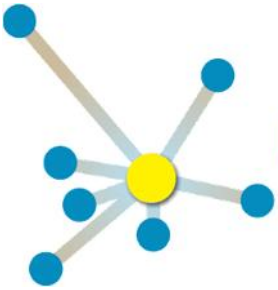
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination

Connecting Europe Facility

Deliverable D3.2.6
Task 3

ELRC Workshop Report for Germany

Author(s): Alexandra Soska
Andrea Lösch
Andreas Witt
Eileen Schnur
Stefania Racioppa

Dissemination Level: Public

Version No.: <V4.0>

Date: 2021-05-11

Contents

1 Executive Summary

The 3rd ELRC Workshop for Germany took place on April 20, 2021. It was organised by the German Research Centre for Artificial Intelligence (DFKI) and closely coordinated with the ELRC National Anchor Points for Germany, Alexandra Soska (Federal Ministry of the Interior, Building and Community) and Prof. Andreas Witt (Leibniz Institute for the German Language). Moreover, the event was collocated with the 5th National Workshop of the European Language Grid which provided additional valuable insights into Language Technology (LT) research and development in Germany.

The 3rd German ELRC Workshop aimed to engage participants in a constructive discussion on the readiness and usability of German language technologies 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 analysed.

Overall, the workshop agenda was structured around the following main topics: (i) the state of the art of language centric AI in Germany, (ii) the demands and needs of both public services and SMEs with regard to LT, and (iii) the availability and management of language data in public services and SMEs.

The main findings of the workshop were that machine translation (MT) has by now found its place as a widely used technology both in public services and SMEs, closely followed by information retrieval/search tools. Other technologies such as chatbots and speech solutions are on the rise but not widely taken up yet, presumably because of insufficient maturity with regard to the German language. Also with regard to the sharing of language data, advances were made even though the overall legal framework and organisational constraints (lack of data management plans or even guidelines) limit the sharing of language data. On the other hand, latest developments (e.g. in the form of the new Data Usage Act ("*Datennutzungsgesetz"*)) pave the way for a better standardisation and organisation of the management and sharing of data.

2 Workshop Agenda

2.1 Workshop Programme (German version)

European Language Resource Coordination
Connecting Europe Facility

PROGRAMM

09:30 – 09:40 **Begrüßung durch das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)** *(Andrea Lösch, Gruppenleiterin Daten und Ressourcen, DFKI)*

09:40 – 10:00 **Sprachtechnologie und Künstliche Intelligenz: Neuste Trends und Perspektiven** *(Josef van Genabith, Wissenschaftlicher Direktor Forschungsbereich „Sprachtechnologie und Multilingualität ", DFKI)*

10:00 – 10:20 **Die CEF AT Plattform und ihre Weiterentwicklung** *(Andreas Eisele, Projektmanager Maschinelle Übersetzung, Generaldirektion für Übersetzung, Europäische Kommission)*

10:20 – 10:45 **Sprachtechnologien in und für Deutschland: Textklassifikation und der „Catalogue of Services"** *(Jörg Feuerhake, Referent Maschinelles Lernen und Imputationsverfahren , Statistisches Bundesamt; Andrea Lösch, Gruppenleiterin Daten und Ressourcen, DFKI)*

10:45 – 11:00 Kaffeepause

11:00 – 11:45 **Sprachtechnologien im öffentlichen Dienst und in KMU** *(Alexandra Soska, Übersetzerin, Bundes - ministerium des Innern, für Bau und Heimat; Jochen Hummel, CEO Coreon GmbH)*

11:45 – 12:30 **Erstellung, Management und Teilen von Sprachdaten: Neueste Entwicklungen im öffentlichen Dienst und in KMU in Deutschland** *(Andreas Witt, Abteilungsleiter „Digitale Sprachwissenschaft", Leibniz -Institut für Deutsche Sprache; Alexandra Soska, Übersetzerin, Bundesministerium des Innern, für Bau und Heimat; Ralf Lemster, Vizepräsident, Bundesverband der Dolmetscher und Übersetzer)*

12:30 – 12:45 **Zusammenfassung und Fazit** *(Andrea Lösch, Gruppenleiterin Daten und Ressourcen, DFKI)*

ab 14:00 5. Nationaler ELG-Workshop

2.2 Workshop Programme (English version)

09:30 - 09:40	**Welcome by the German Research Center for Artificial Intelligence (DFKI)** *(Andrea Lösch, Group Leader Data and Resources, ELRC project manager, DFKI)*
09:40 – 10:00	**Language Technology and Artificial Intelligence: Latest trends and perspectives** *(Josef van Genabith, Scientific Director Research department „Multilinguality and Language Technology", DFKI)*
10:00 – 10:20	**Advances of the CEF AT platform** *(Andreas Eisele, Project Manager Machine Translation, Directorate-General for Translation, European Commission)*
10:20 - 10:45	**Language technologies in and for Germany: Classifications of text input with machine learning methods & the Catalogue of Services** *(Jörg Feuerhake, Assistant Head of Section "Imputation and Machine Learning", Federal Statistical Office; Andrea Lösch, Group Leader Data and Resources, Project leader ELRC, DFKI)*
10:45 – 11:00	***Coffee Break***

11:00 – 11:45	**Language Technologies for the public sector and SMEs** *(Alexandra Soska, Translator, Federal Ministry of the Interior, Building and Community; Jochen Hummel, CEO, Coreon GmbH)*
11:45 – 12:30	**Creating, managing and sharing language data: Recent developments in public services and SMEs in Germany** (*Alexandra Soska, Translator, Federal Ministry of the Interior, Building and Community; Andreas Witt, Head of the Department "Digital Linguistics", Leibniz Institute for the German Language; Ralf Lemster, Vice President Public Affairs, BDÜ German Federal Association of Interpreters and Translators)*
12:30 – 12:45	**Summary and Conclusions** *(Andrea Lösch, Group Leader Data and Resources, ELRC project manager, DFKI)*
In the afternoon (from 2 pm)	**5th National ELG Workshop (**https://www.european-language-grid.eu/5th-national-elg-workshop-germany**)**

3 Summary of Content of Sessions

3.1 Welcome and introduction

After a short introduction to the context and practicalities of the 3rd German ELRC Workshop, Andrea Lösch (DFKI) introduced the agenda of the day. Moreover, as a warm-up, a first test of the live poll and online survey was successfully made with the participants, revealing the organisational affiliation of the workshop participants (see Figure 1 below). As Figure 1 shows, almost half of the participants (49%) were from the public sector, 15% were SMEs and another 6% classified were LT/AI providers.

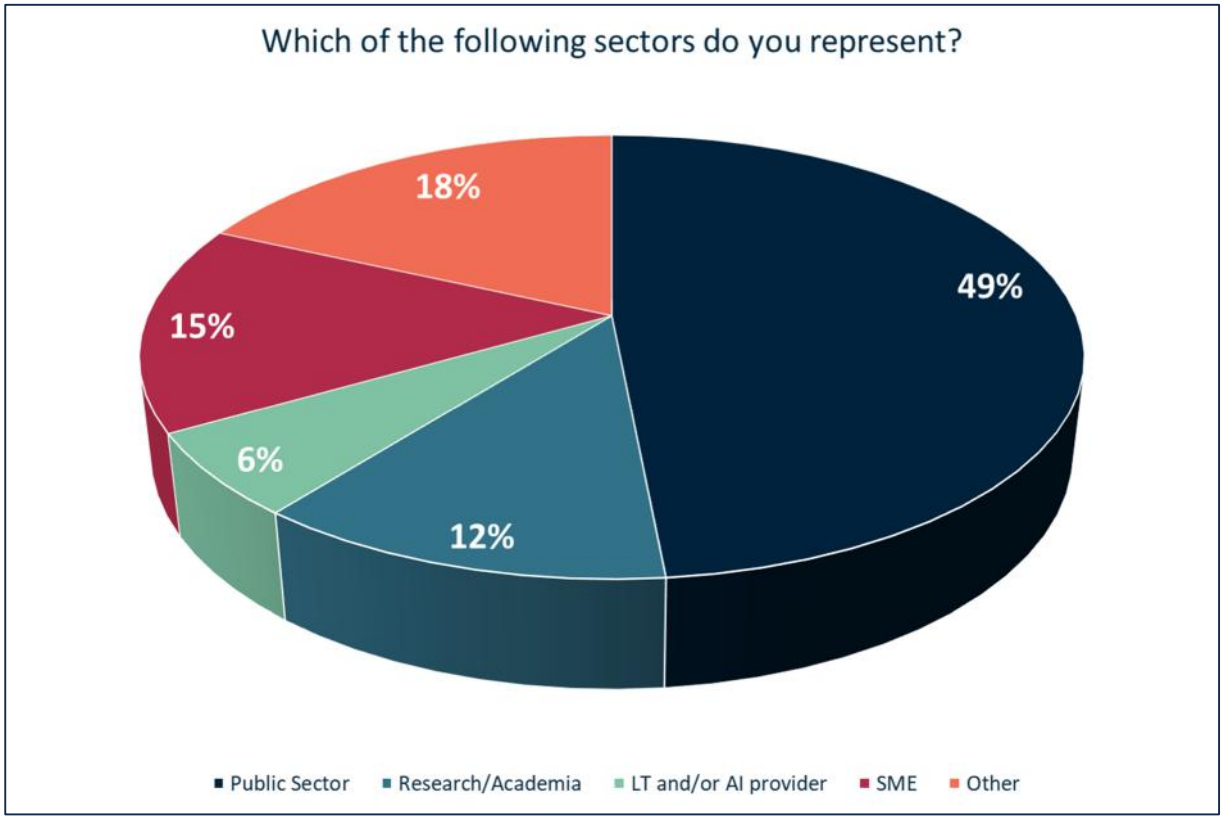


Figure 1: Organisational affiliation of participants

3.2 The potential of Language Technology and AI

The keynote speech by Prof. Josef van Genabith (DFKI) addressed the potential of AI and new trends with regard to making LT work. In particular, Josef van Genabith addressed the question of data and whether it was possible to train a machine translation (MT) system without parallel data. He explained that artificial neural networks work in a very similar way as the biological neural networks that constitute animal brains. Both are based on a collection of connected units or nodes called neurons and each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron that receives a signal then processes it and in turn can signal neurons connected to it. The "signal" at the connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and

edges typically have a weight that adjusts as the learning proceeds. The weight increases or decreases the strength of the signal at a connection. Signals typically travel from the first layer (the input layer), to the last layer (the output layer). Figure 2 below illustrates this process for a system recognising the picture of a dog.

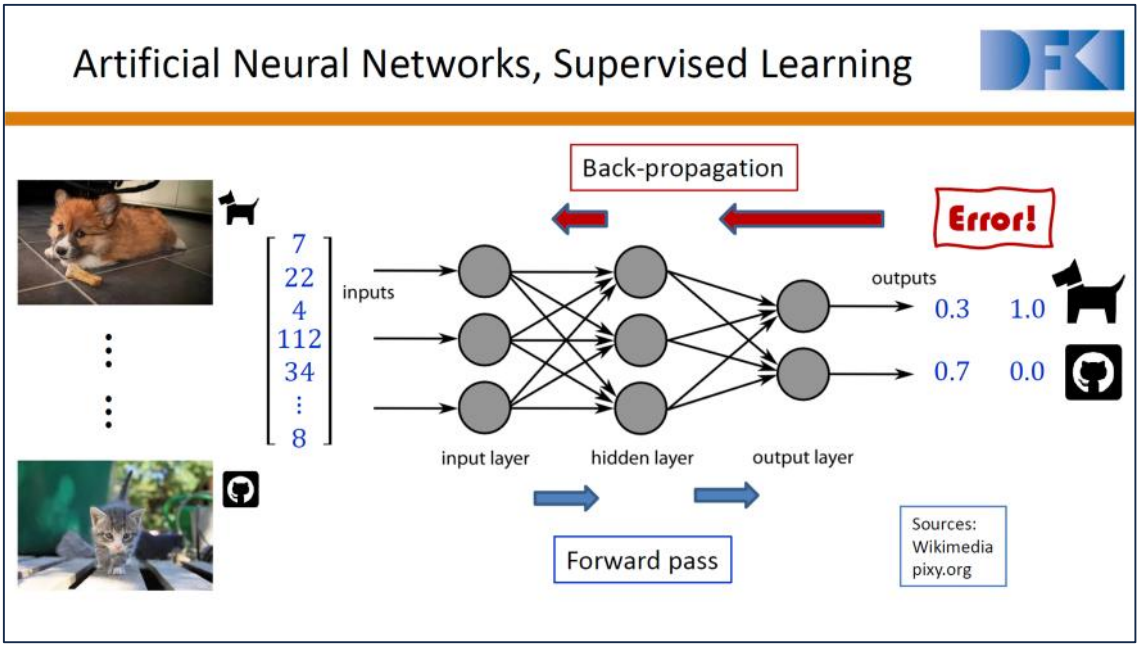


Figure 2: Example of an Artificial Neural Network: Identifying images of dog vs. cat

The same principle applies to machine translation which is also based on the training of artificial neural networks (neural machine translation, NMT). In the context of NMT, however, the training data are parallel texts with input in one language and output in another language. Words are represented by numbers/vectors that indicate their place in a multidimensional room. Such word embeddings allow words with similar meanings to have a similar representation. For instance, the words “sofa”, “armchair” and “couch” would have a very similar representation, while “zebra” and “algebra” would be quite different. As a consequence, this way of representing words also allows to “calculate” with words, e.g. “king – male + female = queen”.

However, as Prof. van Genabith pointed out, the problem of this approach is that many languages are lacking parallel data to train an NMT system. As such, research groups working on so-called “unsupervised MT” try to make use of the massive amounts of monolingual data that is available in order to train a system to translate into another language. This works by first calculating the word embeddings for the monolingual data (see below, Figure 3 (A)) and then to align them (see below, Figure 3 (B) and (C), finally arriving at translations from one word to another (see below, , Figure 3 (D)).

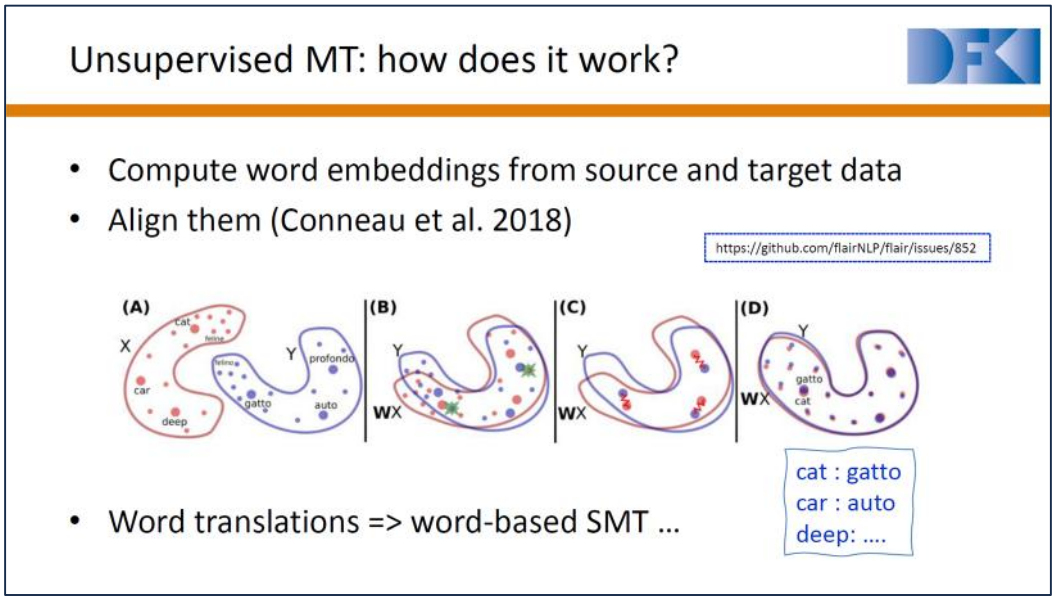


Figure 3: How does unsupervised MT work?

To date, unsupervised MT manages to achieve reasonable results as Prof. van Genabith explains. A new trend that may be able to achieve even better results when relying on monolingual data may be so-called "self-supervised NMT". In the case of self-supervised NMT, word embeddings are again calculated and aligned and an NMT system is initialised with them. For all sentences, the system will then compare the embeddings, and only keep the ones that are reasonably similar, while disregarding the ones which are not of sufficient similarity. In doing so, the system basically only considers the best possible data for its training. Initial results at one of the world's largest conferences on machine translation (WMT) show that this is a very promising approach, even though there are still slight differences between supervised NMT and self-supervised NMT, as the results of the WMT reveal. Prof. van Genabith hence concluded that for now, the availability of relevant parallel language data is the key for systems development to get the best possible output.

Main discussion points:

- **How can self-supervised and unsupervised MT cope with technical texts?** Following Josef van Genabith's presentation, one participant was wondering how the described procedures can cope with technical texts. The answer was that if you have good training data in general language (not specialised text), self-learning and unsupervised learning methods can help to generate synthetic data through back translation, which can be useful to improve the applicability of MT systems to specialised texts.
- **How can MT handle changes in language?** Another participant asked how MT could actually handle language change. Josef van Genabith confirmed that languages are not static and constantly evolving, as they are reflecting our changing reality. He stated that the available language data freezes a certain point in time, but that the training data is also steadily increasing, and that the corpora used to train MT are continuously growing. As an example, Josef van Genabith explained that weighting and using recently added language pairs more than older resources could help to reflect language change in MT to some extent.

- **What are actual low-resourced languages or languages for which there are too few resources?** One of the participants wanted to know where he/she can find an overview of the languages where there are no or only few resources available. The speaker answered that big international companies such as Google, Facebook or Amazon can provide useful hints on that. As an example, he mentioned that Google is currently supporting approximately 100 language pairs (some of them are covered by using pivot languages) and that there is a huge interest in extending the language coverage from the 100 to the 1000 most widely used languages in the world. Unsupervised or self supervised MT would be a starting point to reach that, but they need to be developed further to make it work. Andrea Lösch added that the **ELRC-SHARE repository** can also be a useful source in this respect, as it provides an overview of the resources available for European languages and shows the number of resources that are available for each language/language pair: https://elrc-share.eu/ Last but not least, one of the participants hinted on TAUS, which is dealing with "long-tail language pair data", sending the following link to the chat: https://datarade.ai/data-products/taus-parallel-text-colloquial-domain-english-low-resource-see-description-taus.

3.3 The CEF AT platform at work

The CEF AT platform was presented by Andreas Eisele (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.

CEF eTranslation supports all official EU languages, Norwegian, Icelandic, Russian, Chinese (Mandarin), Turkish, and Arabic and provides not only a general language engine, but also domain-adapted engines, such as the EU formal language engine, health, culture etc.

Andreas Eisele 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 was very good in translating formal EU language. On the other hand, he pointed out that translations may not be of the same quality 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, Andreas Eisele noted that the EC is working on extending the domain coverage (e.g. scientific texts), as well as on supporting additional non-EU languages of social and economic importance, and regional languages.

Last but not least, Andreas Eisele also illustrated the extension of the **CEF AT platform** itself by adding more 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/.

Finally, the presentation concluded with an overview of the eligible users (Public Administrations, Universities, CEF-funded projects, SMEs) and the steps and links to register and use eTranslation were presented. The following links were shared with the audience:

- Self-registration via https://webgate.ec.europa.eu/etranslation/public/welcome.html
- Web service (API) Technical documentation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/How+to+submit+a+translation+request+via+the+CEF+eTranslation+webservice
- CEF eTranslation Service Desk: help@cefat-tools-services.eu
- Access to eTranslation web user interface: https://webgate.ec.europa.eu/ETRANSLATION

There were no questions following Andreas Eisele's talk.

Most interestingly, the results of the live poll and online survey show that the vast majority (97%) of the workshop attendees has already actively used machine translation (see Figure 4 below). This means that independent of the usage scenario (SME, public administration, research/academia), MT has become a standard tool to support their daily work.

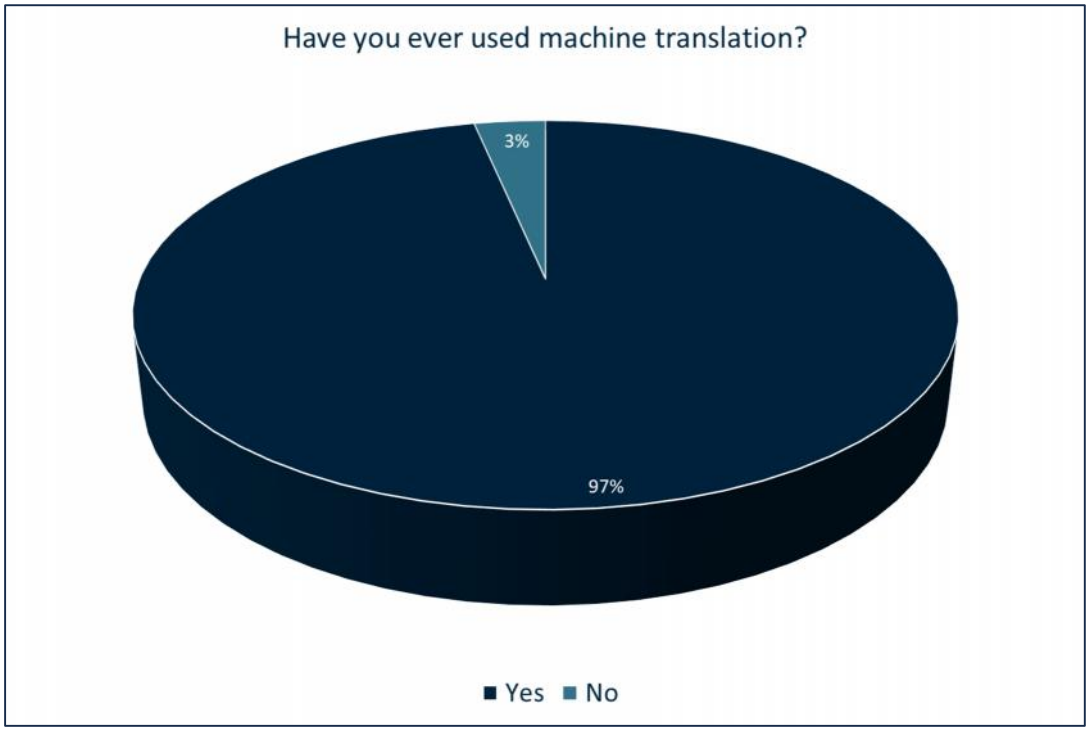


Figure 4: Use of MT as indicated by workshop participants

Even more, half of all workshop participants have been using both eTranslation and other systems (see Figure 5 below). One participant mentioned that in live poll question 2 "Which MT system did you use?", paid machine translation systems are missing as a possible answer and hence should be added to future surveys and/or live polls.

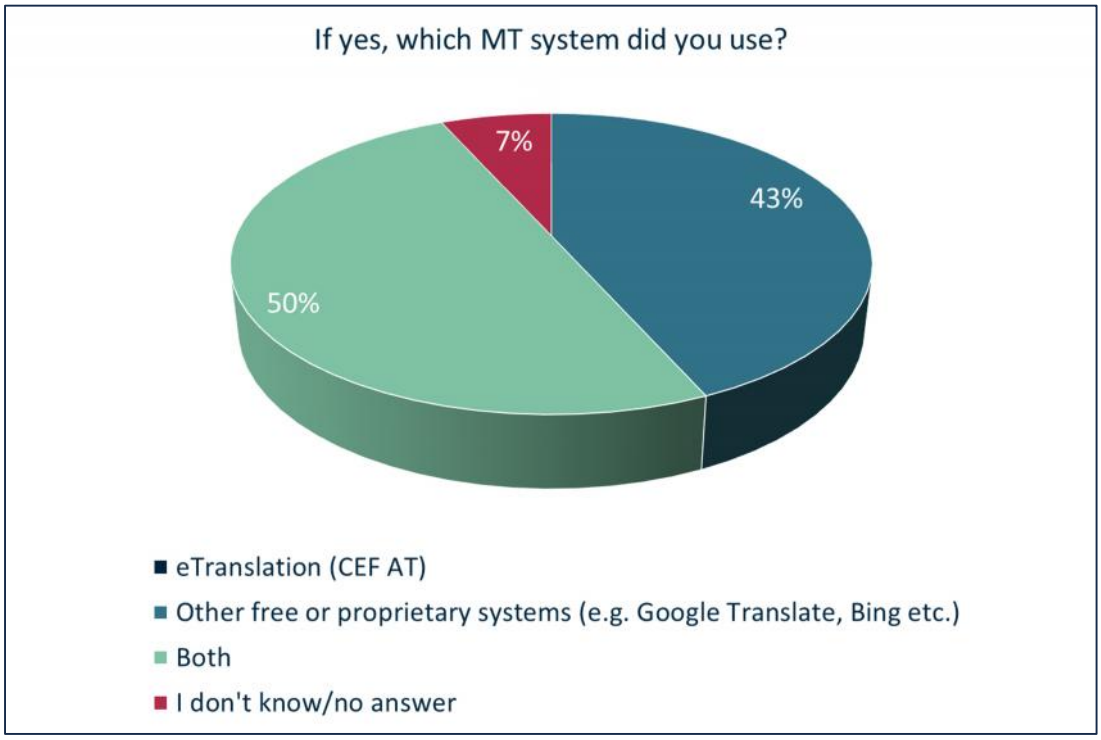


Figure 5: Type of MT systems used by workshop participants

Last but not least, the online survey and corresponding live polls revealed that only 3% of the workshop participants were not satisfied at all when it comes to the quality of machine translation from/into German. Almost 50% were either fully satisfied or very satisfied which shows a great improvement in terms of translation quality compared to earlier years and earlier workshops.

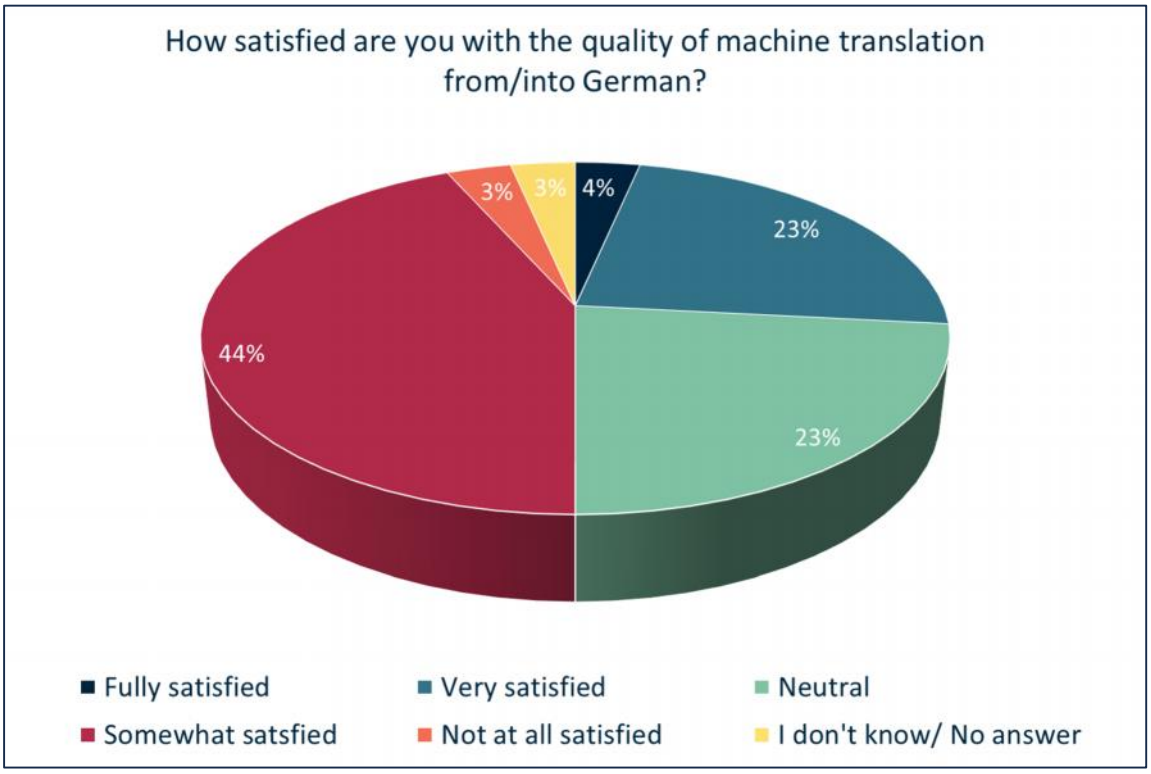


Figure 6: Satisfaction with the quality of machine translation from/into German

3.4 Language Technologies in and for Germany

This session was jointly presented by Jörg Feuerhake (Federal Statistical Office of Germany) and Andrea Lösch (DFKI). Jörg Feuerhake started with an illustration of how machine learning can be used for classification of free texts at the Federal Statistical Office of Germany. He pointed out that classification typically is a very labour- and time-consuming task if it has to be done manually. In addition to that, the overall volume of the texts that require classification has significantly increased in recent years. A typical example of where machine learning can support the classification task are the annual migration statistics. There are about 6 million migration cases every year that are recorded and include information about country of origin, date of birth, religion, and place of birth. Unanimously identifying for instance the place of birth, however, often presents a challenge as the example of "Mumbai" (or Bombay, Mumbai-India, Bombay Maharashtra, etc.) shows. Here, an AI-based classification trained on the records of the 6 million migration cases can help.

Another example concerns the free texts stemming from voluntary household surveys. Here, the free text information provided about expenses and activities were used to create a corresponding classification for systematic capture of expenses resulting in 89% accuracy. As such, AI-based classification approaches may not only reveal interesting information about people's circumstances, activities or behaviour, they can also be used to significantly reduce the time spent on data analysis at the Federal Statistical Office.

Following this very positive example of how language technologies can effectively support processes in public administrations and minimise efforts spent on day-to-day tasks, Andrea Lösch (DFKI) introduced the CEF AT Catalogue of Services (https://cef-at-service-catalogue.eu/), a comprehensive collection of various language technologies, tools and services that enable multi- and/or monolingual communication.

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 their language coverage, domain, type, functionality, etc. Figure 7 below summarises the different types of tools that are currently available. Most interestingly, there are more than 140 tools/services available that were "made in Germany" and more than 130 tools/services for the German language.

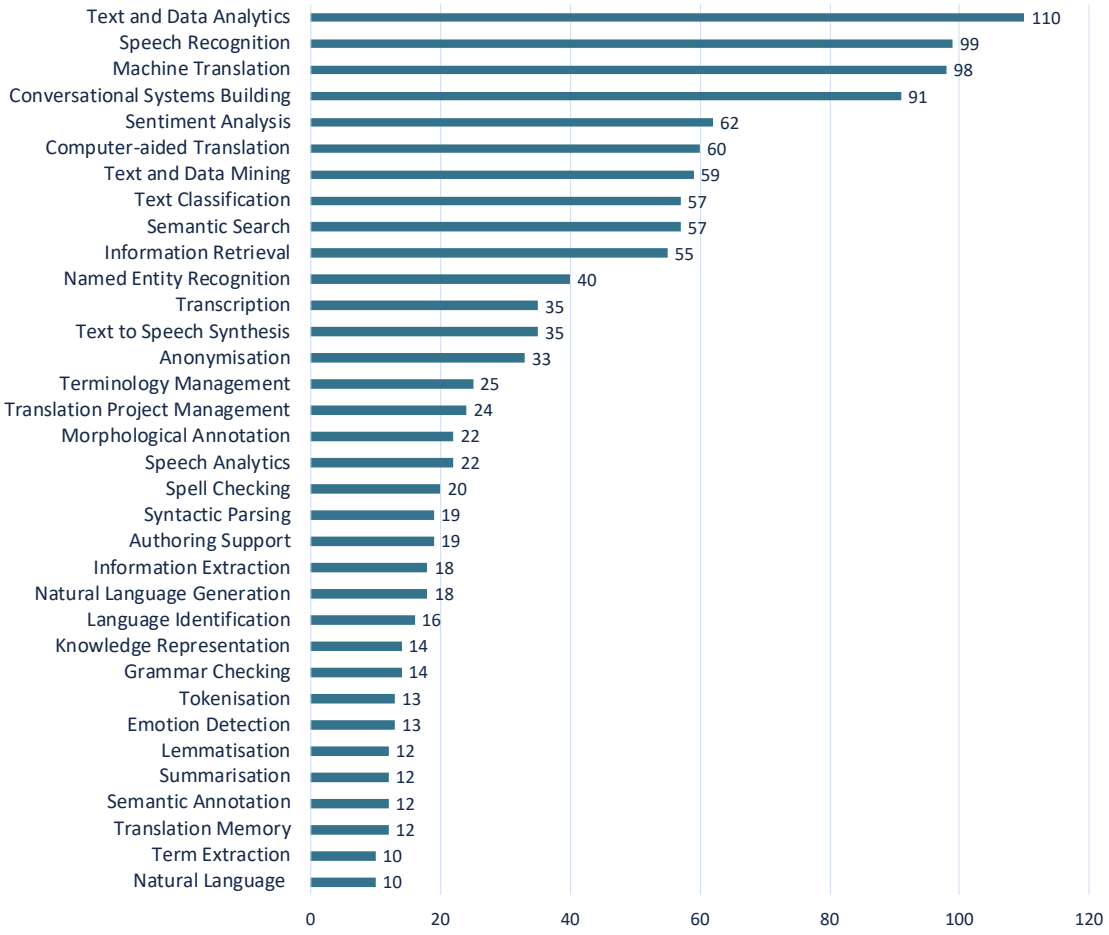


Figure 7: Tools/services available through the CEF AT Catalogue of Services (by type)

Main discussion points:

- **Is it better to use free text input or a solution that provides suggestions for classification when collecting data?** Following Jörg Feuerhake's talk, one participant explained that she is dealing with digitalisation projects, where one of the main topics is meta data of environmental data (e.g. Natura 2000). According to her, there is usually a lot of free text and using the metadata system can be complicated and time-consuming. Referring to the voluntary household surveys as presented by Mr. Feuerhake, she wanted to know if the speaker would stick to using free text instead of smarter solutions such as suggestions or classifications. As a response, Mr. Feuerhake explained that in the case he presented, the classification task for voluntary household surveys also aims at improving the survey tool. He added that such surveys are conducted over several years and that last time (about 5 years

ago), a very simple procedure was used, where people had to enter the information manually in their browser interface. This can be exhausting and a lot of effort. Now, the aim would be to improve the survey tool (e.g. uploading a picture of a receipt instead of entering the information manually, offering 3 to 4 appropriate categories and improving the quality of results). By doing so, they also want to reduce the workload and the amount of follow-up queries.

- **Can the classification software also be used for other purposes?** With regard to the question if the software could also be used for other purposes, Mr. Feuerhake confirmed that this would indeed be possible, as they were exclusively using open source technologies (Phyton R with the usual packages). However, the question would be how much effort it takes to adapt the software to the individual requirements/use case.

The subsequent survey and live poll showed that with regard to the different language technologies used by workshop participants, information search and retrieval and automated translation were by far the most widely used language technologies. Speech recognition and virtual assistants are not common yet. One participant mentioned, however, that the option "Correction of Grammar/Spelling" was missing as an answer to the live poll/survey question. Consequently, it can be assumed that this kind of LT also may be one of the technologies that is frequently used, even though it may not be considered as one of the tools/services to extend CEF AT with.

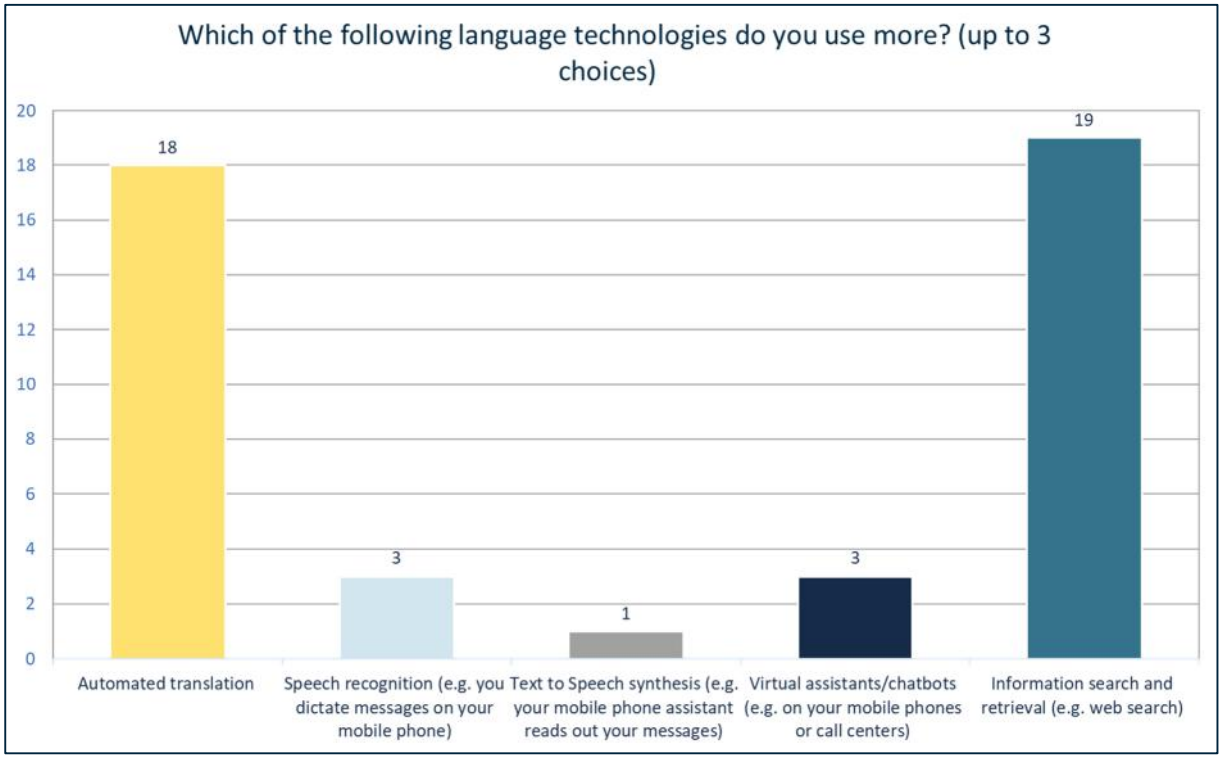


Figure 8: Tools/services most frequently used by workshop participants

In addition to that, the survey/live polls revealed that most participants (68%) used language technologies in their own language, underlining once more the need for LT support in the country's language (see Figure 9 below).

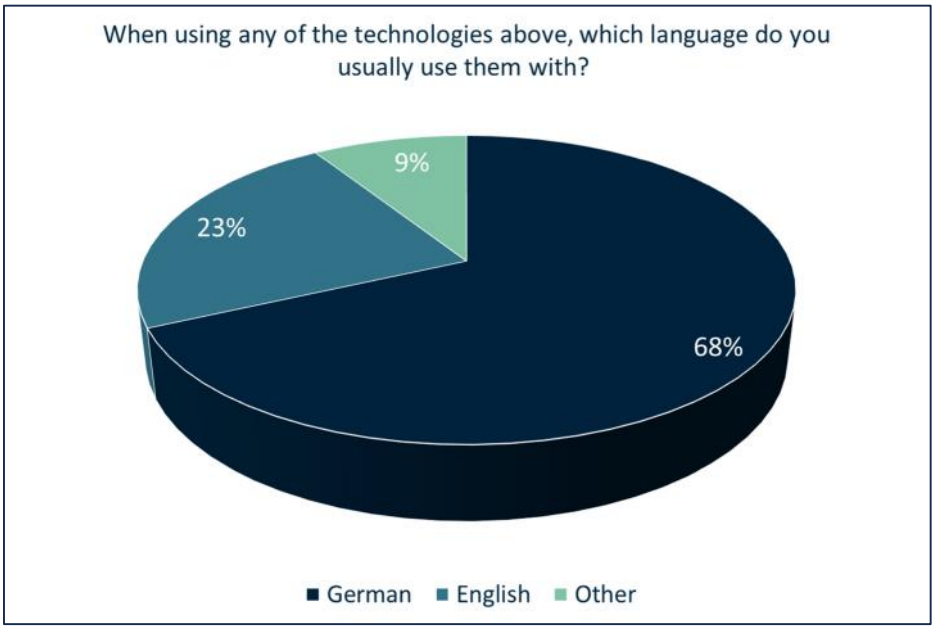


Figure 9: Use of LT in German vs. other languages

As regards the workshop participants' satisfaction with the quality and reliability of language technologies for German, the responses were also positive with 18% being very satisfied, 27% being somewhat satisfied and 36% being neutral. Nonetheless, there is significant room for improvement as the large number of responses selecting "neutral" indicates.

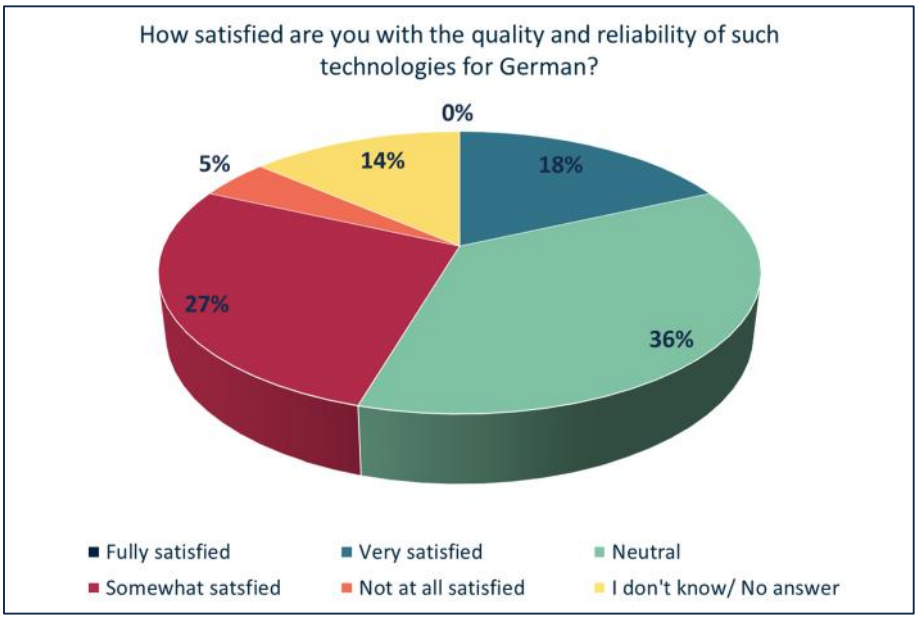


Figure 10: Perceived quality and reliability of language technologies for German

3.5 Language technologies in public services and SMEs

This session was jointly presented by Alexandra Soska (BMI), providing insights into language technologies in the federal administration of Germany and Jochen Hummel (LT-Innovate, Coreon),

illustrating the use of LT in companies across Germany. Alexandra Soska reported on one of the largest evaluation projects in the federal administration of Germany called “Machine translation system for gisting purposes for the federal administration”. The project is targeted at evaluating whether machine translation is capable of serving the needs of the more than 200.000 federal public servants in Germany. In order to achieve this goal, a corresponding working group “machine translation” was set up in December 2018 which consists of 14 representatives from 11 federal authorities. Since then, several important steps were taken, including:

- 2019: Development of (i) employee survey and (ii) evaluation grid
- January – May 2020: Programming of employee survey and tests
- May – July 2020: Conduct of employee survey
- Since September 2020: Statistical analysis of employee survey
- November 2020: Preparation of evaluation, including trial evaluation
- December 2020 – February 2021: Preparation of machine-translated bilingual files
- April 2021: Evaluation by language services

The final report on this endeavour is expected for December 2021. For the evaluation, the working group established its own error classification scheme that should be applied (see Figure 11 below) which builds on the Multidimensional Quality Metrics (MQM) .

Top category	Subcategory
Content	**Addition** Unjustified addition in the target text
	Omission Unjustified omission in the target text
	Untranslated Word Target text contains words that were not translated
	Incomprehensible word creation Target text contains words that do not occur in common language
	Mistranslation Target text contains words that were incorrectly translated
	Numerical value Target text contains a wrong numerical value. Please do not confuse with subcategory "Format"
Language	**Spelling** Target text contains orthographic errors
	Punctuation Punctuation marks were set incorrectly in the target text
	Word Formation (Grammar) Target text contains incorrect word forms
	Sentence structure (Grammar) The structure within the sentence is not correct in the target text
	Reference Incorrect relations between the sentences or phrases in the target text
Style	**Register** The stylistic level of the target text does not comply with the stylistic level in the source text, e.g. too formal/educational/colloquial/informal/vulgar
	Idiom The target text contains structures or formulations that were not adapted to the targeted language or culture
Country and Culture Specifics	**Realia** Country or culture-specific terms and formats are not correctly reproduced
	Titles and Names Target text lacks suitable equivalent for names and titles
	Format Information about date and time, numbers, measurement units or currencies were not reproduced in the format required by the target language.

Figure 11: Assessment of (machine) translations

Regarding the evaluation itself, the following process was employed:

1. The language services send source texts to the MT working group
2. The MT working group creates the corresponding machine translations
3. The MT working group creates bilingual files for the evaluators
4. The evaluators assess the quality of the machine translations
5. The MT working group creates the evaluation for each text

6. The language service creates the report for their authority
7. The MT working group collects the individual reports ands creates the Final Report.

Overall, 24 language services and 126 evaluators participated in the evaluation. 7 MT systems were considered and evaluated based on 371 source texts and 1.450 translations (XLIFF format). The following 16 language directions were assessed: DE-EN, EN-DE, DE-FR, FR-DE, DE-ES, ES-DE, PL-DE, DE-PL, RU-DE, DE-RU, SR-DE, DE-SR, AR-DE, DE-AR, FA-DE, DE-FA. As indicated earlier, the final results will be available by the end of 2021. However, it already became clear how important machine translation is for the well-functioning of the public administration.

Following the example of the introduction and use of language technologies in the federal administration of Germany, Jochen Hummel (CEO Coreon, Vice-President LT-Innovate) gave insights into language technologies in and for SMEs. He pointed out that in effect, language technologies and language intelligence aim at enhancing different functions of the human brain targeted at interacting (see machine translation, speech recognition, speech generation, dialogue management), reasoning (see natural language processing), learning (see finding, storing and retrieving data), or understanding (see intent recognition). As such, LT represents a key technology for enterprises in Germany. Not surprisingly, the majority of AI start-ups in 2021 in Germany could be assigned to the domain of LT companies, being either directly from the computer linguistics domain or supporting customer service and/or marketing functions of businesses.

A prominent example of the great success of LT is the German company Lengoo (www.lengoo.com) that provides enterprise-grade machine-translation services at scale. Thanks to its interactive modelling loop and seamless, continuous localisation, Lengoo won the Technology Fast 50 Award already the third year in a row. Another very successful German LT company is fyrfeed (www.fyrfeed.com) that uses AI to provide fully-fledged, individual social media posts as well as profile optimisation, hashtag analysis and an image design service specifically tied to a customer's company image. An excellent example for yet another very successful LT company in Germany is plusmeta (www.plusmeta.de). Plusmeta primarily operates in the manufacturing sector and provides service assistants, customer service chatbots as well as plug-in-AI for supporting technical documentation tasks (e.g. identification of duplicates, integration of supplier data, automation of workflows, standardisation of meta data, classification). Language technologies, however, can also take the form of Multilingual Knowledge Systems as the case of Coreon (www.coreon.com) from Berlin proves: Coreon combines taxonomies with terminology to create and deploy multilingual knowledge systems, making search, machine learning and IoT (Internet of Things) applications interoperable.

Following his illustration of the LT solutions landscape, Jochen Hummel pointed out that no matter what type of language technology, data is the key ingredient to make it work.

Main discussion points:

- **Will there be only one MT system for the federal administration or several?** As Alexandra Soska explained that in the presented project, several MT systems were evaluated depending on the language direction, Andrea Lösch asked if there were any plans to obtain and use more than one translation system within the language service division. Mrs. Soska confirmed that this would indeed be a possibility and that from experience, the language service already

knows that some of the systems work better/worse for certain languages. Depending on the results, one option would be to obtain several systems and to have them run via one platform. Even though the details are not known yet, this was considered as a possible option.

- **Will the evaluation report of the MT working group of the federal administration be publicly available?** One participant asked if the report will be made available. He hinted on the possibility to automatically evaluate MT using e.g. BLEU scores and mentioned that it would be very interesting to compare the results of human evaluation with those of an automatic evaluation, as this could provide useful insights for improving MT performance (e.g. through error classification). Confirming that this would be a very interesting idea, Alexandra Soska agreed to check if there are any plans to share the report and to get back to the participant individually.
- **Are there further plans for the introduction of LT in the public sector?** As a response to the question which language technologies would be considered next, Mrs. Soska explained that for the Language Service Division, MT can be seen as the first step towards text and speech processing and that processing of e.g. spoken language is currently not taken into consideration. She highlighted that she can only talk from her perspective and that there are other projects by German public administrations where such advanced technologies are already in use, at the same time expressing the wish to connect and to share experiences with them. Referring to the language service division, the speaker stated that if MT will be implemented, this could be the first step to much more, including data collection, the creation of domain-specific systems, etc.
- **To what extent was the quality assessment important in the MT evaluation in the public administration and how will it be considered in the future?** Getting back to human and automatic evaluation, one participant added that this would be an important part when assessing the quality of a translation engine. In addition, she asked if the EU engines were also included in this evaluation. Alexandra Soska answered that the project was not about identifying the best machine translation system, but rather about evaluating how/if the available systems can support the federal administration. Quality assessment would, however, certainly be included in the future. With regard to the question about the assessment of EU engines, the speaker confirmed that they were considered for all languages that are currently being offered by eTranslation (with the exception of Arabic, which was not yet available when the evaluation started).
- **To what extent is AI important for the development of multimodal systems with taxonomies?** Following his presentation, Jochen Hummel was asked what he thinks about the development of multimodal systems with taxonomies. The speaker confirmed that machine learning definitely has potential in this area. He explained that at Coreon, the focus is more on textual data and multimodal data are only considered “decoration” in the knowledge graph. He further elaborated that they can be implemented, but mainly have a supportive function. Mr. Hummel added that for classification and content production, graphs are playing a big role and are useful to disambiguate, to identify domains and to communicate knowledge. In the case of Coreon’s multilingual graph, the focus is, however, mainly on textual data.

Most interestingly, the results of the survey/live polls conducted during the event indicate that many workshop participants (62%) are not satisfied with the digital readiness of German public services and SMEs (see Figure 12 below). 29% of the respondents claimed they were not satisfied at all, and another

33% where only somewhat satisfied. Only 8% were satisfied in this respect! This means that while the availability of language technologies in and for German is ok (see Figure 10 above), their take up and implementation in both SMEs and public services are still not satisfactory for customers, citizens and employees who would benefit from greater digitisation.

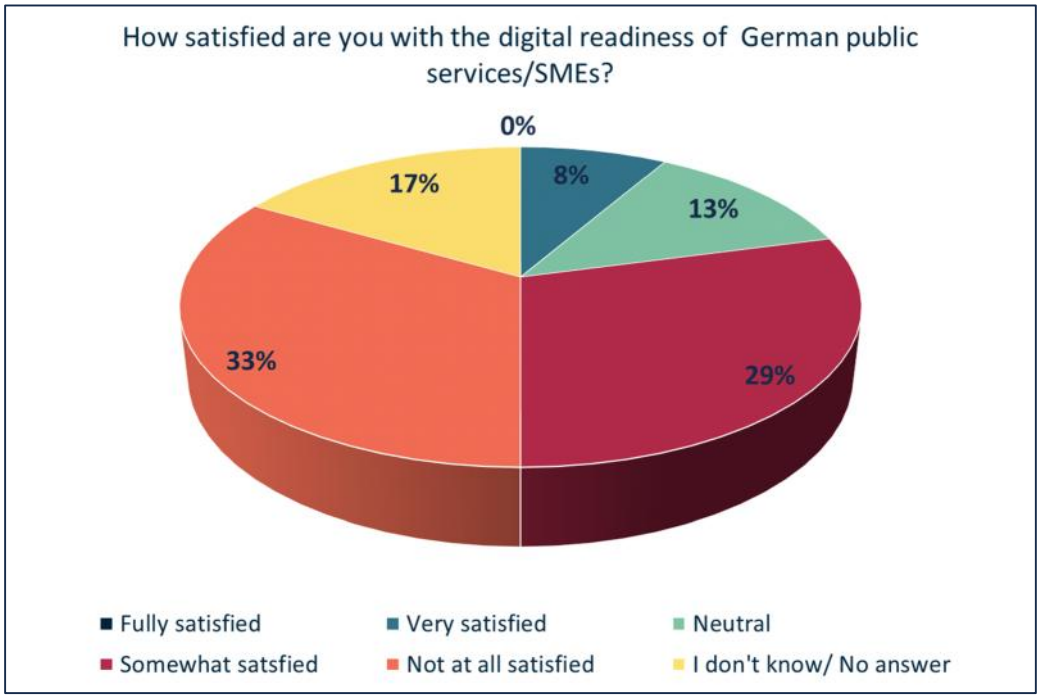


Figure 12: Digital readiness of German public services/SMEs - Level of satisfaction

With regard to the importance of supporting the development of language-centric Artificial Intelligence (AI) in SMEs/public administrations, respondents believe that the major role of these organisations should be the role of a data steward and user, also being a smart buyer and co-developer of such technologies (see Figure 13 below). The role as regulator seemed to be the least important function in this respect and would probably only apply to executive federal public administrations rather than to public services as such.

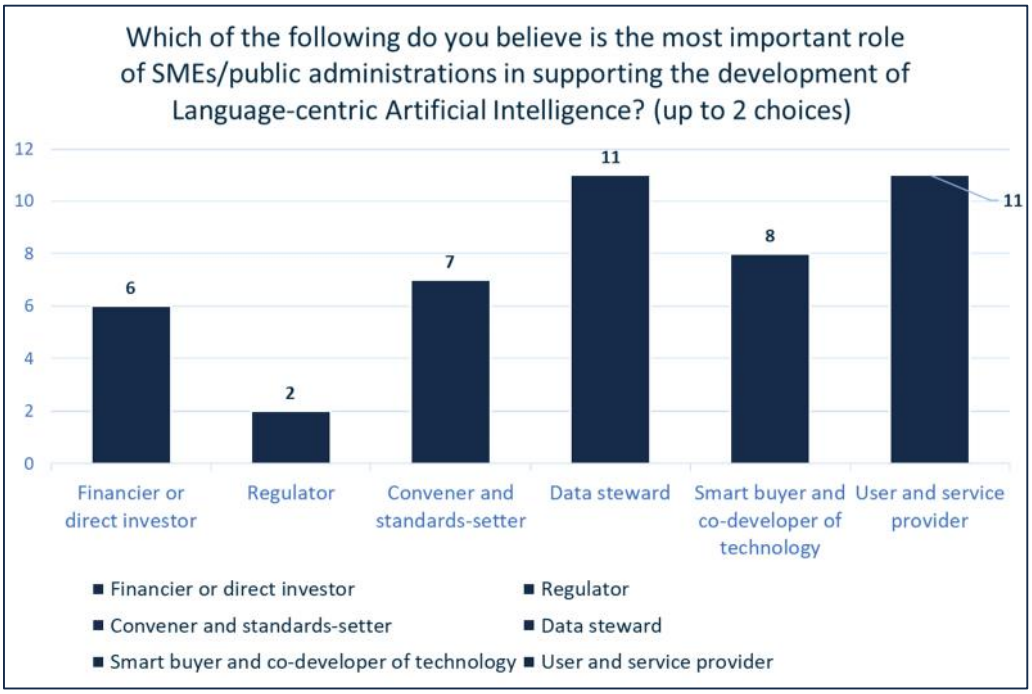


Figure 13: Role of SMEs/public administrations in supporting the development of language-centric AI

3.6 Language data creation, management and sharing

This session was kicked-off by Prof. Andreas Witt (Leibniz Institute for the German Language) who addressed the central legal question of how one can manage to share his/her language data. He started with an introduction to the German copyright law according to which the copyright always belongs to the creator of the work (e.g. a text or translation). In contrast to the U.S. copyright, the copyright in Germany is a non-transferrable right. However, the copyright owner can grant usage rights to others via licenses. In the context of the public sector, it is important to note that public sector texts (e.g. laws, administrative publications, administrative decisions etc.) are typically not covered by copyright. Prof. Witt illustrated several important developments in the European and German law relevant to the sharing of language data, including in particular:

- European law:
  - 2003: PSI Directive
  - 2013: PSI Directive II
  - 2019: Open Data[1]- and Public Sector Information[2] Directive
  - The latter must be implemented into national law by 17 July 2021.
- National law:
  - 2006: Information Re-use Act „Informationsweiterverwendungsgesetz (IWG)“
  - 2013: E-Government Act (changed 2020, further changes planned for 2021)
  - 2021: **Data Usage Act „Datennutzungsgesetz“** (coming soon, see above)

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L1024&from=EN
[2] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32013L0037&from=EN

The main goal of the upcoming Data Usage Act in Germany is to significantly extend the provision of open administrative data on the federal level and to facilitate the re-use of such data (including also commercial usage). The underlying principle is that **public sector information shall be freely usable** and available

- in machine readable format
- in all available languages
- free of charge
- with relevant meta data
- via the National Open Data Portal (www.GovData.de)
- with an open license, where necessary (e.g. CC-BY, DL-DE/zero, DL-DE/namensnennung)

Naturally, personal data are not considered as public sector information and hence shall be excluded (e.g. using anonymisation).

In her subsequent presentation, Alexandra Soska gave practical insights into how language data can be shared in public administrations. At the end of 2019, the probably largest language data collection initiative in the federal administration of Germany began to prepare for the German EU Council Presidency and more specifically, for the development of the German EU Council Presidency Translator (EUCPT - https://presidencymt.eu/#/). 15 public authorities participated in this challenge. Language data made available through the federal public administration included EN, FR, IT, ES, PL and in terms of content, only non-confidential data were used. In order to identify texts suitable for training the EUCPT, potential candidates were first identified by filtering the order management system (Auftragsverwaltungssystem AVS) for particular types of texts. Then, based on the corresponding order numbers, the relevant translations were identified and exported from the CAT tool. The data forwarded to train the EUCPT included both bilingual data (xliff, tmx) and word documents (docx). The actual processing and preparation of the data (alignment, cleaning, anonymisation) was then carried out by Tilde and DFKI. A corresponding **confidentiality agreement** was signed with both organisations. The clean data was then provided back to the federal administrative bodies of Germany. Despite this very successful sharing and re-use of language data, Alexandra Soska pointed out that several questions arose in this context, that need to be addressed with regard to the **future organisation of managing and sharing language data**, i.e.

- How can we best define relevant domains in our context? (Note: This is particularly relevant because the thematic areas of the texts to be translated vary greatly.)
- How can we usefully extend or modify the metadata in our data base to facilitate the export of shareable language data?
- Can the language services actually clean their data themselves in an easy way or will it always be necessary to rely on external help?
- How can we adapt the translation process in a way to enable the re-use of language data for MT training?

Ralf Lemster (Vice President of the German Association of Translators and Interpreters BDÜ) confirmed that the situation in the private sector in Germany is very similar. In contrast to the public sector, language data in the private sector may even be considered as **trade secrets** and/or as information providing a particular **competitive advantage** which is why there is a considerable

reluctance against the open sharing of this data. Even when sharing language data in a definable circle, the issue of **appropriate remuneration** is always present. As such, the situation is slightly more difficult than in the public sector. Nonetheless, Ralf Lemster assumes that a neutral, anonymous platform for sharing language data may at least help to overcome some of these issues and facilitate the sharing of language data in the private sector.

Main discussion points:

- **Can the ELRC-SHARE Repository enable the sharing of language data?** In his presentation, Ralf Lemster mentioned the need for a neutral, anonymous platform, making it easier for SMEs to share their data. Based on that, Andrea Lösch raised the question if the ELRC-SHARE repository could address this need, as it is provided by the European Commission and allows for data sharing and engine creation. Mr. Lemster confirmed that this could be a possibility, adding that the company must be able to choose if they want to share their data e.g. publicly or for internal use only. Andrea Lösch explained that the data can be licensed according to the data owner's preferences, and that it can be e.g. made publicly available, restricted by licences or exclusively shared only with the EC.
- **What about confidentiality agreements?** Ralf Lemster explained that even with the existence of platforms like ELRC-SHARE, other issues remain. He highlighted that when working with companies, external service providers need to sign extensive confidentiality agreements and translators will generally not be allowed to share the data themselves. In order to encourage the companies' openness towards data sharing, more information and more transparency of use would be required.
- **Why are such few language data available through Open Data Portals?** Following up on the Open Data Portal (govdata.de) which Andreas Witt mentioned in his presentation, Andrea Lösch wanted to know why the portal only contains a limited number of translations (or text data in general). According to Mr. Witt, there are two reasons for this: 1) The portal had been released only recently, 2) The national law prescribing that public sector information must be made available (ideally via such portals) has not been issued yet. He added that time will tell if the portal will actually be useful. The most important aspect would be that thanks to the PSI directive, data will need to be made available and can no longer be hindered by legal and technical barriers.
- **What are the major barriers that prevent you from sharing data and how can they be overcome?** To conclude, workshop moderator Andrea Lösch asked the three speakers what they would consider the major barriers which prevent public administrations and/or SMEs from data sharing. According to Ralf Lemster who presented the perspective of SMEs, companies still think of their data as business secrets and want to protect any kind of information which may bring competitive advantage. Andreas Witt added that legal issues also remain to be a major issue. He explained that even though public sector information will need to be made available according to the PSI Directive, most of the information does not fall under this category and will therefore not be shared. In his opinion, it would be important to add more exceptions when it comes to text mining, creating language models, etc. Language models, for example, should be usable in non-academic settings and data and text mining should be possible for copyright-protected data, too. Speaking from the perspective of a translator, Alexandra Soska mentioned confidentiality as one of the major obstacles. She added that on the one hand, data management needs to be improved, making it easier to

identify the data that can be shared. On the other hand, language services spend 90% of their work time translating and not with data preparation. To her, finding the time for this additional task is one of the key challenges.

Looking at the availability of language data in the organisations of the workshop participants, it is interesting to note that the vast majority of workshop participants (90%) confirmed to hold language resources (see Figure 14 below).

Figure 14: Availability of language data in the participants' organisations

Unfortunately, the polls also revealed that less than one quarter (24%) of the respondents' organisations had a data management plan in place (see Figure 15 below).

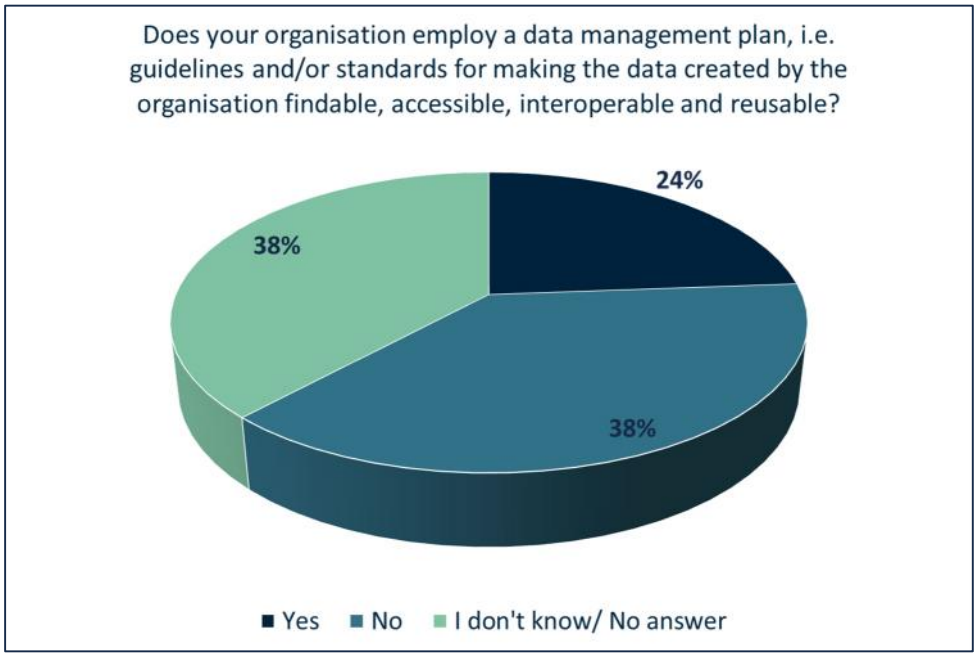


Figure 15: Availability of a data management plan in the participants' organisations

As such, it is not surprising that among the German workshop participants, legal issues were considered as the major factor that may prevent the sharing of language data (19) followed by inadequate practices for the management of language data (12).

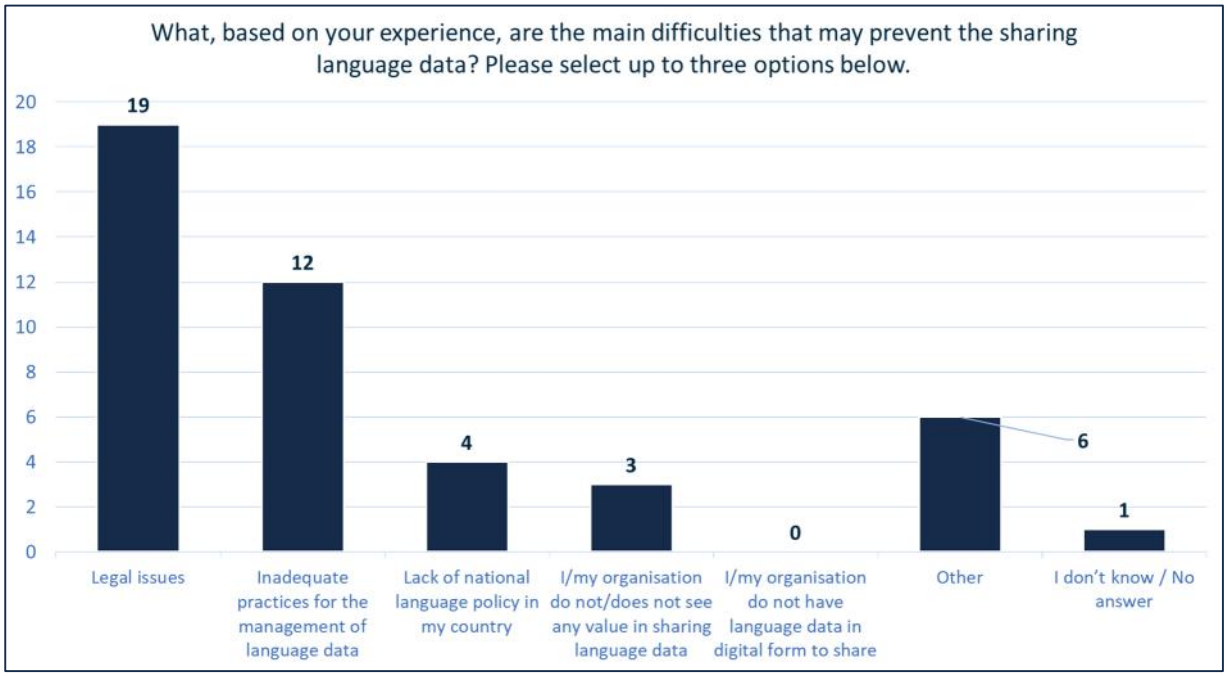


Figure 16: Main difficulties preventing the sharing of language data

3.7 Take-home message and conclusions

The workshop was concluded with a short wrap-up session by Andrea Lösch (DFKI) who stressed that language technologies empower people: they can provide considerable support in our daily work, they can help exchange information across borders and languages, they allow us to communicate, find and analyse information more effectively, and last but not least, they may also carry out routine tasks for us. As such, they enrich and facilitate our daily work and private lives.

However, in order to provide useful support to us (in our particular language and our particular situation/domain), language technologies need data. Not any data, but data from the particular usage scenario. Workshop participants were encouraged to share their data through the ELRC-SHARE Repository and hence to support the further development of CEF eTranslation. Andrea Lösch also hinted on the corresponding Technical and Legal Helpdesk that can support potential data donors, e.g. in finding the right licenses, cleaning the data etc. Also, the Catalogue of Services was mentioned one more time as the place where language technologies from Europe for Europe could be found. Last but not least, participants were encouraged to distribute and participate in the feedback evaluation survey.

3.8 Demo Session: 5th National ELG Workshop

The 5th National ELG Workshop took place in the afternoon, directly following the lunch break and using the same Zoom room as the 3rd ELRC Workshop. Demos included in particular

- an illustration of the ELG platform (https://www.european-language-grid.eu/),
- related pilot projects in Germany (https://www.european-language-grid.eu/open-calls/open-call-1 and https://www.european-language-grid.eu/open-calls/open-call-2/), and
- the ELE project (https://libereurope.eu/project/european-language-equality-ele/).

More detailed information is available through the website of the European Language Grid (https://www.european-language-grid.eu/5th-national-elg-workshop-germany/) and through the corresponding report.

4 Major findings and implications for the German Country Profile

This section summarises the major findings from the German workshop and their implications for the Country Profile of Germany with regard to MT/LT uptake and language data sharing.

4.1 What is the status with regard to MT take-up and acceptance in German public services / SMEs?

97% of all workshop participants had used MT, about half of them both eTranslation and other free systems. Most interestingly, almost 50% were either fully satisfied or very satisfied with the quality of MT for German which shows a great improvement in terms of translation quality compared to earlier years and earlier workshops. As such, it can be concluded that in the past 5 years, MT has become a core technology in and for public administrations and SMEs – a finding that is supported by the recent European Language Industry Survey (ELIS 2021). There was great interest among the workshop participants in the evaluation of MT systems for the German Federal Authorities. Key questions were about the process of the investigation and how the translation quality was assessed. As could be shown, the involvement of translators from various translation services in the review process is vital for the success of such a large-scale investigation: There must be a dedicated working group with staff members from key institutions on the one hand in order to gather and prepare relevant materials and frameworks. On the other hand, translators working as independent evaluators for assessing the quality of the MT outputs are needed. Last but not least, the development of a dedicated framework for quality assessment that would fit the needs of the participating institutions is vital.

4.2 To what extent are other LT relevant to public services / SMEs in Germany?

As became clear from the live polls, Information search and retrieval are the second most important LT according to the workshop participants (MT remains the most widely used technology by German public services and SMEs). In the particular context of the Federal Statistical Office, classification, however, was the key technology. One major issue was for instance how to use classification best in the context of public services and whether to rely on free text input or rather provide some categories upfront for selection. This, however, depends on the situation and there was consensus that even completely different approaches (e.g. sending photos of receipts via mobile phone) might be used for future surveys. Virtual assistants/chatbots, speech recognition and text-to-speech solutions were not yet widely used among the audience. One reason could be the low technological maturity of these technologies, e.g. for German. As shown by the live polls, more than two thirds (69%) of the workshop participants prefer to use LT in their own language (i.e. German). Less than one quarter use them in English. However, more than two thirds were also not really satisfied with the quality and reliability of the LT used in their language. Even more, only 8% of the workshop participants were actually satisfied with the digital readiness of German public services and SMEs. They see the main role of SMEs and public services as data stewards and users/service providers of LT, rather than as a regulator. With regard to solutions side, the Catalogue of Services shows that more than 130 LT solution providers are actually based in Germany. This fact was also underlined when looking at the start-up scene in Germany: The majority of AI start-ups in 2021 in Germany could be assigned to the domain of LT companies, being either directly from the computer linguistics domain or supporting customer service and/or marketing functions of businesses. As such, LT other than machine translation are on the rise, and it can be expected that within the next 5 years, take-up will also increase in Germany.

4.3 What is the situation with regard to the sharing of language resources in Germany?

The situation in Germany with regard to language data creation, management and sharing practices has not significantly changed since the publication of the Country Profile as part of the ELRC White Paper end of 2019. Looking at the availability of language data in the organisations of the workshop participants, it is interesting to note that the vast majority of workshop participants (90%) confirmed to hold language resources. Unfortunately, the polls also revealed that less than one quarter (24%) of the respondents' organisations had a data management plan in place. The practices for the creation of multilingual parallel data (in particular translations) remain fragmented in Germany. There is no formal translation procedure common to all public administrations on the federal level, let alone all public services across the federal states. Some (especially the larger) authorities maintain in-house translation departments, others outsource their translations. In the majority of cases, the outputs are not managed according to a data management plan, translation memories of outsourced translations are not requested back, and, in some cases, translations are only available in word .doc format. Legal issues as well as the absence of data management plans (or even guidelines governing the sharing of language data) in organisations remain the main barriers hindering the sharing of language data in Germany.

Two important developments, however, may help overcome this situation in the future and pave the way for an increased sharing of language data. The first major improvement may be seen in the upcoming Data Usage Act in Germany which aims to significantly extend the provision of open administrative data on the federal level and hence to facilitate the re-use of such data (also explicitly allowing commercial usage of such data). Based on the new Data Usage Act, **public sector information shall be freely usable** and available in machine-readable format, in all available languages, free of charge, with relevant meta data, via the National Open Data Portal (www.GovData.de), and, where necessary, with an open license (e.g. CC-BY, DL-DE/zero, DL-DE/namensnennung). Especially the organised collection (including standardised meta data) and licensing are expected to contribute to the improved organisation, sharing and hence availability of language data. The second important development is the increased take-up of MT (see above) – and hence the pressure, to prepare and share language data. So more and more organisations will need to internally review their processes for language data sharing and above all, address one key question: How to make relevant language data retrievable? This central question covers different aspects of the organisation, including:

- the translation process - How can it be adapted in a way to enable the re-use of language data?
- the classification and meta data descriptions - How can metadata be usefully extended or modified to facilitate the identification/export of sharable data? How can contents be best classified?

With regard to the actual cleaning of data for MT training, organisations can outsource this process using a corresponding confidentiality agreement (in case of a one-time affair) or build corresponding human resources in-house in case this task will frequently arise.

In the private sector, the sharing of language data appears to be slightly more difficult given that such data are often considered as trade secrets and/or as information providing a particular competitive advantage which is why there is a considerable reluctance against the open sharing of this data. Even when sharing language data in a definable circle, the issue of appropriate remuneration is always present.

While it was stressed several times that the availability of sufficient parallel data (e.g. translations) are still key for the development of MT systems in SMEs and public administrations, it could also be shown during the workshop that in cases where sufficient parallel data may not be available, scientific advances may soon provide viable MT solutions in the form of unsupervised MT or even self-supervised MT.

5 Workshop Participants

Overall, almost 140 potential attendees were approached via email and invited to participate in the workshop, including participants of previous workshops and conferences or contacts from ELRC networking activities. In addition, the workshop was disseminated via the ELRC social media channels. In order to maximise the reach and participant rate of the event, the German ELRC National Anchor Points as well as the speakers promoted the event through their channels and networks. A corresponding press release was also published on the DFKI website.

In total, 64 people registered for the 3rd German ELRC Workshop and 54 joined the event on 20 April 2021. This corresponds to a drop-out rate of less than 20%. Separating the attendees by sector, it could be noted that the majority was linked to the public sector (25), followed by Research/Academia (14) and "Other" (11). Examples of people from this category include e.g. freelancers, members of the German Association of Interpreters and Translators (BDÜ) or the German Institute for Standardisation (DIN e.V.). The exact distribution is provided in Figure 17 below.

	Number of participants[3]	Percentage
Public sector	25	46
Research / Academia	14	26
SMEs	7	13
EC and ELRC consortium	6	11
Other	11	20

Figure 17: Participants Distribution by Sector

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