Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3443–3448
Marseille, 11–16 May 2020

Language Data Sharing in European Public Services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures

Lilli Smal[1], Andrea Lösch[1], Josef van Genabith[1], Maria Giagkou[2], Thierry Declerck[1], Stephan Busemann[1]
[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH,
[2]ATHENA Research and Innovation Center in Information, Communication and Knowledge Technologies
[1]Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany, [2]Artemidos 6 & Epidavrou, 15125 Athens, Greece
[1]{lilli.smal, andrea.loesch, josef.van_genabith, thierry.declerck, stephan.busemann}@dfki.de,
[2]mgiagkou@athenarc.gr

Abstract

Data is key in training modern language technologies. In this paper, we summarise the findings of the first pan-European study on obstacles to sharing language data across 29 EU Member States and CEF-affiliated countries carried out under the ELRC White Paper action on *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters*. We present the methodology of the study, the obstacles identified and report on recommendations on how to overcome those. The obstacles are classified into (1) lack of appreciation of the value of language data, (2) structural challenges, (3) disposition towards CAT tools and lack of digital skills, (4) inadequate language data management practices, (5) limited access to outsourced translations, and (6) legal concerns. Recommendations are grouped into addressing the European/national policy level, and the organisational/institutional level.

Keywords: ELRC, LR collection, language data sharing, language barriers, language technology, language equality, multilingualism, Europe

1. Introduction

The European Union has 24 official languages. The European Parliament and the European Commission state that multilingualism is both “an asset and a shared commitment” (European Commission, 2008; European Parliament, 2009). Languages are not only a means of conveying information – they are constitutive parts of our cultures and identities, and thus one of the key pillars of Europe’s rich cultural heritage. At the same time, languages can also create barriers, and overcoming language barriers is one of the main challenges European citizens, public servants and businesses are facing in cross-language communication and trade. The European Commission Directorate-General for Translation has the largest translation service in the world translating EU legislation into all 24 official EU languages. In 2018, the Commission’s Directorate General for Translation translated some 2,255,000 pages (European Commission, 2019, p. 8). To meet the high demand for translation, the Commission developed its own machine translation (MT) system called eTranslation[1] (formerly known as MT@EC), which was launched in 2013. In 2014, the European Commission started the Connecting Europe Facility (CEF) Programme. eTranslation aims to facilitate multilingual communication and the exchange of documents and other linguistic content in Europe between national public administrations on the one hand and between these administrations and EU and CEF-affiliated country citizens and businesses (European Commission, 2018) on the other hand. However, in order to make eTranslation work in the various domains and language pairs relevant to public services and administrations across Europe, corresponding language resources were – and still are – needed.

In order to address this need, the European Language Resource Coordination (ELRC) was launched in April 2015 to collect language data relevant for public services in EU Member and CEF-affiliated States (Lösch et al., 2018). While ELRC has been and is very successful having delivered more than 1.400 language resources and tools covering all EU official languages, it also uncovered a number of important obstacles impeding the collection of language data produced by public services. In order to address these obstacles and to make data collection sustainable, ELRC carried out a Europe-wide study. The results of the investigations are presented in detail in the *ELRC White Paper, Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters*, including Country Profiles for each country (ELRC, 2019). To our knowledge, this is the first pan-European study covering 29 EU Member States and CEF-affiliated countries on obstacles to language data sharing and recommendations on how to overcome these.

Section 2 of the paper provides further background to ELRC, some comparison with other data collection initiatives, as well as details on the methodology of the investigation. Section 3 provides the main findings of the study, detailing in particular the obstacles found that currently prevent the sustainable sharing of language data, specifically in the context of public services in Europe. Section 4 presents recommendations on how to address the obstacles, including corresponding suggestions on the European/national policy level and on the organisational or institutional level. Section 5 summarises and concludes.

[1] For more details on CEF eTranslation, see https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

2. Background and Methodology

ELRC's main purpose is to identify and collect language data from public administrations in all countries participating in the CEF programme to further improve, develop, and expand eTranslation.

In the framework of the Open Data Directive[2], "Member States should ensure that documents, which are held by public sector bodies and accessible according to national access regimes, are re-usable for commercial or non-commercial purposes." (Lösch et al. 2018). Since 2015, ELRC has successfully collected language resources (LRs) and made them available in the ELRC-SHARE Repository (Piperidis et al., 2018).[3] In addition, ELRC-SHARE hosts several language resources that were contributed by other CEF-funded eTranslation projects. As of February 2020, more than 1400 unique LRs are available through ELRC-SHARE of which almost 80% are bi- or multilingual resources. More than 80% of the collected LRs have been made available under permissive licenses and are hence re-usable by the LT community beyond eTranslation. They can be downloaded directly from the ELRC-SHARE repository. Furthermore, Open Data sets are also available through the EU Open Data Portal.[4]

The joint effort by the ELRC consortium and its Language Resource Board (LRB) to identify and collect LRs in European public administrations, unveiled a number of substantial challenges that hinder the continuous flow of language data. Therefore, it is considered essential, as indicated by Lösch et al. (2018), that future work should focus on "further development of stakeholder involvement and data pipeline sustainability". To this end, ELRC initiated the so-called Country Profiles with the aim to investigate, identify and describe the multilingual data creation and sharing infrastructures in each of the participating countries, as well as the key stakeholders, main challenges and concrete country-specific action plans. Based on the findings, ELRC issued a series of general recommendations on how language data sharing infrastructures can be improved (ELRC, 2019).

2.1 Initiatives Collecting Language Data

ELRC-SHARE has the specificity that the language data are provided by governmental agencies, administrations and other (public) institutions that possess a collection of mono- or multilingual data that are relevant for overcoming the language barriers between European institutions and between European institutions and citizens. This type of multilingual data is very important for the further development of the CEF Automated Translation platform as well as of other MT systems employed by public administrations in the Member States. The role of the ELRC consortium is to elicit the submission of such mono- or multilingual data by the organisation of awareness raising workshops, in close cooperation with the National Anchor Points (NAPs)[5], and by providing on-site assistance for helping the language data providers prepare and deliver their data. Before uploading the obtained language data into the ELRC-SHARE repository, the ELRC consortium has to process it (e.g. data cleansing, format conversion, alignment, etc., where relevant). So, for example, parallel data should be encoded in a specific TMX[6] format. Terminological data must be transformed in TBX[7].

In addition to the technical support, the ELRC consortium is offering legal guidance to language data providers, as in many cases their data was not foreseen to be delivered to a repository or to be used for training MT systems. As a consequence, information about the type of licenses that can best suit the purpose of re-using the data is needed .

While from the outside it can seem that ELRC shares similarities with other initiatives like CLARIN[8], there are several differences. First, CLARIN is a research infrastructure which aims to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences (Hinrichs & Krauwer, 2014). In this sense, CLARIN's scope is broader than that of ELRC, which focuses mainly on language resources and technologies for multilinguality. While CLARIN addresses mainly the research community, ELRC targets data owners from the public sector. This difference with regard to the targeted audience, entails different operations for data identification, processing, sharing and re-purposing. Since ELRC is mainly dealing with language data that reside in public organisations, it is heavily engaged in efforts to "unlock" these data through raising awareness of their value and potential for language technology and for machine translation in particular. ELRC is then taking care of the data for making it ready for re-use. In most cases, the result of this transformation, the processed data, can be re-used by researchers or the industry. In the case of CLARIN it is the language technologist, who is pushing her/his data into the portal, stating its relevance for potential applications.

However, CLARIN (and similar infrastructures for LRs, or LR distribution agencies like ELRA[9]) and ELRC can cooperate in the field of metadata, including the information on legal aspects. This is also true for the recently started European Language Grid (ELG) infrastructure project (Rehm et al., 2020)[10]. This initiative is aiming primarily at offering a platform for multilingual,

[2] The Open Data Directive (2019/1024/EU) replaced the Public Sector Information Directive (2003/98/EC), which was amended by the Directive (2013/37/EU).

[3] https://www.elrc-share.eu/

[4] https://data.europa.eu/euodp/en/home

[5] see http://www.lr-coordination.eu/anchor-points

[6] TMX (Translation Memory eXchange) is an XML specification for the exchange of translation memories. See https://sourceforge.net/projects/tmx/

[7] TBX (TermBase eXchange) is an international standard (ISO 30042:2019) for the representation of terminological data. It also supports the representation of multilingual terms associated with a concept in the terminology base.

[8] https://www.clarin.eu/

[9] http://www.elra.info/en/

[10] https://www.european-language-grid.eu/

cross-lingual and monolingual technologies supporting the emergence of the multilingual Digital Single Market. The focus is also not on MT for European institutions and public administrations in the European Union and CEF-affiliated countries. Language technology vendors are the main players for uploading their resources and tools to the ELG platform. However, ELG is also involved in the specification of a large set of metadata, further building on the META-Share initiative[11], which can be re-used by ELRC-SHARE.

2.2 Methodology

The ELRC Language Resource Board is a unique EU-wide network. It comprises a Technology and a Public Services National Anchor Point (T-NAP and P-NAP) of each country participating in ELRC.[12] The T-NAPs are highly regarded experts in the field of language technology, Artificial Intelligence, translation or language studies for their respective languages. The P-NAPs represent the national governments or public administrations and have extensive expertise related to translation or digitalisation. The NAP tandems are jointly driving change in their respective countries.

Since 2015, ELRC has conducted 57 local country-specific workshops in all CEF-affiliated states. As part of these workshops, national stakeholders, translation practices and needs as well as translation data and their availability were identified and discussed. The workshop reports are available on the ELRC website.[13] Moreover, since November 2017, the language data creation and sharing infrastructures as well as the main challenges for sharing language data were regularly investigated and analysed as part of the bi-annual meetings of the Language Resource Board and the regular regional Q&A online sessions.

The aforementioned instruments (LRB, Country Workshops, dedicated Q&A Online Sessions) were used to inform this investigation. Apart from feedback gathered at the ELRC workshops, and further consultations with the NAPs in face-to-face and virtual meetings, a structured questionnaire was designed to gather and organise country-specific feedback. The questionnaire comprised open-ended questions that sought to investigate the current situation in terms of (multilingual) data creation and sharing. These included for instance the following: “How are translation needs in the public sector met?”; “Is there any exchange of data on a national level?”; “Is public procurement data openly available?” etc. In addition to the open-ended questions, through a series of semi-closed questions, the respondents were requested to define and rank (in terms of relevance for their country) the issues that they considered as main obstacles or challenges for sustainable language data sharing, as well as the actions that should be undertaken in order to address them.

The questionnaire was answered by all NAP tandems, resulting in a filled-in questionnaire per country. Based on first findings, the corresponding country profiles were set up by the ELRC consortium and then assessed and extended by the NAPs.

The recommendations are based on the collective expertise of the ELRC NAPs and the consortium as well as on best practices.

3. Main Findings

The translation practices as well as the language data sharing infrastructures and processes vary greatly across the CEF countries. Yet, six main challenges that hinder the sharing of language data by public services in most countries were identified. These are:

1. Lack of appreciation of the value of language data.
2. Structural challenges.
3. Disposition towards CAT tools and lack of digital skills.
4. Inadequate language data management practices.
5. Limited access to outsourced translations.
6. Legal concerns.

3.1 Undervalued Perception of Language Data

The fact that translation data (or more broadly: language data) is not regarded as a high-value asset can be considered as the most important obstacle for creating sustainable language data sharing infrastructures in European public services. 24 out of 29 countries indicated that the seemingly low value of language data is the most or second most important challenge (ranked first or second) that needs to be addressed and overcome. No country considers this challenge as least important (ranked 5th or 6th). This is illustrated in Figure 1 below.

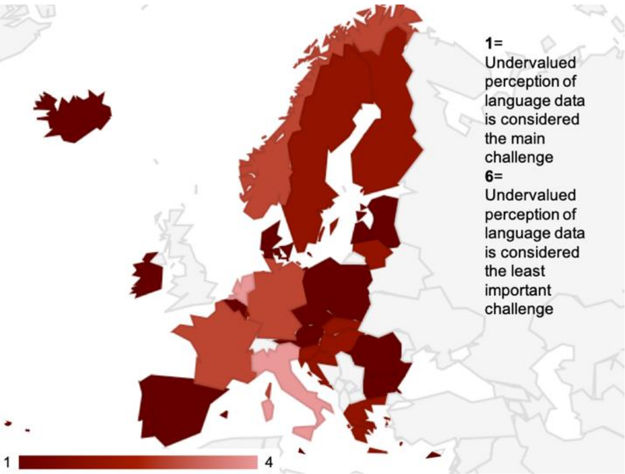

1=
Undervalued perception of language data is considered the main challenge
6=
Undervalued perception of language data is considered the least important challenge
1
4

Figure 1: Undervaluation of language data as the main challenge across Europe.

In addition, there is little awareness of the concept of language resources and the added value of shared language

[11] http://www.meta-share.org/

[12] http://lr-coordination.eu/anchor-points

[13] ELRC Events: http://lr-coordination.eu/events

data. Last but not least, the disregard of language data as an asset also leads to a number of subsequent issues.

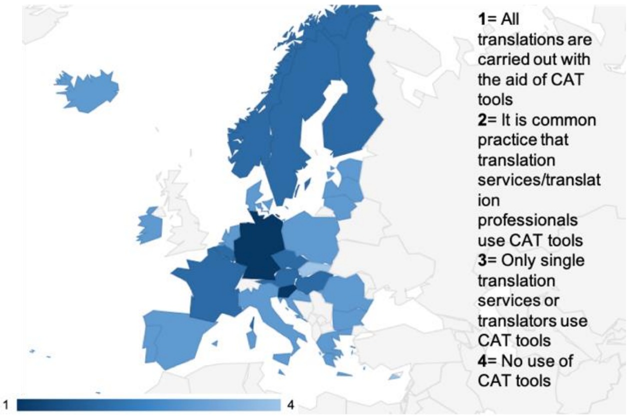
3.2 Structural Challenges

In most countries, there is no public administration, ministry or body which is responsible for the collection and curation of translation or language data. Therefore, even though a considerable volume of language data is being produced by the public sector every day, public administrations cannot leverage on it, as the data is not managed, procured or curated in an appropriate way. In addition, there are frequent changes in such positions. Hence, even when single translation services or translators are willing to share their translations, often they cannot identify the responsible person or body to authorize the sharing of the data.

With regard to policy level, only 19 of the 29 countries actually have a language policy and only eight countries (Estonia, Iceland, Ireland, Latvia, Norway, Slovenia, Spain, Sweden) have a dedicated financial plan for the development of language technologies. This also contributes to the fact that in many countries there is no appropriate infrastructure for sharing language data yet.

3.3 Disposition Towards Language Technologies and Lack of Digital Skills

Another issue faced by a large number of countries is the lack of digital skills with regard to using computer-assisted translation (CAT) tools, but also with respect to the process of preparing language data for sharing and the actual sharing process. The translation processes vary greatly across the countries and also across the administrative levels. Figure 2 indicates the use of CAT tools in the translation process at the federal/national level in the respective countries.



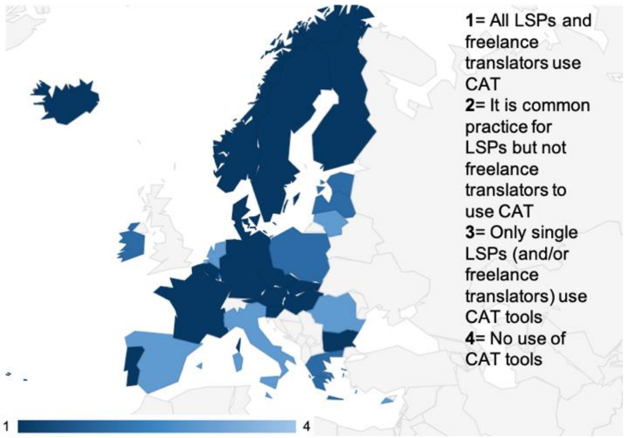
1= All translations are carried out with the aid of CAT tools
2= It is common practice that translation services/translation professionals use CAT tools
3= Only single translation services or translators use CAT tools
4= No use of CAT tools
1
4

Figure 2: Use of CAT tools by public administrations in Europe.

Only two countries (Germany and Slovenia) indicated that all translations in public administrations were carried out with the help of CAT tools. Nine countries claimed that it was common practice to use CAT tools in the translation process and 16 countries said that only single translation services or translators used CAT tools. Two countries indicated that CAT tools were not used at all (Cyprus and Slovakia).

In contrast, in 15 countries, it is standard practice for language service providers and freelancers to use CAT tools in their translation processes. Seven countries indicated that it was very common to use CAT tools and only seven countries indicated that only single LSPs used CAT tools.

Figure 3: Use of CAT tools by Language Service Providers in Europe.



1= All LSPs and freelance translators use CAT
2= It is common practice for LSPs but not freelance translators to use CAT
3= Only single LSPs (and/or freelance translators) use CAT tools
4= No use of CAT tools
1
4

According to ELRC's investigation, the limited use of CAT tools by translators in public administrations has a number of reasons. In some countries CAT tools are considered too expensive, in others translators are not trained to use them. In some cases, a general resistance to using CAT tools can be noted or even a combination of all three reasons.

The use of a MT system by translators in public administrations in Europe is even less frequent. Only eight out of 29 countries stated that one or several public administrations use a MT system. In no country, however, is it a common practice to officially integrate MT into the translation process apart from the (sometimes unauthorized) use of free online MT services.

3.4 Inadequate Language Data Management Practices

The above-mentioned challenges all feed into inadequate language data management practices, i.e. language data practices that do not allow for easy sharing of translations. Reasons for this include the following:

- Translators do not use Translation Memories (TMs) and therefore do not produce TMX files.
- Confidential texts or texts containing personal information are not proactively indicated or tagged as such in TMs, making it impossible to distinguish them from the sharable content contained within the same TM.
- The copyright of the translation belongs to the translator; no provisions are taken to either transfer the copyright or license the texts for further reuse.
- Translation Memories and other by-products of the translation are not transferred to the contracting body when translations are outsourced.

3.5 Limited Access to Outsourced Translations

Some public administrations rely heavily on LSPs and freelancers for translation services. Yet, as mentioned above, the TMs are not requested back. Five countries (Finland, Poland, Germany, Luxembourg and France) stated that TMs were requested back for most outsourced translations. All other countries either do not request back TMs at all (13 countries) or only some public administrations do so (11 countries).

3.6 Legal Concerns

In addition to the lack of licensing and transferring of Intellectual Property Rights to the contracting authority of outsourced translations, even for in-house translations, there is no rights management of translations, therefore complicating the sharing of these texts. The updated General Data Protection Regulation (GDPR)[14] also added to the resistance to share language data because of fear of infringement.

4. Main Recommendations

To overcome the identified challenges, a number of recommendations addressing the European and national policy level, as well as the organisational/institutional level were identified.

4.1 Recommendations at the Policy Level

The recommendations formulated by the ELRC consortium and the Language Resource Board include the following:

- One of the main reasons for limited sharing of language data within European public administrations is due to the fact that language data is not regarded as a high-value asset in general and is not referenced as a data category in the Open Data Directive (2019/1024/EU). To initiate long-term changes, it is vital to explicitly include language data in the Open Data Directive.
- In order to identify and quantify the value of language data for citizens, public administrations and businesses, a corresponding study should be commissioned by the European Commission to further raise awareness about the direct and indirect impact of language data.
- Language data management and procurement should be included in the national digital strategy, strategy for Artificial Intelligence or Open Data strategy.
- It is vital that one public administration or public body is mandated to coordinate all matters related to multilingualism, translation and the development of language technologies. It is also essential to develop long-term funding schemes.
- The inclusion of obligatory language data management plans in all relevant national funding policies and calls for proposals will not only increase the amount of available language data but it will also sharpen the general understanding of the need for Open Data.
- The conduct of national surveys assessing the translation practices in public administrations on all administrative levels is deemed necessary to find tailored solutions to improve translation practices in each country and at all administrative levels. Most importantly, the requirements and feasibility for a sustainable platform or mechanism for sharing language data on the national level should be investigated more closely.

4.2 Recommendations at the Institutional and Process-Level

Change can only be driven on both the policy level and the operational level of the institutions. As such, the following recommendations addressing the organisational or institutional level are necessary and complementary to the above-mentioned policy-level suggestions. They are based on collective expertise of the Language Resource Board and the ELRC consortium as well as best practices.

- Translation and data management practices need to be adapted and improved in order to allow for easy/easier language data sharing in the future, including appropriate licensing of translations. It is vital that confidential information is separated from public information at an early stage in the translation process.
- Investments in human capital must be made, including in particular the provision of technical and legal training for translators and translation managers.
- Investments in IT infrastructures, equipment and tools are necessary, in particular the provision of CAT tools, MT, data anonymization methods and similar.

5. Conclusions

The European Commission has made large investments in the past to support and foster multilingualism in Europe. However, only through further investments in language technologies will all official EU languages be able to contribute to and create a truly multilingual Digital Single Market.

Language data is crucial to develop language technologies for all languages regardless of the number of speakers. At the same time, public administrations in Europe create a large amount of bi- and multilingual data on a daily basis. Yet, a lot of this language data is not re-used and hence its potential remains untapped. To overcome the identified challenges for language data sharing, it is indispensable to raise the awareness of the high impact language data can have in "deliver[ing] sustainable economic and social benefits from a digital single market based on […] interoperable applications" (Kolodziejski, 2013). It is necessary to include language data as a high-value data category in the Open Data Directive which will then be transposed into national legislation.

Multilingualism including translation needs should be coordinated, regulated and streamlined at the Member State level as well as between Member States. Technical infrastructures that enable language data sharing need to be

[14] https://ec.europa.eu/info/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en

improved, created or extended depending on the current status of the Member State.

It is also vital that sufficient and affordable training opportunities are provided to all actors involved to acquire the necessary skills to either use CAT tools in the translation process, to share language data or to procure translations in the most efficient way, both in terms of cost- and time-efficiency.

6. Acknowledgments

Work underlying this paper has been created under the ELRC contract with the European Union (SMART 2015/1091 Lot2). The opinions expressed are those of the contractor only and do not represent the contracting authority's official position.

We thank all members of the ELRC consortium for their continuous and intense work that has contributed significantly to the results of the investigation.

We are indebted to the National Anchor Points for their most valuable voluntary contributions. Without their support, the White Paper would not have been successfully completed. Their names and affiliations are made available as given in (ELRC, 2019) through the link below.[15]

We also thank the anonymous reviewers for their very helpful comments on the submitted version of this paper.

7. Bibliographical References

ELRC (2019). ELRC White Paper. Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters. Available at http://lr-coordination.eu/infopoint.

European Commission (2008). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM(2008) 566 final.

European Commission (2018). eTranslation – Connecting Europe. Available at: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Media+library++Infographics?preview=/82773424/84413846/eTranslation_English.pdf [last accessed: 05.03.2020]

European Commission (2019). 2018 Annual Activity Report Directorate-General for Translation, Ares(2019)2432753 - 05/04/2019.

European Parliament (2009). European Parliament Resolution of 24 March 2009 on Multilingualism: an Asset for Europe and a Shared Commitment (2008/2225(INI)).

Hinrichs, E. & Krauwer, S. (2014): The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, 1525–31.

Kolodziejski, M. (2013) Digital Agenda and the Economic Development of European Regions; http://www.europarl.europa.eu/RegData/etudes/divers/join/2013/513984/IPOL-REGI_DV(2013)513984_EN.pdf

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri K., and van Genabith, J. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Piperidis, S., Labropoulou, P., Deligiannis, M., and Giagkou, M. (2018). Managing Public Sector Data for Multilingual Applications Development. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA)

Rehm, S., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J. Hamrlová, J., Kačena, L., et al.. European Language Grid: An Overview. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, 5 2020. European Language Resources Association (ELRA). Accepted for publication.

[15] Contributors: Available at http://lr-coordination.eu/infopoint