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

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.12
Task 8

ELRC Workshop Report for the Netherlands

Author(s): Jan Odijk & Carole Tiberius

Dissemination Level: Confidential

Version No.: V1.0

Date: 2018-11-27

Contents

1 Executive Summary

This document reports on the second ELRC Workshop in the Netherlands, which took place in The Hague on the 5th of October 2018 at the Huis van Europa. 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 25 participants.

The dedicated event webpage can be found at http://lr-coordination.eu/l2netherlands .

2 Workshop Agenda

08:30 - 09:00 **Registration**

09:00 - 09:10 **Welcome and introduction**
Jan Odijk (UU, National anchorpoint)
Steven Krauwer (First executive director of CLARIN ERIC, chairman of the day)

09:10 - 09:15 **Welcome by the EC**
Hugo Keizer (European Commission, representative for The Netherlands)

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**
Alexandra Wesolowska (Project Officer, Directorate-General for Communications Networks, Content and Technology, European Commission) - video presentation

09:35 - 09:55 **National initiatives for digital public services and (open) data**
Paul Suijkerbuijk (Open data expert at the learning and expertise centre Open Government)

09:55 -10:40 **The CEF eTranslation platform @ work**
Hugo Keizer

10:40 – 11:30 **CEF in The Netherlands: an outlook into current and future challenges – Panel session**
Moderator: Jan Odijk
Panelists:

- **Piet van den Berg**, Project manager international at the **RINIS Foundation**
- **Kornelis Drijfhout**, Head of Unit eInvoicing for the Netherlands Ministry of Economic Affairs and Climate Policy, Information manager for TenderNed at **PIANOo**
- **Xander van der Linde**, Coordinating Advisor at ICTU a.o. involved in national implementation strategy eDelivery, program Implementation European Health-services and Single Digital Gateway Resolution

11.30 – 12:00 *Coffee Break*

Session 2. Engage: hands-on data

12:00 – 12:25 **The European Language Resource Coordination (ELRC) action**
Khalid Choukri (General Secretary ELRA, CEO ELDA)

12:25 – 12:40 **ELRC in The Netherlands**
Jan Odijk /Carole Tiberius (INT)

12:40 – 13:40 *Lunch Break*

13:40 – 14:05 **Can language data be shared and how? National and European legal framework**
Annemarie Beunen (Copyright lawyer, National Library of the Netherlands)

14:05 – 14:40 **Preparing and sharing data with the ELRC repository – and what happens next**
Khalid Choukri (General Secretary ELRA, CEO ELDA)

14:40 – 15:15 **Identifying and managing your data: Questions & Answers**
Jan Odijk / Steven Krauwer

15:15 – 15:30 **Conclusions**
Jan Odijk / Steven Krauwer / Carole Tiberius

15:30 – 16:30 *Coffee Break and Networking*

3 Summary of Content of Sessions

3.1 Welcome and introduction

Jan Odijk opened the second ELRC workshop in the Netherlands by thanking everyone for their attendance and by introducing Steven Krauwer as chairman of the day.

He briefly sketched the context of this workshop and its predecessor workshop held in 2016: EU countries want to offer an increasing number of services digitally across country borders, hence also often across language barriers. Machine translation can help alleviate the language barriers that currently impede such cross-border digital public services. For this reason, the main goals of this workshop are: showing how a public service can benefit from CEF eTranslation and how CEF eTranslation can benefit from a public service, viz. if it makes available textual data that can be used to improve machine translation for the specific domain and vocabularies of the public service. A first characterization of the types of textual data that are needed was given. Jan Odijk emphasized that it is important to be aware of the fact that running natural language text is also valuable data. Unfortunately, that awareness is often not present yet.

3.2 Welcome by the EC

Hugo Keizer, the European Commission representative in the Netherlands, welcomed everyone on behalf of the EC and the DG Translation in particular. He provided some background information on the work of the representatives of the European Commission in the Netherlands and briefly introduced the CEF (Connecting Europe Facility) programme in the Netherlands and gave a few examples of CEF-projects in the Netherlands, e.g. the PortLiner project which aims at the uptake of zero emission shipping based on electric propulsion, targeting inland waterway vessels.

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. She presented the Connecting Europe Facility, with special emphasis on the Digital Service Infrastructures (DSIs) and particularly the CEF Automated Translation building block. Aleksandra concluded with the need for the involvement and connection of the national public administrations with eTranslation and the current funding opportunities.

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

After the video, Paul Suijkerbuijk, who is an Open Data expert at the Learning and Expertise centre Open Government, talked about open data and multilinguality. He illustrated this with several examples, e.g. OVRadar, i.e. public transportation information, information on quality of schools, information on city councils, data on waterway locations, openstreetmap, and Europeana.

Although he observed that multilinguality may be a topic, no systematic attention is given to it in the Netherlands. He did show a few open data sets with multilingual content, e.g. openstreetmap, but many of these are restricted to terms or names in metadata that consist of a single word or a few words, but do not involve running natural language text. Jan Odijk suggested that for such examples multilingual vocabularies and term banks are more appropriate than full-fledged MT.

Paul Suijkerbuijk emphasised the importance of metadata and noted a potential cross-jurisdictional issue as the freedom of information law differs per country, and even something as simple as a license

plate is connected to a car in the Netherlands but to a person in Belgium, and is for this reason subject to different legal and ethical restrictions

3.5 The CEF eTranslation platform @ work

Hugo Keizer 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.

In the discussion after Hugo's presentation the issue came up whether the system automatically adapts itself after manual corrections of automatically generated translations by the user. The answer to this question was negative. The influence of a single correction would be negligible anyway, and the system is generic and not specific to a single user or group of users.

3.6 CEF in the Netherlands: an outlook into current and future challenges – Panel session

The panel session hosted representatives of RINIS, PIANOo and ICTU.

- **Piet van den Berg**, Project manager international at the RINIS Foundation. RINIS is a hub for fully-automated electronic data exchange in the public domain. RINIS is an acronym of Routerings Instituut (inter)Nationale InformatieStromen (the Routing Institute for National and International Information Streams). Piet showed a short animation to introduce RINIS. He characterized RINIS as a kind of digital mailman.

- **Kornelis Drijfhout**, Head of Unit eInvoicing for the Netherlands Ministry of Economic Affairs and Climate Policy, Information manager for TenderNed at PIANOo.
  He introduced PIANOo through a short presentation**.** Through TenderNed administrations can digitally announce and publish calls for tenders. Companies can make proposals and submit them digitally through TenderNed. TenderNed is fully connected to the European Tenders Electronic Daily (TED), so that all information is available across all of Europe. A call for tenders is typically a highly detailed legal and technical document, a summary of which is posted on TenderNed. European tendering rules apply. The administration can post the summary in any language they prefer, usually it is in Dutch, but English also occurs. 60% of an announcement consists of structured metadata and 40% is running natural language text. Calls for tender are publicly available but the submitted proposals are confidential and cannot be used for improving MT.

- **Xander van der Linde**, Coordinating Advisor at ICTU a.o. involved in the national implementation strategy eDelivery, program Implementation European Health-services and Single Digital Gateway Resolution. He sketched a number of on-going activities.
  One is the Programme for Implementation of International Care Services, i.e. the implementation of a National Contact Point eHealth in the European Network for exchange of patient summaries between healthcare professionals. The patient summaries are available in the original language and are translated to English and Dutch. Five hospitals are the first group where the Emergency departments (SHE) will be connected to the NCPeH.

A second one concerns the *Single Digital Gateway Resolution Regulation*, which is applicable in all Member States, according to the "once only" principle (ensures that citizens and businesses are asked to submit information only once to a public administration). It affects all government layers in the Netherlands. All EU citizens and businesses are given access to information and procedures from other Member States. Member States shall make the information accessible in an official language of the Union that is broadly understood by the largest possible number of cross-border users. That is, "without discrimination" between national residents and other EU residents. The envisaged implementation period is expected to be between 2 and 5 years. Machine translation appears to be of special importance for this.

3.7 The European Language Resource Coordination (ELRC) action

Khalid Choukri, ELDA CEO and representative of the ELRC consortium, presented the consortium and its goals, its activities and the current situation as regards to data collection at the European level, the repository developed, and the services offered by the helpdesk to data contributors and users. ELRC is a coordination body founded in 2015, and headed by 4 organisations: Tilde, ELDA, DFKI and ILSP. It is also supported by 60 National Anchor Points (NAPs): For the Netherlands, the technical NAP is Prof. Dr. Jan Odijk, and Dr. Carole Tiberius is substitute technical NAP. For the workshop organisation, they are supported by a student from Utrecht University: Irene Kramer. There is no public administration NAP yet.

"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, translation memories, 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. Translators need support. And how to make it (MT) work? In-domain text that has been translated by experts is the key. Khalid Choukri concluded his presentation by repeating that help is available for any data holder who needs it, which can be accessed via the online helpdesk[1].

3.8 ELRC in the Netherlands

Carole Tiberius, substitute technical NAP, presented the ELRC project and its achievements at the national level since 2016. She gave a brief history of the project and recapitulated the main goals of the ELRC initiative. She then presented an overview of the Dutch data identified and collected in the framework of ELRC, and the problems and issues faced by ELRC representatives during the collection process. Among them the lack of awareness of the importance of textual data, the difficulty in finding the right people to give permission to share the data and the fragmentation of the translation processes at the public administrations in the Netherlands. Next, there are legal and ethical issues,

[1] http://lr-coordination.eu/helpdesk

and finally there is the importance of obtaining the correct rights on your translations, especially when they are outsourced.
Carole Tiberius then dedicated some time to explaining the importance of a so-called country profile. A country profile should provide an insight in the language resource infrastructures at public administrations in the Netherlands. It should provide information on the translation workflow (e.g. what documents are translated, which language pairs, in-house or outsourced, (CAT-)tools used) and what data is included in any of the (open) data portals that exist in the Netherlands (e.g. Dataportaal van de Nederlandse Overheid, Open Data Portaal van de Tweede Kamer, Open Data Nederland). Within ELRC, country profiles are defined for each participating country.

She concluded her presentation by saying that 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.

3.9 Can language data be shared and how?

Annemarie Beunen, copyright lawyer at the National Library, introduced and discussed the applicable laws in sharing data for the Netherlands:

- Auteurswet (Dutch Copyright Act)
- Databankenwet (Dutch Databases Act)
- Wet hergebruik overheidsinformatie (WHO, Act on re-use of Public Sector Information)

All three laws are based on EU regulations that are currently being revised in Brussels.

She discussed the three laws, with special attention to the regulations that apply to public information and to translations.

The EU is currently revising the guideline for *Copyright in the Digital Single Market* (DSM guideline), in particular it is considering a *Text & Data Mining* (TDM) exception for research organisations. This is probably less relevant for use of data by CEF eTranslation but may be important for use of the same data by organisations such as universities and research institutes.

A very important directive is the *Directive on the re-use of Public Sector Information* (PSI Directive), which holds since 2003, was sharpened in 2013 and is currently also being revised. Its main purpose is to stimulate the use of public sector information by other parties. It has been implemented in the Netherlands through the *Wet Hergebruik Overheidsinformatie*[2] (WHO). Organisations can make a WHO request to a public sector organisation, which has to decide within 4 weeks whether these data fall under the WHO, and if so, it is obliged to supply the data.

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

In this presentation, Khalid Choukri 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

[2] https://www.open-overheid.nl/open-overheid/handleiding-wet-hergebruik-van-overheidsinformatie-een-nieuwe-versie/ (in Dutch)

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

There were many questions and answers, and many suggestions for potentially relevant data:

- Alice Dijkstra pointed out the texts produced by the national research funding organisation NWO: it produces most of its publications (web site pages, brochures, calls for proposals, etc.) both in Dutch and in English
- Henk van den Heuvel inquired whether texts from universities would be relevant. The answer was positive, so we will contact Henk for Radboud University web site texts and brochures, and we will contact people at other universities as well.
- People from the UWV[3] were present (Tom Koppe, Rob van Luinen) and they have big translation needs and have a lot of textual data that they very likely can share. It involves sharing information with foreign public services and communication via social media, which sometimes happens across country and language borders. Their website is available in Dutch and in English. They indicated their interest explicitly.
- A representative of the Tax Service was present (Oele Koornwinder), and we agreed to investigate with him about the needs of the tax services and the availability of relevant data there.
- Neil Gouw of Autoriteit Consument & Markt[4] (ACM) was attending and an arrangement for a follow-up meeting was made immediately.
- Several people pointed out the Rijksvoorlichtingsdienst[5] and the Dienst Publiek en Informatie[6] (both residing under the Ministry of General Affairs) as potential data providers. For example, the sites https://www.rijksoverheid.nl/ and https://www.government.nl/ appear to form a parallel corpus that describes the government structure of the Netherlands.
- Translators of the Ministry of Defense were attending, and they are surely also worth further investigation (Kees Bakhuijzen, Peter Janssen). They indicated their interest explicitly.
- Maybe we can intensify our data requests with KOOP through Bieke van der Korst.
- Paul Suijkerbuijk even made a suggestion for a candidate public NAP. We will surely investigate this suggestion further.
- Xander van der Linde pointed out that ELRC and CEF eTranslation are insufficiently known within the ministries and other public organisations. It should be made much better known and it should be made clear where and with whom people can inquire for more information, supply data, discuss opportunities etc.
- Marcel Hopman (Ministry of the Interior and Kingdom Relations) pointed out that one should carefully analyse where there are problems for which MT can offer a solution or be part of a solution, and where it cannot.
- Though no one of SVB[7] was attending the workshop, there was contact via e-mail and a new attempt will be made to get their data available.

[3] Employee Insurance Agency
[4] Authority for Consumers & Markets
[5] Government Information Service
[6] Public Information and Communications Service
[7] Sociale VerzekeringsBank, Social Security Bank

For texts from websites ideally the underlying source data is used, and such data are especially valuable if the source and target texts are already aligned.

3.12 Conclusions

Jan Odijk concluded the workshop. He stated that he was not very optimistic at the beginning of the workshop about its success. The number of participants is small, and though many more registered many canceled their registration shortly before the workshop. It was also not clear that we had succeeded in reaching all important players. But during the workshop it turned out that many important players were present, and many suggestions for new potential data providers were made, both explicitly in the meeting but even more informally during lunch or via e-mail.

The workshop also succeeded in raising awareness for the importance and the enormous value of running natural text as data, as an essential ingredient for the production of high quality MT systems, including CEF eTranslation, which will increase its quality and its suitability for public service documents when it is fed with such documents and their translations in the training phase.

4 Synthesis of Workshop Discussions

We succeeded in raising awareness of the importance and value of running natural language text as data in the workshop. We also succeeded in raising enthusiasm for investigating options to make use of CEF eTranslation and to contribute to it by providing data, and we identified over 8 new potential sources of data. Many participants suggested their own organisation, but we were also pointed to an important new potential source by several people in the audience, viz. the *Rijksvoorlichtingsdienst* and the *Dienst Publiek en Communicatie*, for which no representatives were present.

Though the absolute number of attendees was lower than in 2016, the number of highly relevant participants was much higher (many participants from ministries and implementing bodies).

We repeat some other points that were mentioned earlier:

- Xander van der Linde pointed out that ELRC and CEF eTranslation are insufficiently known within the ministries and other public organisations. It should be made much better known and it should be made clear where and with who people can inquire for more information, supply data, discuss opportunities etc.
- Marcel Hopman (Ministry of the Interior and Kingdom Relations) pointed out that one should carefully analyse where there are problems for which MT can offer a solution or be part of a solution, and where it cannot.

4.1 ELRC and Open Language Data in the Netherlands

The data portal of the Dutch government (https://data.overheid.nl/) mainly contains non-textual data. Textual data account for less than 2%. By raising awareness of the importance of textual data, we hope this number will increase. The section on laws (https://wetten.overheid.nl) contains a lot of text, and it is not clear whether these are already in use by CEF eTranslation. If not, these surely can be added.

4.2 Success stories and lessons learnt

- The Dutch Open Data portal is already a good source of information, although the metadata should be improved to identify textual data.
- The level of awareness of the importance of textual data has been raised. But so far this awareness is present with too few people. We should really announce our existence, the opportunities that CEF eTranslation offers, and persons and systems that can be contacted for donating data inside the ministries and other public service organisations, e.g. by advertising there in their internal newsletters, being visible on their internal communications channels (intranet, expertise books, etc.)
- After the last workshop, ACM (Authority of Consumers and Markets) has expressed interest in donating multilingual data.
- It would be good to include more information on “what’s in it for the data provider” in the workshop brochure.
- As there are not a lot of services yet where eTranslation is used, it would be good to have a concrete (mock-up) example of a workflow with eTranslation that could be presented during the workshop.

4.3 Workshop Presentations

The presentations have been made available online on the ELRC website: http://www.lr-coordination.eu/l2netherlands_agenda.