri who illustrated the practical side of sharing data. He showed us the website and a detailed example of one of the shared language resources. Once more, he stressed the need for language- and domain-specific data. While the EU already has a lot of data, this does not suffice; not if the MT is supposed to work for national public services as well.

He gave an overview of all types of data that are of interest: from monolingual corpora, to parallel corpora, to term bases. He took care to mention the preferred data formats. He also reminded everyone that ELRC provides on-site assistance to those who require it, without extra charge.

3.11 Identifying and managing your data: Questions & Answers + conclusion and discussion

Khalid Choukri, Stijn De Smeytere and Prof. Dr. Véronique Hoste answered questions from the audience:

Q: "Due to English as a lingua franca, many domains are almost disappearing in other languages, so collecting data in these domains is not possible anymore."

Véronique Hoste: "Yes, this is an important issue in language policies. For instance, at universities, there is an increasing demand for classes taught in English. While I like teaching in English, this does present a potential problem, with a loss of non-English terminology in certain domains. I teach language and translation technology myself and often catch myself using the English terminology simply because I do not know a good Dutch alternative. Luckily, this is somewhat countered by the strict language policies here in Belgium, where we public administrations are very strictly bi- or even trilingual."

Khalid Choukri: "We should also be optimistic, that maybe technology can be the tool to help us keep language-specific terminology, to avoid this code-switching."

Q: "How do you deal with language variants, such as Dutch in Belgium vs. the Netherlands and French in Belgium vs. France and German in Belgium vs. Germany? Some things can be correct in one variant and not the other."

Khalid Choukri: "I don't think there really is a solution for this problem. It is the richness of language and part of the charm of language."

Véronique Hoste: "That is just one more example of why it is important to share your own data: so the MT will know your variants as well!"

Q: "Will domain-specific engines be made? Or just the one engine for all domains?"

Jeroen Aspeslagh: "There are definite plans for domain-specific engines as well."

Q: "In the Netherlands, we had cases where people were willing to share data, until they went to talk with their lawyers. After that, problems arose with licenses and legal issues…"

Khalid Choukri: "Such problems did indeed arise after the PSI directive was implemented. Exclusive contracts aren't possible. Belgium and the Netherlands should, however, profit from their "open by default” strategies. This entire system ought to be a win-win situation: any effort needed to donate the data is compensated by the freely available MT that uses this data.”

Q: “There seems to be very little enthusiasm for different countries that share a language to cooperate. This would be very useful, since all of our national languages are spoken in other EU countries as well, sometimes several. What can we do to stimulate this?”

Khalid Choukri: “The EC funds 75% of such cooperative projects between EU members.”

Q: “For now, all of these projects are still messy and vague. When you ask a terminologist how many terms there are, you will never get a straight answer. Don’t we need to know what we are up against, before we go any further? Have a definite goal?

Véronique Hoste: “Concerning the number of terms, I do not think it is necessary to be able to count them. Homo universalis doesn’t exist anymore: we are all very specialised and there are so many fast evolving domains, that this number would change constantly. What we do need, are flexible tools that can handle these trends; tools that can find terminology automatically, such as the ones Ayla Rigouts Terryn, is working on for her PhD. There isn’t even consensus about the definition of a term, so it really is impossible to start counting. So really, the focus should be on technology that can do this for us, that can help us get an idea of such language characteristics.”

4 Synthesis of Workshop Discussions

In general, the audience of the workshop seemed impressed by the ELRC project in general and the workshop specifically. They were pleased to see the improved quality of neural versus statistical MT and appeared generally convinced of the importance of domain-specific data. One of the primary concerns voiced throughout the day, was how to find and convince the relevant people through the complicated hierarchies of Belgian public services. The individual public services seemed up-to-date with the different CEF digital building blocks, though little use was made of them so far. They seemed content to watch any evolutions in case anything relevant to their own situation was developed, or to take one of the building blocks as a starting point to build their own tools. Finally, there appeared to be some interest for the collaboration projects where the EU could fund up to 75% of the costs. For information, the last call for such project proposals under H2020 has just been evaluated. There will be no further calls until the new programme comes into force in 2021.

4.1 ELRC and Open language Data in Belgium

- https://data.gov.be/ is the Belgian open data portal, which contains linguistic data as well
- The Belgian policy for public service data sharing is “open by default”
- Copyright is determined by “originality” and “creativity” and is assigned to the “maker” of the work (in the case of translations, the translator holds the copyright). To change this, you need to contractually determine that the translator offers the client the copyright over his/her translation.

See summaries above and presentations for more information.

4.2 Success stories and lessons learnt

- The Belgian open data portal is already a good source of information
- It can be difficult to identify the relevant people, especially since Belgium is such a complicated country, with many different (layers of) administration
- The last version of eTranslation is a definite improvement over MT@EC
- After the last workshop, 1 translation memory, 2 parallel corpora and a term base were donated: all multilingual resources.

4.3 Workshop Presentation Material

The presentations are published on the Belgian workshop agenda webpage (http://lr-coordination.eu/l2belgium_agenda)