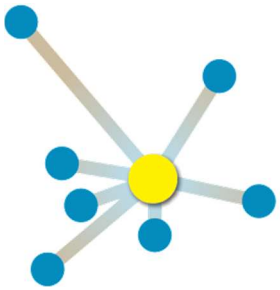
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.10
Task 3

ELRC Workshop Report for the Netherlands

Author(s): Carole Tiberius
Jan Odijk

Dissemination Level: Confidential

Version No.: <V1.0>

Date: 2021-08-03

Contents

1 Executive Summary

This document reports on the third ELRC Workshop in the Netherlands, which took place online via Zoom on the 11th of June 2021 because of the COVID-19 pandemic situation. The workshop language was Dutch and an interpretation service was offered from Dutch to English for non-Dutch speaking participants. We particularly thank Livewords for the interpretation services. There were overall 65 participants who attended the online event. Most participants came from the Netherlands, some from Belgium and a small number from other countries.

The 3rd ELRC Workshop in the Netherlands aimed to engage participants in a constructive discussion on the readiness and usability of language technologies for Dutch for small and medium-sized enterprises (SMEs) and public administrations. Developers, integrators and users of LT, both from the private and the public sector, shared their experiences, requirements and ways for transforming digital interaction in an increasingly multilingual environment with the help of LT. Also, the value of language data was illustrated, and practical ways of sharing language data were discussed.

The ELRC office, Andrea Lösch, Eileen Schnur and Stefania Racioppa, fully supported the local Technical National Anchor points, Carole Tiberius (Dutch Language Institute) and Jan Odijk (Utrecht University), in the organisation of the event. This facilitated a smooth conduction of the workshop without any technical difficulties.

The following section includes the agenda of the event (Section 2); Section 3 briefly informs about the content of each presentation of the workshop (Subsections 3.1-3.7). In Section 4, there is a summary of the discussion raised during the workshop. In Section 5, we focus particularly on the Country Profile for the Netherlands. All presentations by the speakers are available at https://www.lr-coordination.eu/thenetherlands3rd.

2 Workshop Agenda

The workshop agenda was as follows:

09:15 - 09:30	**Opening Zoom session**
09:30 - 09:40	**Welcome and introduction** *Carole Tiberius, Instituut voor de Nederlandse Taal*
09:40 - 10:00	**The potential of Language Technology and AI – where we are, where we should be heading** *Khalil Sima'an, Universiteit van Amsterdam*
10:00 - 10:20	**ELRC, Language Technology and AI for Dutch** *Carole Tiberius, Instituut voor de Nederlandse Taal & Jan Odijk, Universiteit Utrecht*
10:20 - 10:30	**Coffee Break**
10:30 - 11:00	**The CEF AT Platform** *François Thunus, Europese Commissie*
11:00 - 11:40	**Language technologies by/for the public sector - Panel session** • *Emma Hartkamp, Europese Commissie, directoraat-generaal Vertaling (Moderator)* • *Catia Cucchiarini, Taalunie* • *Arjan van Hessen, Universiteit Twente* • *Oele Koornwinder, Belastingdienst* • *Harvey van der Meer, Gemeente Tilburg*
11:40 - 11:45	**Short Break**
11:45 - 12:15	**Language data creation, management and sharing: existing practices and challenges - Panel session** • *Steven Krauwer, Universiteit Utrecht & CLARIN ERIC (Moderator)* • *Franciska de Jong, Universiteit Utrecht, Executive Director CLARIN ERIC* • *Hans Overbeek, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties; Kennis- en Exploitatiecentrum Officiële Overheidspublicaties (KOOP)* • *Ted van der Togt, Koninklijke Bibliotheek, Afdeling Onderzoek* • *Vincent Vandeghinste, Instituut voor de Nederlandse Taal, ELG*
12:15 - 12:30	**Conclusions**

Apart from a slight delay towards the end, the agenda was followed as planned; there were no changes or major technical difficulties.

3 Summary of Content of Sessions

3.1 Welcome and introduction

Carole Tiberius, Technical NAP for the Netherlands (Instituut voor de Nederlandse Taal), welcomed the participants and explained the practicalities of the 3rd ELRC Workshop in the Netherlands. She then set the context of the event and introduced the agenda of the workshop.

At the end of the introduction a live poll was conducted, revealing the organisational affiliation of the workshop participants. As Figure 1 shows, both research/academia (38%) and the public sector (30%) were well represented, accounting for more than half of the participants. A smaller number of participants represented the other sectors: 13% were coming from SMEs, 8% were LT/AI providers and another 11% classified themselves as representing another sector.

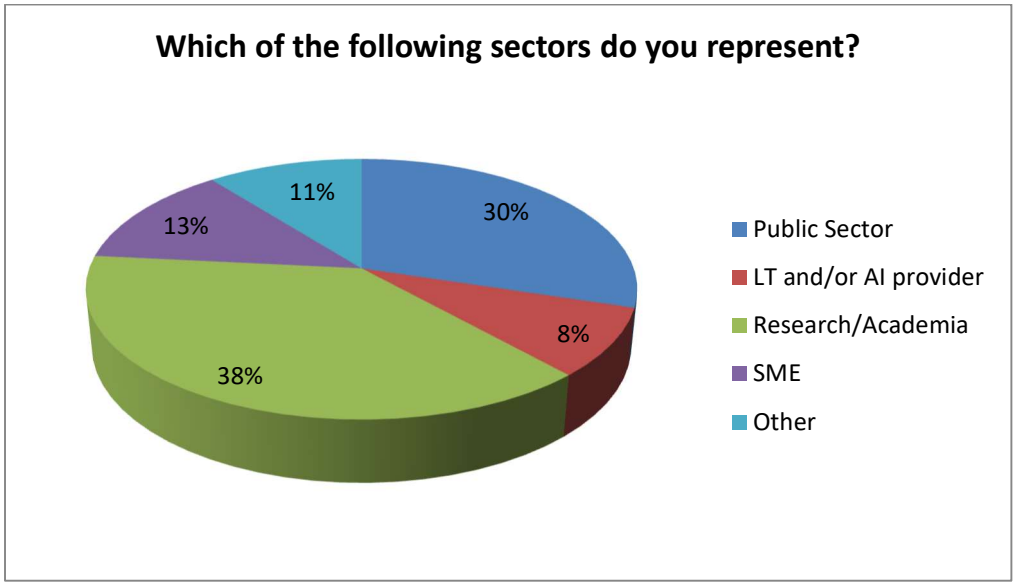


Figure 1: Organisational affiliation of the participants[1]

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

The keynote speech by Prof. Khalil Sima'an (University of Amsterdam) addressed the potential of AI and new trends with regard to making LT work. In particular, Prof. Sima'an presented a personal perspective for the future, with a focus on language and the role of data. AI represents one of the greatest opportunities for global societal and economic progress. There are many initiatives in Europe with a focus on AI, e.g. AI4EU.

AI is also becoming more and more common in our daily life (e.g. digital personal assistants, chatbots, intelligent cars). AI will play an increasingly important role and the crucial components are human communication, classification, prediction and also prediction/decision under uncertainty.

The level of digitisation is crucial for developing and applying AI. When we look at the DESI index (of 2019), we see that the Netherlands are at the forefront in Europe when it comes to digital performance.

[1] Please note that some of the participants assigned themselves to more than one sector, which is why they were counted twice

Prof. Sima'an then went on to discuss the positioning of AI in relation to Machine Learning and in particular to Deep Learning. Machine Learning and Deep Learning demand a large amount of training data. He also stated the importance of language communication in what it means for a system to be called intelligent. As Alan Turing rightly stated, if a conversation with a device cannot be differentiated from a similar conversation with a human being, then the device can be called intelligent. To achieve this goal, however, is not easy at all. To describe the complexity of everyday human experiences we need language, but natural language is rather ambiguous.

Prof. Sima'an illustrated how, from 2015 onwards, Deep Learning revolutionised developments in machine translation. Statistical machine translation (SMT) systems needed large amounts of parallel data. Sentences were split into smaller units called phrases, and the system estimated a probability on how a phrase translates into other phrases. Given an input sentence, the sequence of phrases with the highest probability would be provided as translation.

With Deep Learning, internal representations of words, phrases and whole phrases are learned from data as vector representations. These can be learned in one task and reused for a variety of tasks. Just like semantic representations, in that sense.

Beside Machine Translation, Deep Learning is used for various other language technologies such as Speech Recognition or Question-Answering. Prof. Sima'an emphasised that data is the fuel for Deep Learning as a large amount of data is needed to develop models.

Towards the end of his presentation, he showed the Gartner's 2019 Hype Cycle for Digital Government Technology which indicates that cloud office will make the introduction of capabilities including everyday artificial intelligence (AI), mobile collaboration, collaborative content authoring and workstream collaboration part of the mainstream. Furthermore, within the coming 5 to 10 years, chatbots and dialogue systems are expected to trigger increased interest and become central goals for research and technology.

Prof. Sima'an concluded with a few requirements for future AI which he listed as follows: AI has to be Explainable, Verifiable, Physical, Collaborative, Integrative and Humane. AI systems need to be able to (1) "understand" people, (2) adapt to complex environments and (3) communicate adequately in complex social situations and environments. Language is the main source of expression of human experiences. Prof. Sima'an stressed that language is embedded in the situations in which these experiences occur and therefore understanding language implies a need to understand the situations in which it is used.

Following Prof. Sima'an's presentation there was an interesting question from the audience: will AI adapt to humans or vice versa? According to Prof. Sima'an, humans adapt to their environment, but they also adapt their resources to achieve efficiency, so AI is not used to imitate as many people as possible, but to achieve practical goals.

3.3 ELRC, Language Technology and AI for Dutch

This presentation was given by Carole Tiberius. She began by providing some information about the background and history of the ELRC project in the Netherlands. She briefly discussed the Country Profile for the Netherlands as included in the ELRC White Paper which was published at the end of 2019. The next part of the talk presented the current contents of the ELRC Share repository for resources including Dutch. She noted that although there is a fair number of resources for Dutch; Flemish in the repository, the majority of resources comes from Belgium.

Carole Tiberius then moved on to introduce the CEF AT Catalogue of Services (https://cef-at-service-catalogue.eu/), a comprehensive collection of various language technologies, tools and services of which the ELRC-SHARE repository is part. To date, the Catalogue of Services contains more than 690 different tools and services from more than 540 providers with headquarters in the EU. A corresponding browse and search function allows to find the right tools based on language coverage, domain, type, functionality, etc. Figure 2 below summarises the different types of tools that are currently available. Most interestingly, there are more than 73 tools/services available that were "made in the Netherlands" and more than 97 tools/services for the Dutch;Flemish language.

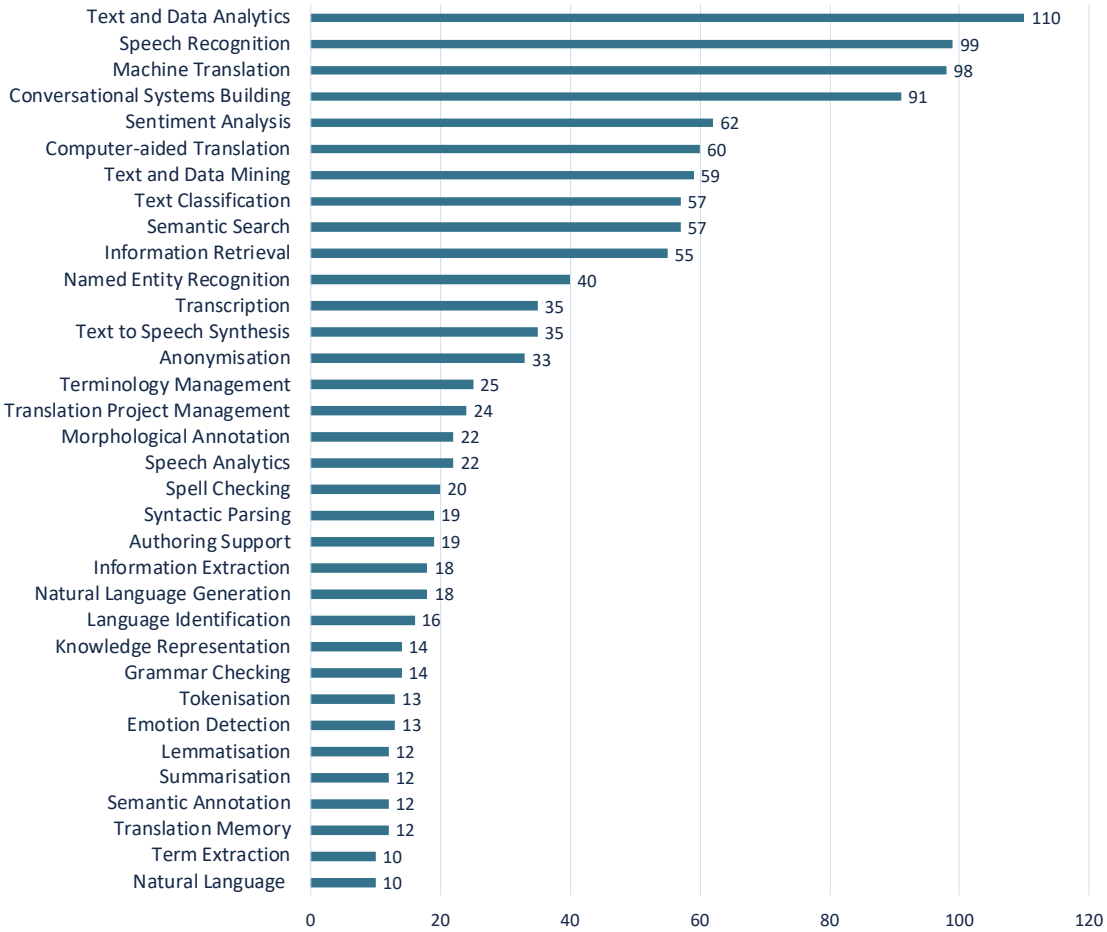


Figure 2: Tools/services available through the CEF AT Catalogue of Services (by type)

Carole Tiberius then concluded with two recent initiatives in the Netherlands: the NL AI Coalition and the NAIN project. The NL AIC has been set up to substantiate and stimulate AI activities in the Netherlands. It is a public-private partnership in which the government, the business sector, educational and research institutions, as well as civil society organisations collaborate to accelerate and connect AI developments and initiatives in the Netherlands. The NAIN (Netherlands AI for the Dutch language) project is one of the use cases of NL AIC focusing on language. It aims to set up an infrastructure for AI for the Dutch language, covering both speech and text.

At the end of this session, another live poll was launched to find out how familiar the workshop participants were with the various projects that had been introduced so far. The results from the poll showed that more promotion of the various projects is needed.

Did you already know about ELRC before the workshop?	
Yes	13
No	13

Have you already used tools or services from the Catalogue of CEF eTranslation services?	
Yes	4
No	22

Were you already familiar with the NL AI Coalitie before the workshop?	
Yes	9
No	17

Were you already familiar with the NAIN project before the workshop?	
Yes	4
No	22

3.4 The CEF AT platform

The CEF AT platform was presented by François Thunus (DGT, European Commission). He presented the evolution of the EC's machine translation system from the statistical to the neural paradigm and its development to cover more language technologies through the CEF AT platform. The target users of the CEF AT platform (in particular CEF eTranslation) are:

- Translators and staff of the EU Institutions
- Digital services of the EU Institutions
- CEF Digital Service Infrastructures
- Pan-European digital public services
- Public administrations in EU Member States, Iceland and Norway
- European SMEs (as of March 2020)

The CEF eTranslation service can be accessed either through:

- a web user interface to automatically translate documents and text snippets or
- an API to integrate machine translation in workflows, websites, digital services, etc.

François Thunus stressed that CEF eTranslation supports not only all official EU languages, Norwegian, Icelandic, but also Russian, Chinese (Mandarin), Japanese and Turkish (and more languages are coming). The system provides not only a general language engine, but also domain-adapted engines, such as the EU formal language engine, health, culture, etc.

François Thunus subsequently commented on the translation output quality of CEF eTranslation, underlining that, since the system had been trained on a huge database of translated official EU texts, it performs very well in translating formal EU language. On the other hand, he pointed out that quality level of translations may not be the same when it comes to non-standard or creative texts. However, the availability of the general language engine, which is trained on respective non-official texts, delivers high-quality output already now. The need to select the appropriate domain-adapted engine according to the text type to be translated was highlighted. Regarding the future development of the CEF eTranslation service, François Thunus noted that the EC is working on extending the domain coverage (e.g. scientific texts), as well as on supporting both additional non-EU languages of social and economic importance, and regional languages. He also mentioned the extension of the CEF AT platform to include additional language technologies such as speech recognition, anonymisation, named-entity recognition and a basic Computer-Aided Translation tool. Some of these tools have already been made publicly available at https://language-tools.ec.europa.eu/.

At the end of his presentation, François Thunus addressed some of the questions which had been raised by participants prior to the workshop or which had been collected out of the survey on the use of machine translation held before the workshop. These related to the quality and reliability of the translations. François Thunus answered that automatic translation should always carry a label (however good the translated text may be) and that the CEF eTranslation system is regularly tested and evaluated with standard tools and benchmarks, such as BLUE (BiLingual Evaluation Understudy) score and TER (Translation Error Rate) score which match translations with a known good translations. He also mentioned ongoing work on a specific tool that will allow to ascertain the quality of a translation using AI.

To a question raised about security, François Thunus answered that security is that what sets CEF eTranslation apart from other machine translation systems. CEF eTranslation is completely secure. The texts that are submitted for translation are not preserved, but are erased immediately after translation, not even staff can see them.

Another question related to the different varieties of Dutch, i.e. Dutch in the Netherlands and Dutch in Belgium was posed. François Thunus answered that at the moment no distinction is being made as they have no way of knowing this for the data from Euramis. At European level, there is only one Dutch. It is however possible to make a different engine for Dutch for the Netherlands and Dutch for Belgium provided that there is enough external data with specific tagging.

The last question that could be addressed during the workshop concerned feedback. François Thunus answered that feedback can be sent to a wiki page: https://ec.europa.eu/cefdigital/tracker/plugins/servlet/desk.

During the workshop, there was unfortunately no time to present the results from the online survey on CEF eTranslation that was held prior to the meeting. Here we briefly summarise the results that show that the CEF eTranslation is not yet widely used by the respondents. We received 24 responses and the majority indicated not using or not having used CEF eTranslation.

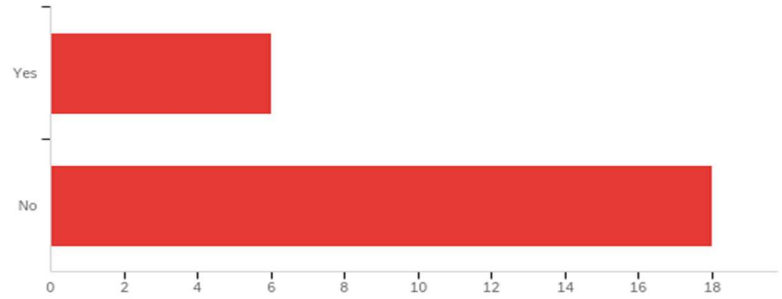


Figure 3: Are you currently using CEF eTranslation or have you used CEF eTranslation in the past?

Most respondents answered using other systems, as can be seen in the figure below:

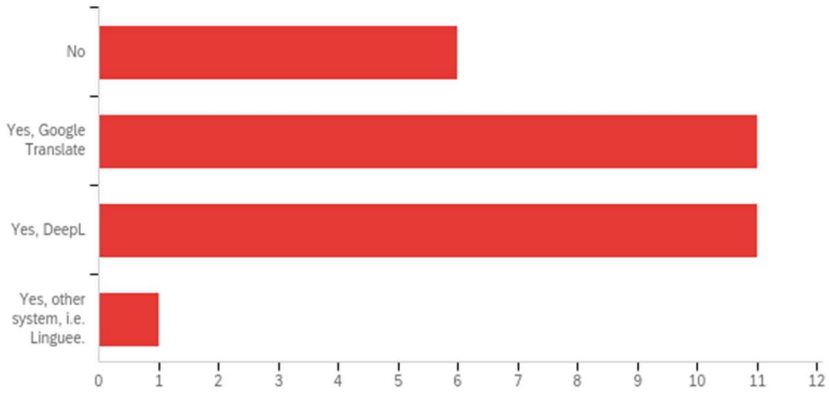


Figure 4: Do you use other machine translation systems?

Other systems that are used are mainly Google and DeepL. One respondent indicated using Linguee.

The survey revealed that those using CEF eTranslation use it for translating different types of text ranging from general texts, policy texts, parliamentary proceedings, administrative texts to technical manuals. Gisting was also mentioned here. The use CEF eTranslation for translating documents is more common than using the system for the translation of snippets. The wish for the support of sound files (MP3) was expressed.

English-Dutch is the most used language combination, followed by Dutch-English. Other language combinations that were mentioned are English-French and Greek-French.

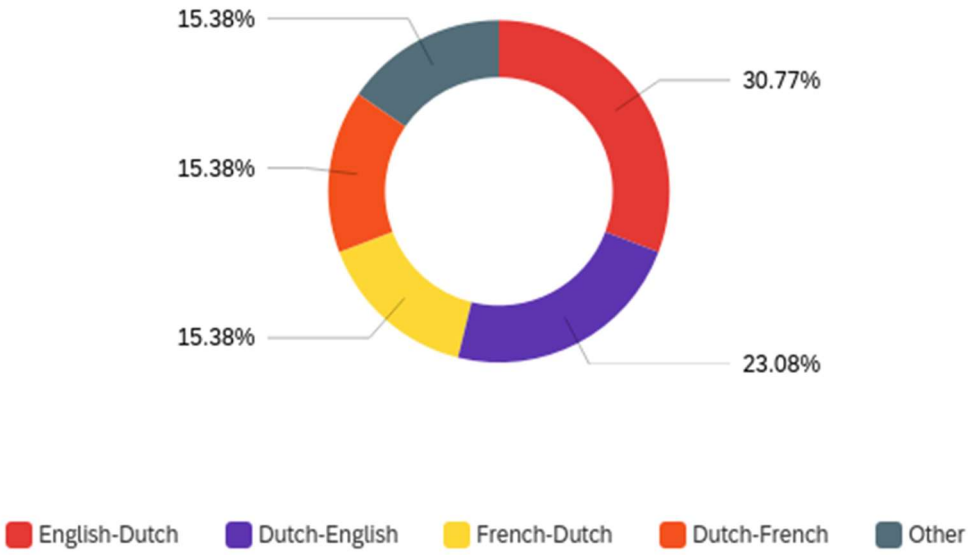


Figure 5: Language combinations used for translating with CEF eTranslation

In general, those using CEF eTranslation are reasonably satisfied with the quality of the output from the system. Overall CEF eTranslation is considered easy to use and it has a good processing time.

The respondents particularly liked the following features:

- grammatical correctness;
- the fact that the EU is trying to keep up with developments in machine translation;
- well suited to policy texts, mostly accurate terminology;
- clear user interface;
- the people behind the machine are very approachable for explanations/tips/suggestions.

The respondents considered the following as less good:

- terminological inconsistency;
- clumsy formulations, uncertainty about terminological correctness/exactness;
- translation at sentence level instead of at text level;
- recently, in the NMT, especially for NL-EN, negative influence of ill-considered data collection via web scraping;
- currently no distinction between NL-NL and BE-NL (but that is a task for the Member States and the eTranslation team should not be blamed for that).

3.5 Language technologies by/for the public sector (Panel session)

The first panel session addressed the demand side, specifically the demands and needs of the public sector for LT-enhanced digital services in the Netherlands. The panel was moderated by Emma Hartkamp (Europese Commissie, directoraat-generaal Vertaling in Den Haag). She initiated the session introducing the four panelists:

- **Catia Cucchiarini**, Senior Policy Officer at the Taalunie (Dutch Language Union), Senior Researcher at Radboud University, Nijmegen
- **Oele Koornwinder**, Business Analyst at the Belastingdienst (Tax Office)
- **Arjan van Hessen**, Speech Technologist at Twente University, Utrecht University and the SME Telecats
- **Harvey van der Meer**, Strategic Advisor and Innovation Lead at the Gemeente Tilburg (Tilburg Municipality), GEM Product Owner

Before looking at concrete examples of LT applications (being) developed by the panelists and their needs for further LT support, Emma Hartkamp briefly presented possible roles of the government in the development of AI and LT. Governments can act as financier, regulator, convener and standards-setter, data steward, smart buyer and co-developer, and user and service provider. She also described the situation in the Netherlands. Although there are many ongoing initiatives in the Netherlands and various policy documents have been published (Digitaliseringsstrategie ”Digitisation strategy” (2020) and the Strategisch Actieplan voor Artificiële Intelligentie “Strategic Action plan for Artificial Intelligence” (2019)), it is not always clear who takes care of the implementation. She noted that maybe a separate ministry for Digital Affairs is needed.

To trigger discussion, Emma Hartkamp gave a few examples of interesting MT and LT-based services in Europe, i.e. Re-open EU, Conference for the Future of Europe, Plata (the MT platform of the Spanish government), MT-HUB (a platform for public services in EU-countries), Hugo.lv (the language technology platform of the Latvian government) and eJustice.

Then, Catia Cucchiarini gave her perspective on the use of LT in the public sector. Many “nice” tools have been developed, but they are not being used as much as they could and should. The Taalunie could help to further promote their use. One of the projects that the Taalunie is currently involved in is Netwerk Begrijpelijke Overheid ”Network Understandable Government” which focusses on improving communication between government and citizens. This network could be used to raise further awareness and to promote CEF eTranslation and other LT tools. Catia Cucchiarini also suggested to carry out a systematic assessment of needs in the public sector (similar to assessments carried out in the past by the Taalunie), to identify what the needs are and how these needs can be fulfilled.

Harvey van der Meer presented GEM: a virtual conversational assistant for citizens. Municipalities are the first point of contact to the government for citizens. This contact should be easily accessible and reliable such that Dutch citizens can get an answer to their questions 24/7 regardless of the municipality they belong to. The goal is to automate what can be automated in order for employees to have more time for tasks that cannot be automated. To make GEM easily accessible, GEM is omnichannel. It can be accessed via WhatsApp, Google Assistant (experimental phase in Tilburg) and by phone (also experimental). Recently they have started integrating translation in GEM using DeepL, but they will also look into the possibility of using CEF tools. Reliability of the translation is of course an important criterion.

Oele Koornwinder gave a short introduction to the Belastingdienst (Tax Office). The Belastingdienst is currently undergoing a transition to become more customer-oriented. According to Oele Koornwinder, language and speech technology can support this. Think for instance of automatic text classification, writing assistants (model letters), virtual assistants and making the website accessible in multiple languages. He noted that domain-specific tools, large anonymous datasets, customatisation of tools and a database of public sector terminology are needed and that security, privacy and transparency are important criteria for software and tool selection.

Arjan van Hessen explained that currently quite good results are being achieved with speech recognition. One of the projects he is working on is a project on providing subtitles for the parliamentary debates. He noted that going from speech to text is fine as long as a literal transcription is required. However, the next step, going from text to report, i.e. understanding the text, is still very difficult. Spoken language includes unfinished sentences, ungrammatical sentences, etc. For humans, it is perfectly understandable as long as some context is provided, but if the context is taken away, it is much more difficult to understand what is meant. As Khalil Sima'an already pointed out in his keynote, we need to be able to understand the meaning.

In the last part of the panel session, Emma Hartkamp focused on a few important criteria for LT tools in the public sector:

Quality: Are the available tools good enough for the task at hand?

For instance, in the context of 'Easy Language'[2], will LT tools yield sufficient quality to help automating the task? Catia Cucchiarini thinks that it is possible to use LT to support this process. There are many tools available and it should be possible to train them for the specific purpose of retranslating to an easier language. There is a clear need for this.

User Experience: How are the tools perceived by the users? Has user experience research been done?

This question was specifically addressed to Harvey van der Meer. Do citizens appreciate using a virtual assistant? Harvey van der Meer answered that this varies. There is a trade-off between automating what can be automated and what cannot. How can we make information as easy accessible as possible for continuous service providing. They found that low-literate people (who often cannot handle all the information found on government websites) particularly appreciate using a virtual assistant as it involves less reading.

Customisation: Why is customisation needed?

A concrete example of the need for customisation was given by Oele Koornwinder. Letters often contain a call of action, i.e. the action the recipient of the letter needs to take. If this call of action is not clearly stated, the recipient will not understand it. The letters and the corresponding calls of action are different for different organisations and departments. Therefore a general tool will not work, but the tool will need to be customised to the individual organisations and departments' specifities.

The final question addressed in the panel session related to distortion of competition.

[2] See Handbook of Easy Languages in Europe: https://www.frank-timme.de/verlag/verlagsprogramm/buch/verlagsprogramm/bd-8-camilla-lindholmulla-vanhatalo-eds-easy-language-in-europe/backPID/easy-plain-accessible.html

To what extent should national governments or the European government develop tools? Will this not result in distortion of competition for companies?

Arjan van Hessen acknowledged the importance of this aspect while recognizing that it is extremely complex and cannot be answered easily. Governments and especially the EU have a lot of money. The CEF tools are great, but if you can do the same with commercial tools, e.g. Google or DeepL, then it is difficult to say why you would choose one and not the other. Maybe CEF tools should be made commercially available to companies. Arjan van Hessen finally noted that public procurements generally lead to open source tools. This mitigates some of the distortion. However, another workshop would be needed to discuss this in detail.

Emma Hartkamp concluded the session noting that one of the main challenges preventing data sharing is often still the lack of data management plans. This was discussed in the next panel session.

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

The final panel session of the ELRC workshop in the Netherlands focused on language data and sought to investigate the policies, legal framework and infrastructures for sharing language data in the Netherlands. The panel was moderated by Steven Krauwer (Universiteit Utrecht & CLARIN ERIC). He briefly presented the main discussion topics and format of the panel. The panel was structured in three rounds.

The panel brought together different stakeholders, two that could be considered as producers and providers of data, and two representing platforms for sharing data with others.

In the first round, the 4 panellists introduced themselves.

Ted van der Togt works at the Research Department of the KB, National Library of the Netherlands. The National Library has been collecting publications since 1789 and its ambition is to make as many publications openly available as possible. The National Library collaborates extensively with the National Archive, which also hosts a large collection of archives including private archives, and as data provider faces similar challenges to the National Library.

Hans Overbeek is advisor contents standards at KOOP (Kennis- en Exploitatiecentrum Officiële Overheidspublicaties) which is placed under under the umbrella of the Ministry of Interior and Kingdom Relations. KOOP is the Dutch publications office. It serves as the official publisher of the central and local governments of the Netherlands. These publications can be found on **officielebekendmakingen.nl, overheid.nl** and **data.overheid.nl**.

Vincent Vandeghinste is a Senior Researcher at the Instituut voor de Nederlandse Taal (Dutch Language Institute) and represents the Netherlands within ELG (European Language Grid). ELG is a platform somewhat comparable to the CEF catalogue of services, but structured per company/data provider and providing direct connections to APIs, tools and services.

Franciska de Jong, Executive Director CLARIN ERIC, introduced CLARIN, the Common Language Resources and Technology Infrastructure that exists since 2012. It is an ERIC-type research infrastructure, which means that countries are the primary participating parties. CLARIN is organised as a network of centers (currently more than 60), of which 25 are CTS certified data centers. There is a strong focus on FAIRness and interoperability (by federated login, central metadata harvesting). CLARIN focusses on language data, both spoken and written.

After the introductions, the first discussion topic was tackled: **What is the framework for sharing language data in the Netherlands?**

Ted van der Togt explained that in principle it is the National Library's ambition to bring all printed publications from the public domain online (via services like Delpher and DBNL). First and foremost such that people can see the publications, but ideally also such that the data can be studied and analysed. In this context, the project "Web Publications for digitized content" was mentioned, which is carried out together with the Nationaal Archief, TU Delft and Bureau van Leeuwen & van Leeuwen, and aims to investigate to what extent digitized material can be made even more accessible online while complying to the W3C standard Web publications. Machine Translation could potentially play a role here. For out of commerce and copyrighted material, the National Library has to negotiate with rights holders. The National Library also delivers data services and Linked Open Data via data.bibliotheken.nl. Furthermore, they collaborate at national and international level - Clariah, Future Library Lab, European Digital Reading Lab, Cultural AI Lab. Data from the National Library can be found at KB LAB https://lab.kb.nl/. Data from the Nationaal Archief can be found at *NA Datalab* https://www.nationaalarchief.nl/over-het-na/datalab-nationaal-archief.

Hans Overbeek then discussed the framework for sharing data from the perspective of KOOP. He noted that, in the Netherlands, Dutch is the only official language for official publications and that the need for translated national legislation is underestimated. There are translations, often made by third parties, but they are not easily retrievable (e.g. there is no clear link between the translation and the publications on the websites of the central and local governments). Two legislations are important in this context. First, "Wet open overheid" (Open Governement Act) which implies an active disclosure of "everything" via PLOOI (PLatform Open Overheid Informatie by KOOP). Second, "Wet Elektronische Publicatie" (Electronic Publication Act) which implies that all legislation and regulations become available online.

Looking at institutional level, data is made available at data.overheid.nl. The open data platform provides access to a catalogue of all data sets (open and closed) on the one hand and on the other hand, it functions as a data broker to help finding and disclosing hidden data sets.

Concerning Language Technologies, KOOP does not use Machine Translation, but does use entity extraction and semantic web technology (SKOS and OWL) for metadata and reference data.

Vincent Vandeghinste explained that ELG is a Horizon 2020 project that aims to set up a long-term platform with a focus on the commercial sector to reduce fragmentation of the LT business environment in Europe (currently high number of specialised companies). The project is actively collecting data from companies and organisations that should be included in the ELG platform.

Franciska de Jong then provided an insight on the framework for sharing data from the CLARIN perspective. She noted that policies differ in different countries and determine what is possible in each country. In the Netherlands there are some important developments. Investments in a national AI programme have recently been granted (NL AI Coalition). The government also encourages Open Data according to FAIR principles (https://www.openscience.nl/) and researchers are also encouraged to make their data accessible where possible. CLARIN supports data sharing and promotes the use of standards. CLARIN is also contributing to the EOSC federation of services (see Figure 6) with so-called thematic services for language data, which increases visibility of harvested data.

After this rather positive view of existing frameworks for sharing data in the Netherlands, the panel focused on **the main challenges for sharing data in the Netherlands and possible solutions to overcome these**.

Ted van der Togt first discussed the situation at the National Library. The National Library is creating a New Digital Storage to enhance accessibility of data. Standards and description of data quality are essential for this. In this regard, OCR data constitute a challenge for the National Libray. Many OCR data sets need to be curated before becoming really useful (for AI but also for other purposes). However, who does the curation? Where do we store this curated data and what standards do we use for this curated data?

Hans Overbeek focused on the situation at KOOP. The main challenge KOOP faces is the switch from passive disclosure (Wob) to active disclosure (Woo). More data bring another challenge of making all data retrievable and "findable". KOOP continues to improve its infrastructure. Metadata and standards are important as well as the use of common reference data. Hans Overbeek concluded: *No AI (Artificial Intelligence) without IA (Information Architecture)*.

Vincent Vandeghinste presented his own perspective describing the obstacles encountered when trying to obtain data for research. These obstacles concern, amongst others, limited availability (research only, commercial use, online service versus download). He noted that the difference between source data and online availability is not always clearly understood by the data provider. Furthermore it is often not clear what can be done with the derived data. Legal issues also prevent data sharing. Legal departments tend to use the GDPR as an excuse for not sharing data.

According to Vincent Vandeghinste some of the challenges could be overcome by:

- Building automated checks to keep information up to date
- Keeping regular contact with contact persons at each resource provider
- Making prototype contracts for different situations available to everyone
- Making prototype informed consent forms available
- Providing examples of large organisation that make their data / tools available
- Providing tools for automated anonymization / pseudonymization, such that everyone uses the same tools (with the same quality).

Franciska de Jong confirmed that data sharing is further complicated by unclarity and uncertainty on how to comply with the GDPR framework. There is also a certain contradiction as working on open science can be at odds with sharing data with commercial bodies. Further challenges are posed in research contexts where there are not always clear guidelines as for where to deposit data to comply with rules for the management of research data. Also there is limited institutional capacity for supporting researchers, and finally there is a focus on in-house development which is not always aligned with emerging standards rooted in developments elsewhere.

According to Franciska de Jong, clarity on roles and responsibilities at the various relevant levels - national organisations (e.g. OCW, NWO, VSNU, etc.), individual universities and academic institutes; faculties and departments; individual researchers and disciplinary communities – must be addressed urgently.

Data sharing can also be stimulated by including incentives in the assessment and by introducing a reward system for academic researchers generating data sets.

To conclude, Franciska de Jong showed a picture of EOSC, the European Open Science Cloud, a cloud of disciplinary clouds. The picture shows that within sub-clouds, in particular in the Humanities Cloud within the SSHOC cloud, there are clear collaborations to coordinate services to researchers from the various sub-disciplines. Collaboration is key to future service providing.

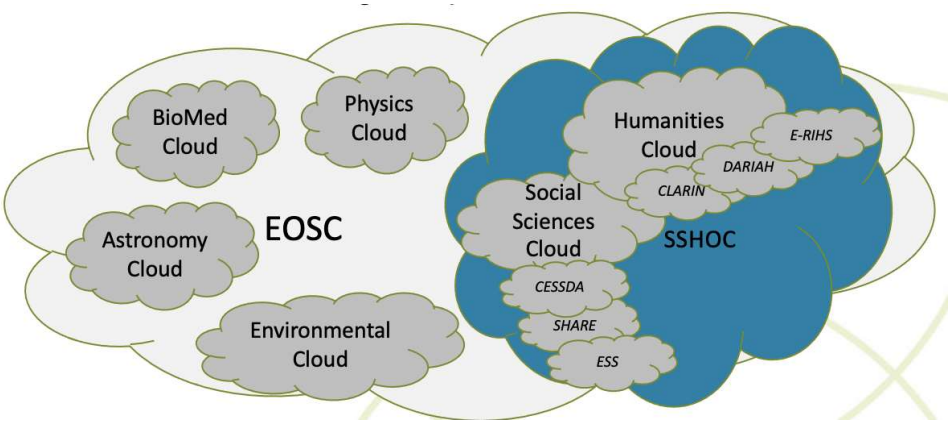


Figure 6: The European Open Science Cloud[3]

3.7 Conclusions

The workshop was concluded with a short wrap-up session by Carole Tiberius (Instituut voor de Nederlandse Taal) who stressed once more the need for data in order to realise the full potential of AI and language technologies for Dutch. Workshop participants were encouraged to share their data through the ELRC-SHARE Repository and hence to support the further development of CEF eTranslation. Carole Tiberius also mentioned the ELRC Technical and Legal Helpdesk that can support potential data donors, e.g. in finding the right licenses, cleaning the data, etc. Last but not least, participants were encouraged to participate in the feedback evaluation survey.

[3] https://marketplace.eosc-portal.eu/
Related question from the chat:
12:30:17 From Oele Koornwinder : Welke FAIR-principes worden er in de European Open Source Cloud toegepast. Die van Go Fair (focus op interoparability)? Ander perspectief is fairness en tegengaan van bias: belangrijk bij trainen van eerlijke en inclusieve modellen: transparantie over de samenstelling van de dataset.
12:35:07 From Franciska de Jong | CLARIN : @Oele: Go FAIR; wat niet wil zeggen dat responsible data science (o.m. tegengaan van bias) niet ook aandacht krijgt

4 Synthesis of Workshop Discussions

The workshop succeeded in bringing together relevant stakeholders from the public sector, research and academia and to a lesser extent from SMEs. We also succeeded in raising further awareness on the importance of language data for AI and LT and of shared repositories. The workshop also showed that ELRC and CEF eTranslation are still insufficiently known and more promotion is needed. The ELRC Technical National Anchor Points will join efforts with Emma Hartkamp (Europese Commissie, directoraat-generaal Vertaling) and the Taalunie (Catia Cucchiarini), and Edwin de Koning (Category Manager Interpreting and Translation Services, Ministry of Justice and Security) to pursue active promotion of CEF eTranslation (and other CEF tools). The ELRC Technical Anchor Points will also continue collaboration with NOTaS (Nederlandse Organisatie voor Taal- en Spraaktechnologie; foundation representing research institutions and application developers in the Language and Speech Technology in the Netherlands) to further promote the ELRC initiative. We repeat some of the main points raised during the workshop.

- **A major obstacle to data sharing is that there is no organised or centralised exchange of language data at national level.** There are no clear roles and responsibilities at the different levels. Maybe a separate ministry for Digital Affairs is needed.
- **For application and development of LT tools at national level, the distinction between Dutch as used in Belgium and Dutch as used in the Netherlands is important.** This distinction is needed as both countries have their own terms for specific concepts. This distinction may not be important at European level, but it is important at the national level. It would be good if data repositories could include this information in the metadata to increase reusability of the data.
- **Standards for metadata**[4] are essential for sharing data.
- **Data sharing** can be improved by providing tools at national or European level for e.g. automated anonymisation / pseudonymisation, such that everyone uses the same tools (with the same quality) instead of different tools with different quality.
- Important criteria for using LT tools in the public sector are **security** and **privacy**. These aspects should be emphasised in promoting CEF eTranslation as an alternative to e.g. Google.
- A major challenge for future AI and LT is being able to truly understand the meaning of texts.
- **Collaboration** between different platforms and infrastructures is key to future service provision.

[4] Related question from the chat:
12:24:58 From Oele Koornwinder : Standards of Metadata kun je ook verbinden aan een databank voor overheidsterminologie inclusief ontologische relaties (taxonomie / wordnet).
12:30:19 From Hans Overbeek : @oele: inderdaad, er is samenwerking op het gebied van gegevenscatalogi, met definities van termen die gebruikt worden in (basis) register

5 Country Profile: Language data creation, management and sharing

The situation in the Netherlands with regard to language data creation, management and sharing practices has not changed significantly since the publication of the Country Profile as part of the ELRC White Paper end of 2019.

Legal issues as well as the absence of data management practices (or even guidelines governing the sharing of language data) in (public) services remain the main barriers hindering the sharing of language data in the Netherlands.

The workshop brought together relevant stakeholders. Future collaboration is key to good service providing.