K•PLEX

KNOWLEDGE COMPLEXITY

Deliverable Title: Report on Multilingual Big Data and Language Technology

Deliverable Date: 31st March 2018

Version: V1.0

Project Acronym :	KPLEX
Project Title:	Knowledge Complexity
Funding Scheme:	H2020-ICT-2016-1
Grant Agreement number:	732340
Project Coordinator:	Dr. Jennifer Edmond (edmondj@tcd.ie)
Project Management Contact:	Michelle Doran (doranm1@tcd.ie)
Project Start Date:	01 January 2017
Project End Date:	31 March 2018
WP No.:	5
WP Leaders:	Rihards Kalnins and Dr. Andrejs Vasiļjevs, TILDE

Authors and Contributors (Name and email address):	Rihards Kalnins (Rihards.kalnins@tilde.com) Dr. Andrejs Vasiļjevs (andrejs@tilde.com) Dr. Mārcis Pinnis (marcis.pinnis@tilde.lv)
Dissemination Level:	PU
Nature of Deliverable:	R = report

Contents

Objectives of WP5

The central objective of WP5 was to conduct research relevant to the topic of language technology and linguistic data, as well as to interact with experts in the field of language technology and language resources (including but not limited to translators) to understand the gaps between their practices and current technological norms.

To reach these objectives, WP researchers explored the state of knowledge and practice regarding the representation of language as data. In completing this task, researchers analyzed the current situation in respect to coverage for language resources, paying particularly attention to the current state of availability, coverage, and development of language resources and tools for each EU language.

This included an overview of studies and research reports that examined:

- Coverage and availability of language resources (Open Data) for European languages and various domains
- Coverage and availability of publicly available linguistic tools (tokenizers, parsers, etc.) for European languages
- Language resources used to built publicly available machine translation engines (CEF Automated Translation platform)

Researchers also conducted an analysis of policy documents such as the Strategic Research Agendas (SRIA) for the Language Technology Community and the Big Data Value Association (BDVA), as well as a review of past and current language technology (LT) industry strategy documents.

This analysis helped KPLEX researchers to understand what was missing from these previous studies, and therefore to formulate comprehensive surveys that would compile more information on the current state of multilingual data in the language technology industry.

In order to compile more information, researchers conducted two in-depth surveys with LT industry experts – both from the global language technology community (including localization professionals and language technology providers), as well as researchers and specialists in the field of language resources and processing.

Taken together, these analysis and surveys helped KPLEX researchers to reach the overarching objective of WP5 – namely, to formulate clear multilingual policy recommendations for the European Commission, allowing policymakers to draft more comprehensive ICT work packages in the future that bridge the current gaps in language technology coverage.

In so doing, the objective of WP5 was to ensure that language technology will become more robust and provide a more nuanced view of Big Data and multilingual content, helping to open up multilingual information from the widest possible range of European data sources.

Introduction

In order to transform culture into data, its elements have to be classified, divided, and filed into taxonomies and ontologies. This process of "datafication" robs them of their polysemy, or at least reduces it. Datafication of culture can be analyzed from various perspectives – e.g., datafication of cultural practices, personal interactions, religious practices, artistic production, and other phenomena.

In WP5 of KPLEX, researchers focused on the datafication of language – the transformation of the ambiguous, polysemic, conflicting and contradictory phenomenon of language into data. Language data is by definition unstructured text, which makes up a sizeable (but by no means dominant) portion of so-called Big Data landscape.

The language technology (LT) industry serves as an ideal test case for examining issues surrounding the datafication of language and the availability of language data (as well as gaps in coverage), as well as the impact of language data on technology, infrastructure, and employment. LT solutions are developed with language data as input material, therefore data issues—such as errors, noise, and inconsistencies in coverage—have a crucial impact on the quality of services.

Input data and language coverage issues become even more acute when neural networks are utilized in development. AI-based solutions like Neural Machine Translation (MT) are more sensitive to mistakes in input data, often treating them as linguistic phenomena. These mistakes are exacerbated by data scarcity and data inequality, particularly for smaller languages and overlooked domains.

The report attempts to illuminate these language data issues in LT, exploring the consequences in terms of technology, infrastructure, and employment. By exploring LT as a test case, the report intends to show how data inequality will potentially become a major theme in Big Data.

Furthermore, once it has been examined in the context of LT and Big Data, the overarching political consequences of data inequality will certainly become apparent, helping to inform possibilities for policy decisions on the part of EU institutions.

Findings of previous LT reports

To understand the impact of data on LT, KPLEX researchers analyzed data availability for EU languages, including large-scale corpora, multilingual open data, and resources available for the European Commission's eTranslation platform. The researchers also analyzed the effects of data inconsistencies on Neural MT. The study was supported by an analysis of several crucial policy documents for LT in Europe today:

- *Language Technologies for Multilingual Europe – Towards a Human Language* Project, prepared by the Cracking the Language Barrier federation;
- *Strategic Research Agenda for Multilingual Europe*, presented by the META Technology Council in 2013;
- *Language Equality in the Digital Age*, a Science and Technology Options Assessment prepared for the European Parliamentary Research Service.

Data inequality

These analyses uncovered several key findings. The first is that data inequality is a growing problem for LT. Some languages have large amounts of data, while others have little. This has a direct effect on the quality of LT solutions. Data inequality also applies to ownership: larger corporations and countries have access to large volumes of data, while smaller companies and nations are left behind. Data inequality has serious consequences, as access to data has become a social issue—in effect, data is making the big bigger and the small, smaller.

In its report on the LT landscape in Europe, entitled *Language Technologies for Multilingual Europe – Towards a Human Language Project*, researchers found that “many European languages other than English are heavily under-resourced, i.e., there are almost no resources or technologies available” (*Language Technologies for Multilingual Europe – Towards a Human Language Project* p.26). The report stresses the importance to “increase the size and improve the quality of available language resources by giving continuous support for management, preservation and evolution” (*ibid.*, p. 13).

These statements are based on a large survey conducted in June of 2017, which generated a total of 634 responses with a wide demographic reach from 52 countries. The report states that:

> “According to the survey, around 16% of the respondents see the biggest challenge that the European LT community is currently facing being the neglect of smaller languages. This is a severe threat, which is leading to a fragmented rather than a united and multilingual Europe. Around 90% state that they work with English in their research (not exclusively though) since they are often given little incentive to solely focus on smaller or minority languages. For instance, when it comes to publishing research results there is a strong bias towards incorporating results for English. Other challenges include the insufficient amount of data resources (approx. 13%)…” (*ibid.*, p. 16).

The findings of *Language Technologies for Multilingual Europe* were also backed up by the large-scale study *Strategic Research Agenda for Multilingual Europe*, presented by the META Technology Council in 2013. This study, in turn, was based on the META-NET White Paper Series *Europe’s Languages in the Digital Age*, which describes the current state of language technology support for 30 European languages published in the summer of 2012. White Papers were written for the following 30 European languages: Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, and Swedish.

The report found that "differences in technology support between the various languages and areas are dramatic and alarming" (*Strategic Research Agenda for Multilingual Europe*, p. 28). The Strategic Agenda for Multilingual Europe concludes:

> "While there are good quality software and resources available for a few larger languages and application areas, others, usually smaller or very small languages, have substantial gaps. Many languages lack even basic technologies for text analytics and essential language resources. Others have basic resources but the implementation of, for example, semantic methods is still far away. Therefore, a large-scale effort is needed to attain the ambitious goal of providing high-quality language technologies for all European languages. … Currently no language, not even English, has the technological support it deserves. Also, the number of badly supported and under-resourced languages is unacceptable if we do not want to give up the principles of solidarity and subsidiarity in Europe" (*ibid.*, p. 28)

Based on these findings, the report *Language Equality in the Digital Age*, prepared for the European Parliament, promoted the need for EU policy changes: "In order to bridge the technology gap, policies should focus on fostering technology development for European languages other than English, particularly the smaller ones or less-resourced ones, and also on language preservation through digital means" (*Language Equality in the Digital Age*, p. 38).

Data limitations

The reports also found that the data limitations in LT have a strong potential to affect users of AI applications. The use of AI, of course, has increased steadily in recent years, building on huge breakthroughs in AI research and application.

The report *Language Technologies for Multilingual Europe – Towards a Human Language Project* found that "We are currently witnessing a highly relevant commercial and industrial interest in Artificial Intelligence, Machine Learning and also Language Technology solutions, especially with regard to technologies based on neural networks. … Many experts in AI perceive cracking human language to be the next barrier and also goal for the next generation of AI technologies" (*Language Technologies for Multilingual Europe – Towards a Human Language Project*, p. 2).

The report also states that, based on these breakthroughs in AI, "AI is rapidly taking over many sectors that previously relied on human interaction. Banks are increasingly using chatbots to answer customer queries. For instance, it is suggested that Artificial intelligence will be the main way that banks interact with their customers within the next upcoming years" (ibid, p. 2).

Likewise, the report *Language Equality in the Digital Age* ascertained that AI-based virtual assistants have been gaining traction in customer service and other commercial applications.

> "Currently, sophisticated applications can accept widely varied and highly complex caller requests, enabling fully automated transactions or customer self-service including, but not limited to, accepting payments and entertainment ticketing, banking transactions or collecting personal information. In fact, nearly every

> industry segment (communications, financial services, government, healthcare, retail, tourism, etc.) has now implemented automated speech dialogue at some level, from simple call routers to fully automated self-service to even purchase/transaction applications" (*Language Equality in the Digital Age*, p. 22).

The study goes on to note that: "[w]ith respect to Text-to-Speech systems, improvements in this technology, combined with platforms requiring interactivity (such as mobile or gaming), are opening new opportunities for speaking applications. Some notable features are naturalistic voices in many more languages, which are used in education and gaming environments, and interactive access to the web" (ibid., p. 23).

Language Equality in the Digital Age also asserts that this trend shows no signs of slowing down: "Looking into the future, according to eminent voices such as Google CEO Sundar Pichai, we are moving from a mobile-first to an Artificial Intelligence (AI)-first world. Spoken LT are part of many AI scenarios that are quickly becoming mainstream…" (ibid., p. 23).

KPLEX Surveys of the EU Language Community

To expand the findings of the aforementioned reports, KPLEX sought to collect new data from two comprehensive surveys, covering a wide swath of the language technology community. By conducting these reports KPLEX attempted to hone in on several aspects that were missing from previous reports, namely, the use of language data processing and the corresponding level of skills in respondents' organizations.

Language Technology Survey

This survey was intended as a broad, far-reaching survey of many members of the wider language technology and localization services community in the EU.

Methodology

To gather respondents, researchers conceived and prepared a multi-question survey (questions in Annex 1) intended for a general audience of language specialists. Following an extensive validation process, the survey was sent out to 3648 individuals who are active in the language technology community and had signed an open letter to the European Commission calling for a multilingual Digital Single Market. The addresses represented members of the language industry in general – including localization specialists, language technologists, translators, researchers, and business managers.

Respondents

Over 500 individuals (approximately one in seven addressees) responded to the Language Technology Survey, taking up to ten minutes to complete the comprehensive survey. Respondents included individuals that work for the following organizations:

- Language technology providers
- Localization/translation services vendors
- Corporate localization departments
- Public sector institutions
- University/research institutions

- Self-employed

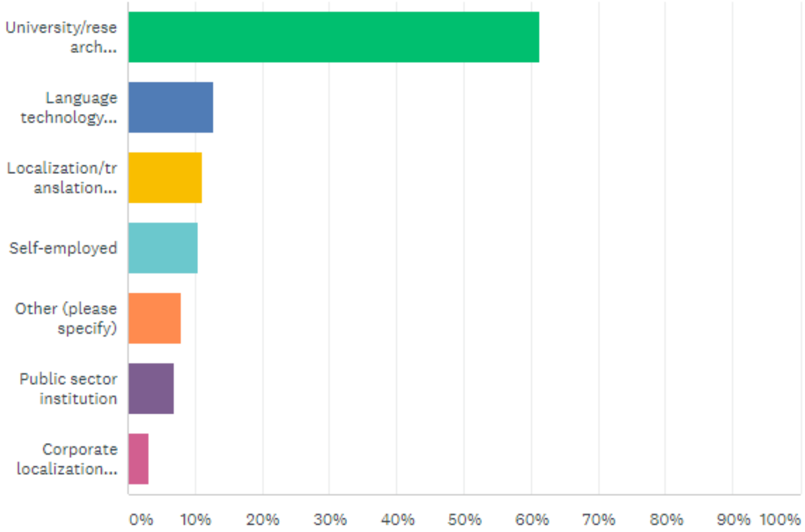

Figure 1: Respondents by organization

As seen in Figure 1, the majority of respondents were employed at universities and/or research institutions (61%). The second and third largest groups were, respectively, language technology providers (12%) and localization/translation services vendors (11%).

At these organizations, the respondents held the following positions:

- Researcher
- Teacher/professor
- Translator/linguist
- Manager
- Developer
- Editor
- Terminologist
- Student
- Government official

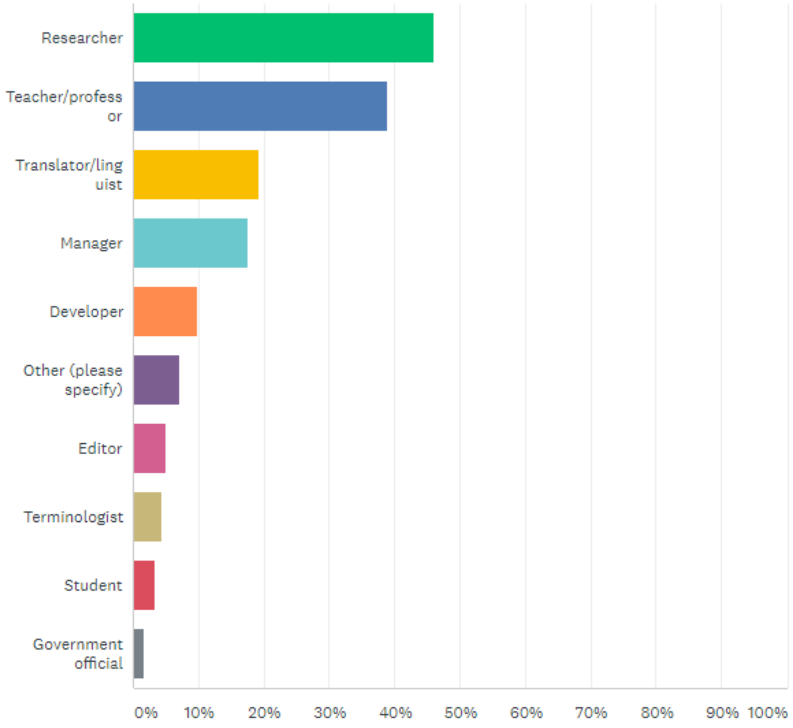

Figure 2: Respondents by occupation

As seen in Figure 2, the majority of respondents were occupied as researchers (46%) and teachers/professors (38%), followed by translators/linguists (19%), managers (17%), and developers (9%).

Key findings

The findings of the Language Technology Survey reveal the respondents' use of machine translation, providing a solid overview of how often these LT services are used, for which languages, and in what way.

The majority of respondents revealed that they use machine translation in the range of a few times per week (30%) to a couple times a month (39%).

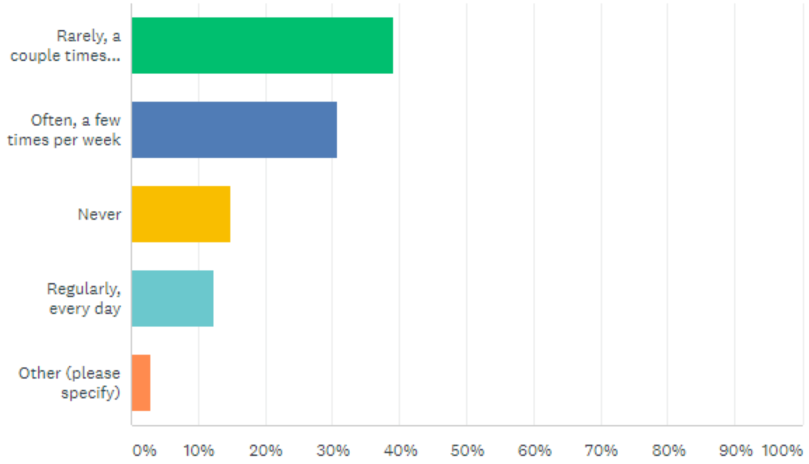

Figure 3: Frequency of machine translation usage

These figures validate the importance of machine translation in the everyday work of members of the language community, with nearly 70% using the tools on a nearly daily basis.

Perhaps not surprisingly, then, the majority of these users employed machine translation for the world’s larger languages, as seen in Figure 4. English was the language most frequently used for machine translation (66%) followed by the so-called “FIGS” languages: German (44%), French (38%), Spanish (31%), and Italian (20%). These languages – the most widely spoken in Europe – were followed by two of the world’s largest languages: Russian (19%) and Chinese (18%).

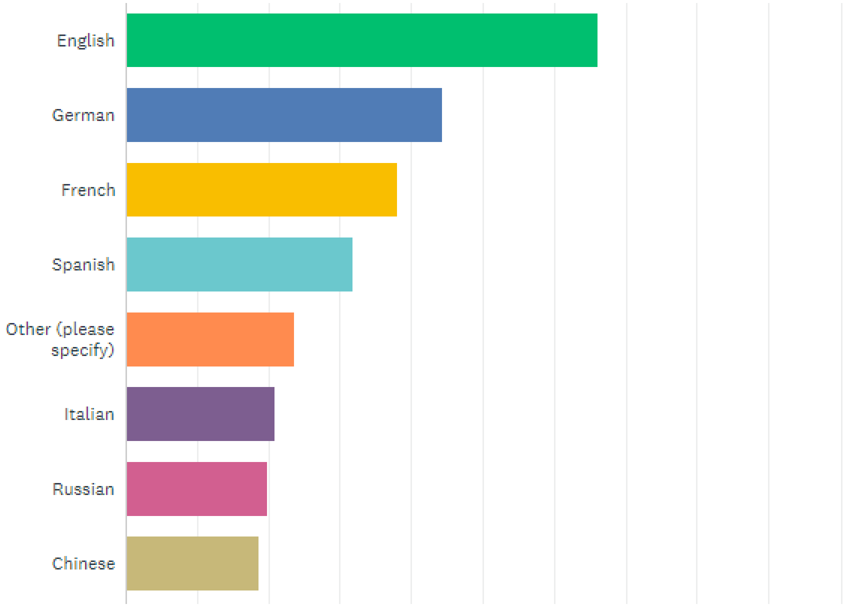

Figure 4: Languages used for machine translation

Though the usage of machine translation for these languages mirrors the predominance of these languages online,[1] it also falls in line with the availability of language resources for developing these systems, as we will find later in the Language Resources Survey.

The utilization of MT for larger languages, though positive in its reflection of the increased utilization of language technology for crossing language barriers online, also foretells another troubling trend: the underserving of machine translation for the world's smaller, more complex languages.

However, the usage of machine translation – though, as we see above, mostly for larger languages – does, for the most part, satisfy its users in terms of the applicability of MT for its purposes. The Language Technology Survey finds that, when asked to rate the extent to which MT is fit for purpose, i.e., serves their needs, the average response was satisfactory – slightly above 3 out of a possible score of 5.

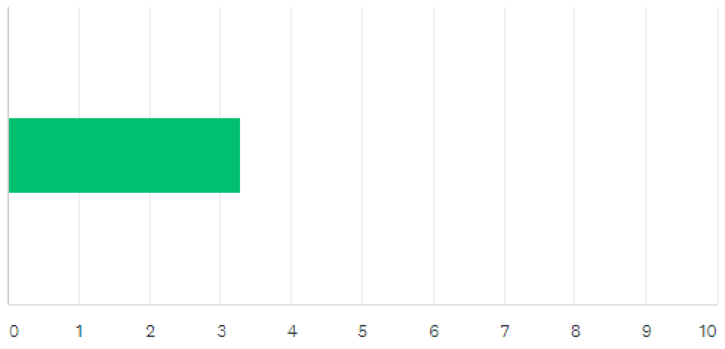

Figure 5: Average response to question: "On a scale of 1-5, to what extent is the MT you use "fit for purpose"?

When asked where they utilized machine translation – i.e,. which platforms and tools – respondents also reflected the importance of generic MT systems available online. Ass seen in Figure 6, the majority of respondents (87%) use MT in online translation sites like Google Translate. As Google Translate puts primary effort into developing high quality MT systems for the world's larger languages,[2] the correlations between usage of Google Translate, seen in Figure 5, and the usage of MT engines for larger languages, in Figure 4, should come as no surprise.

However, what is interesting about the findings from this survey question is the high number of respondents (21%) who reported using MT systems in computer-assisted-translation tools, otherwise known as CAT tools. These specialized programmes are intended for professional translators, helping to break down documents into individuals segments, or "strings," and enabling faster translation through the use of translation memory software and machine translation engines.

[1] http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/netlang_EN_pdfedition.pdf

[2] https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

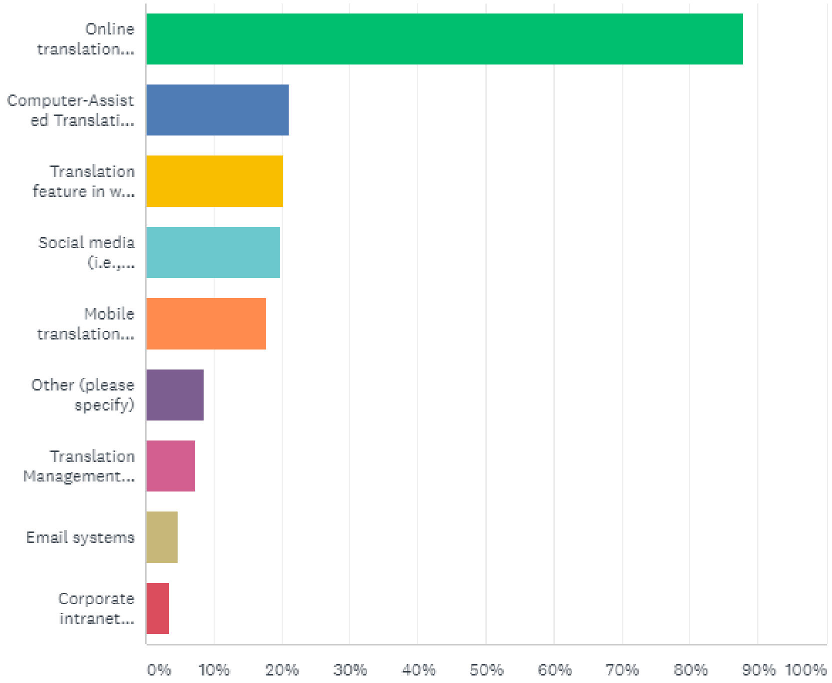

Figure 6: Platforms where users accessed MT systems

Machine translation has experienced a rocky adoption by professional translators in the localization industry, with many translators reluctant to utilize new tools that, at least according to popular mythology, seemingly "threaten" to replace human translators.[3] In recent years, however, the technology has been increasingly applied successfully to boost productivity in localization, dispelling these myths and helping to establish MT as a productivity tool for translators, like CAT tools themselves.[4]

The results of the KPLEX Language Technology Survey underscore this acceptance of MT in CAT tools, reflecting the embrace of the technology throughout the localization industry.

While this adoption of MT in localization is reflected in the survey, the majority of respondents continue to utilize MT to translate more "casual" texts, as seen in the next set of graphs. Over 85% utilize MT for translating "words and short texts"; 30% for full documents; and 30% for website translation (see Figure 7).

This trio of text types – short texts, documents, and websites – is directly in keeping with the functionality offered by online translation sites like Google Translate, which (as seen in Figure 6) is the most popular platform for utilizing machine translation. This further underscores the importance of such online translation sites for providing high quality MT for a wide range of languages.

[3] https://www.theguardian.com/education/2014/sep/19/tech-removing-language-barriers-jobs-lost-translation

[4] http://cracker-project.eu/csa-mt-report/

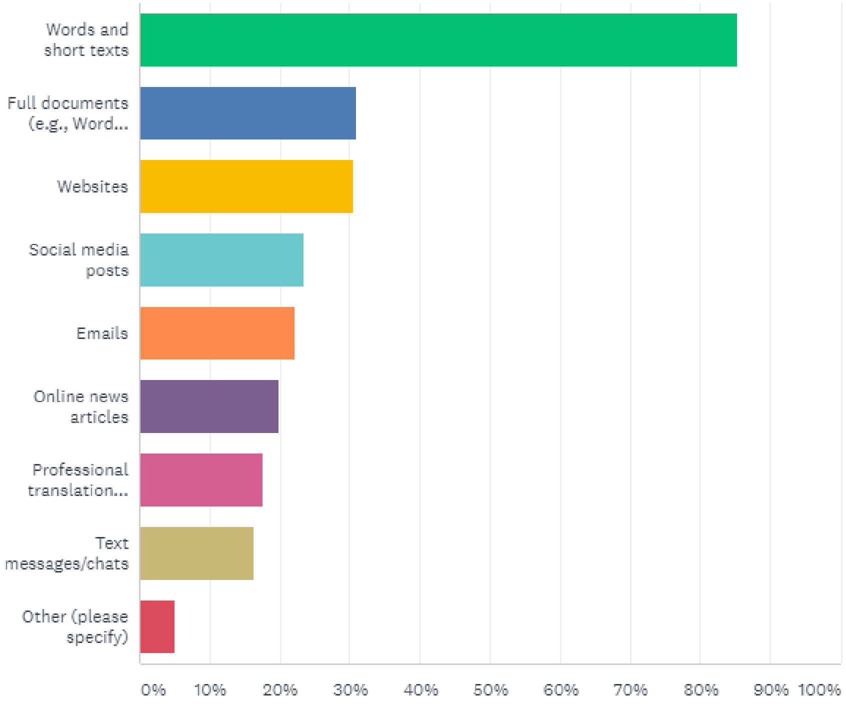

Figure 7: Types of texts translated with MT

Finally, rounding out the importance of MT use for larger languages in online translation sites for translating texts, documents, and websites, the survey found that the top purpose for utilizing MT was for the “reading and analysis of information” (59%) and the “preparation of written texts in other languages” (46%).

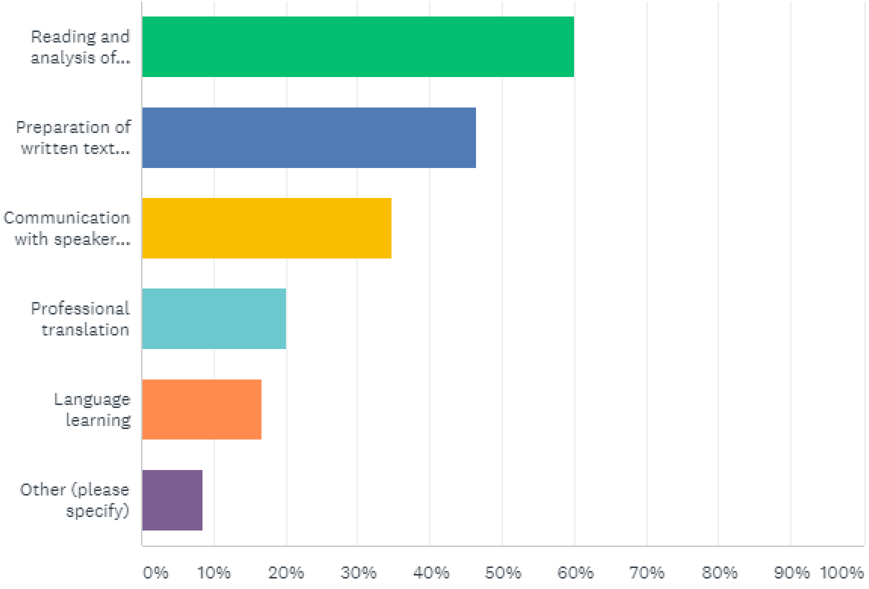

Figure 8: Primary purpose for using MT

Though the responses to the Language Technology Survey clearly demonstrate the widespread usage of MT for larger languages, in online translation sites, and for reading and analyzing texts in multiple languages, the survey also draws attention to respondents' awareness of the limitations of MT.

As seen in Figure 9, the majority of respondents (71%) acknowledged that they trust MT results "only somewhat," agreeing that they "usually double-check machine translation results with other sources." This statistic reflects a healthy understanding of the limitations of MT and its proper usage in online translation sites – that is, as a tool for reading texts at a so-called gisting level.

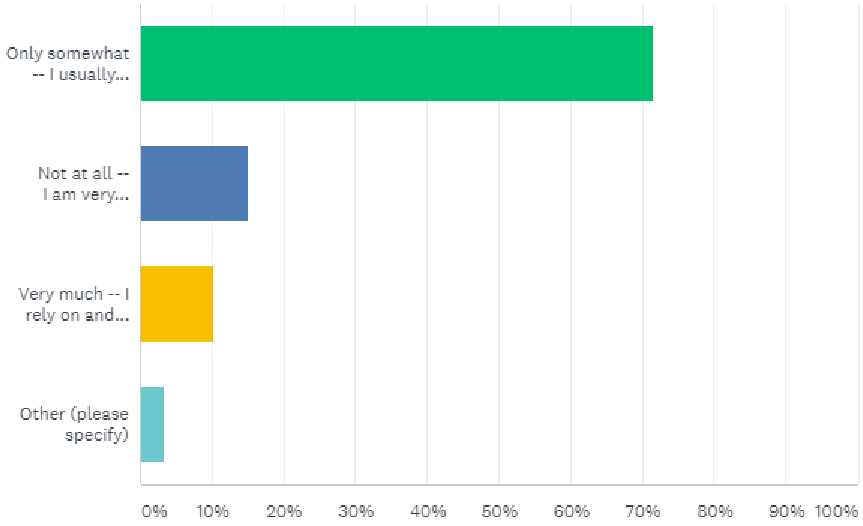

Figure 9: Responses to question: "How much do you trust MT in helping you read a text and make a decision about its content?"

While the survey found convincing evidence of the widespread use of machine translation – with 76% of respondents claiming that they had used MT for their native language – the KPLEX survey revealed a much lower rate of usage for other LT services, such as Automated Speech Recognition (ASR) and chatbots/virtual assistants.

Though these two LT solutions are written about in much detail in the abovementioned studies and reports on the language technology landscape (see the section: Findings of previous LT reports) – namely, in the reports *Language Technologies for Multilingual Europe – Towards a Human Language Project*, *Strategic Research Agenda for Multilingual Europe*, and *Language Equality in the Digital Age* – our survey found actually very limited usage of such services.

In fact, the majority of respondents answered that they do not use ASR (53%) for their native language; an even higher majority answered that they do not use chatbots/virtual assistants for their native language (75%).

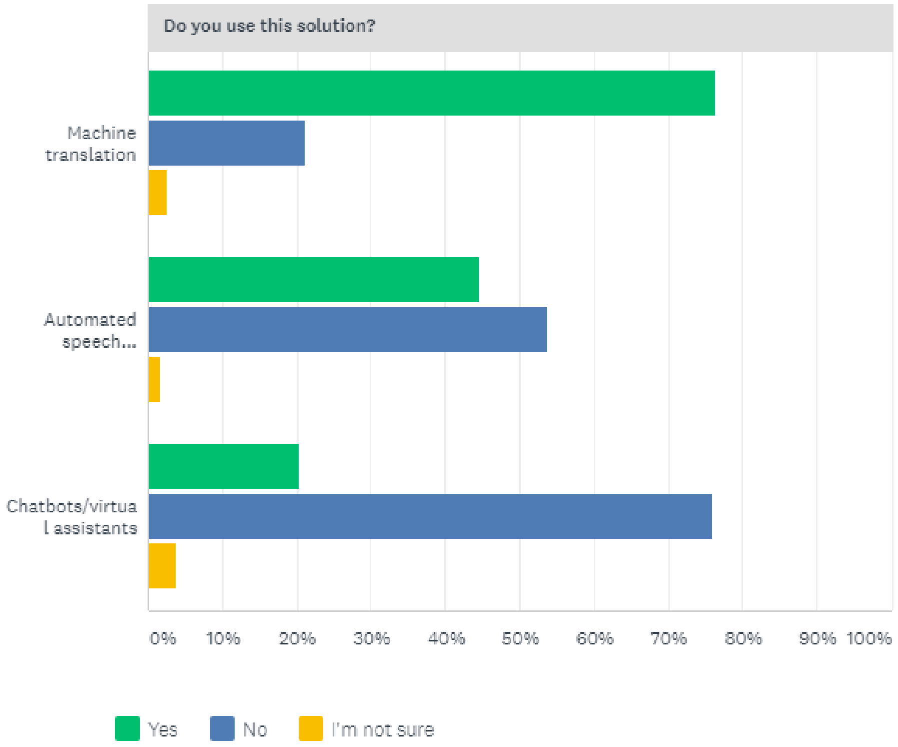

Figure 10: Usage of various LT solutions

These findings are therefore rather troubling for LT, as it indicates that the level of support of ASR and chatbots for EU languages is very unevenly spread. Both ASR and chatbots have the potential to serve as powerful tools for accessing information (see the section: Findings of previous LT reports), by leveraging advances in speech technology and AI. But without adequate coverage for a wide range of languages, usage will be limited to just a few major languages and inhibit usage by users in their native languages.

This claim is further backed up by the data accumulated in two crucial questions about how respondents felt about the importance of LT support for their native language. When asked to rate, on a scale of 1-5, how importance it is to ensure LT support for their native language, the average response was 4. This clearly indicates that the development of LT tools is fundamentally important for users in their native language – not just English.

At the same time, when asked "To what extent do you agree with the following statement: 'My native language has full support from LT services like MT, virtual assistant chatbots, and voice controlled devices'?" the majority answered “partially” (55%) and, most troublingly, “not at all” (24%).

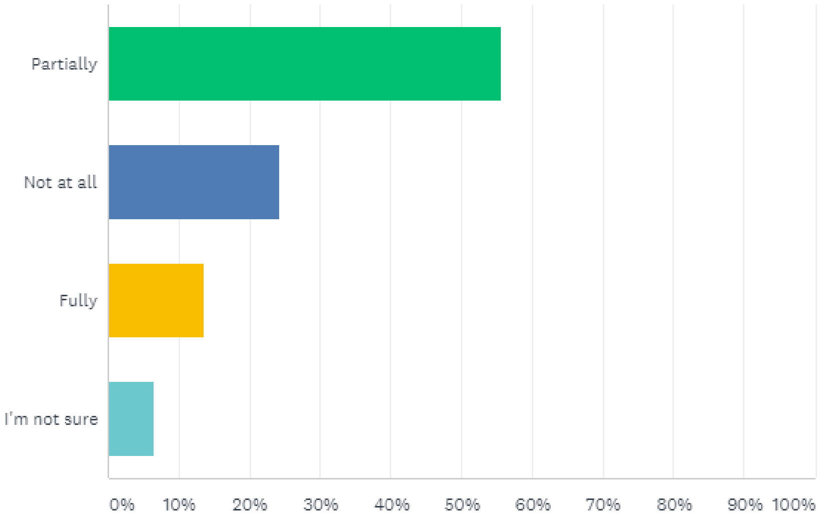

Figure 11: Response to question: "To what extent do you agree with the following statement: 'My native language has full support from LT services like MT, virtual assistant chatbots, and voice controlled devices'?"

Based on the responses in Figure 10, we can infer that most respondents are unsatisfied with the support provided by virtual assistants and voice-controlled devices – crucial LT solutions that have the potential to significantly help users to access information and navigate digital environments with more ease.

At the same time, as seen in Figure 11, respondents indicated several issues that they feel are problems with online machine translation services. Nearly all respondents answered that they felt "quality issues" were a concern (90%); 51% pointed to "language coverage issues"; and 30% pointed to "security issues." A relatively small percentage pointed to "integration issues" (18%) and "cost" (13%).

What is interesting here is the strong predominance of quality and language coverage as concerns over, say, security. This stresses the powerful importance of quality and language coverage in usage of machine translation – these issues are even more important than the much-discussed issue of data security for users of machine translation, even though data security and machine translation is an issue that has been frequently written about in the global media.[5]

[5] https://slator.com/technology/translate-com-exposes-highly-sensitive-information-massive-privacy-breach/

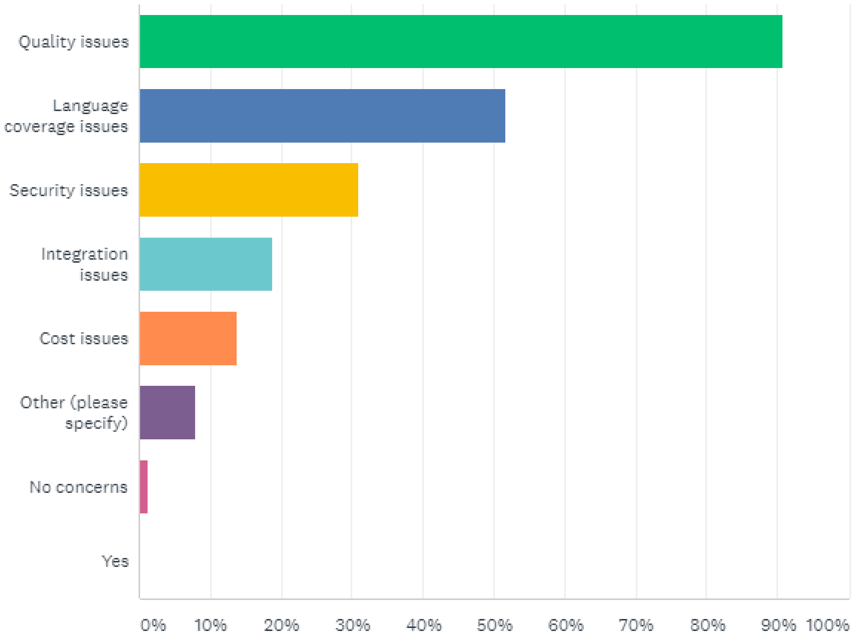

Figure 12: Main concerns with online MT services

Data skills in the LT community

A subset of questions in the Language Technology Survey was conceived to reveal another crucial area of information that was inadequately covered in previous LT reports – namely, the types of data management processes and skills at organizations and the staff members qualified to process and collect language data.

As we saw in the key findings, particularly in Figure 12, users of LT solutions like MT have found serious quality issues and language coverage issues when utilizing such solutions. To mitigate these issues, language data management processes are crucial, as are specialists trained to prepare data and to collect new data. Only by implementing new processes and employing highly skilled specialists can organization hope to overcome the problems they have encountered with LT solutions.

However, when asked to assess the language management skills in their organization (with language data including parallel texts, documents, audio files, terminology, translation memories, etc.; and data management encompassing the collection, processing, administrations, and workflow administration associated with language data), only 34% called them “strong.” 24% called their language management skills “sufficient” and a troubling 21% of respondents – nearly 1 out of 5 – referred to their organization’s language management skills as “poor” (see Figure 13).

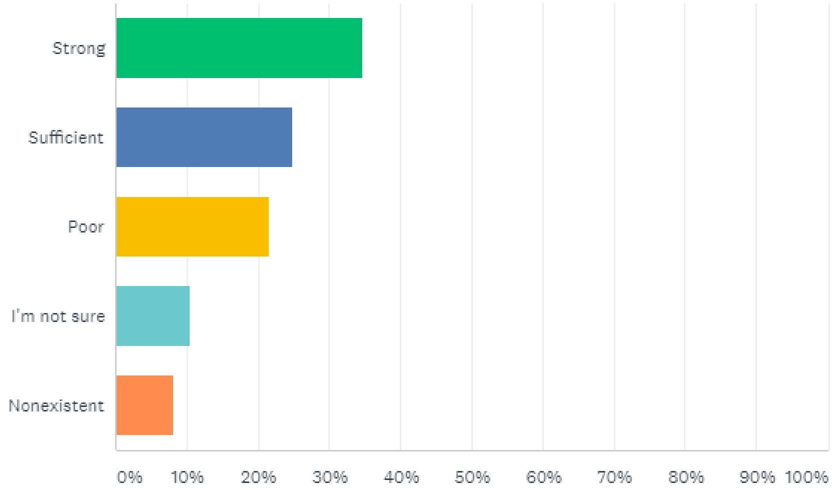

Figure 13: Assessment of language management skills at respondents' organizations

The fact that only a third of respondents feel that their organizations have "strong" language management points to a major potential problem in overcoming issues with language solution quality and with data inequality in terms of language coverage.

This finding is further supported by the results of the subsequent questions. When asked to characterize the language data management processes at their organization – an attempt to qualify the general assessment sought in the previous question – the majority called them either "loosely structured" (27%) or "ad hoc/case by case" (27%). Only 23% of respondents referred to these processes as "highly structured and regulated."

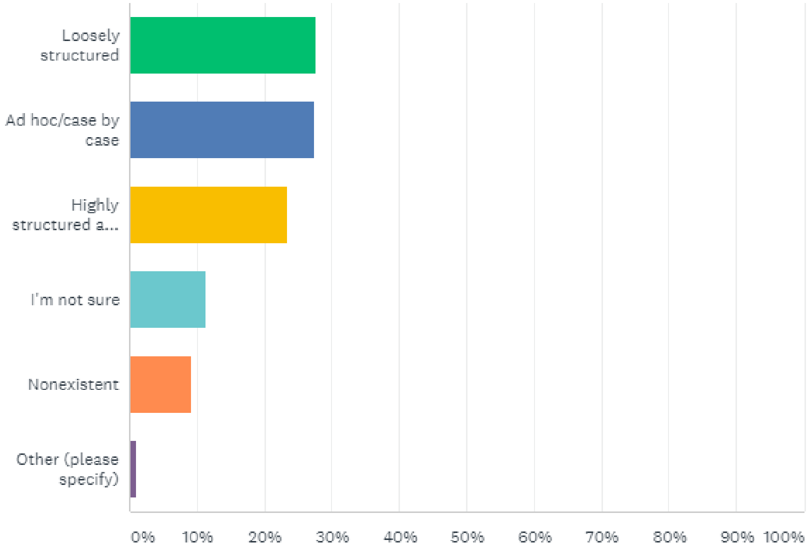

Figure 14: Characterization of language management processes at respondents' organizations

However, as was also underscored above, the only way to truly guarantee the proper leveraging of language data – and thus ensure the development and maintenance of strong language technology solutions – is to implement strong language management processes at organizations.

This negative trend toward under-implementation of language data processes at organizations is fully brought home in the final questions, which enquires about which of the following data specialists are employed on staff:

- Data processing specialist
- Data engineer
- Data manager
- Localization engineer

These specialists are highly skilled at not only managing data but also leverage data to be used for developing and deploying powerful LT solutions. Though respondents to the survey work in the language industry – with data at its core – the majority of respondents (57%) answered that they organization's employed "none of the above," meaning they don't employ data processing specialists, data engineers, data managers, or localization engineers.

In fact, only 26% of respondents said they employed data processing specialists on staff, while 22% employed data engineers and 19% employed data managers.

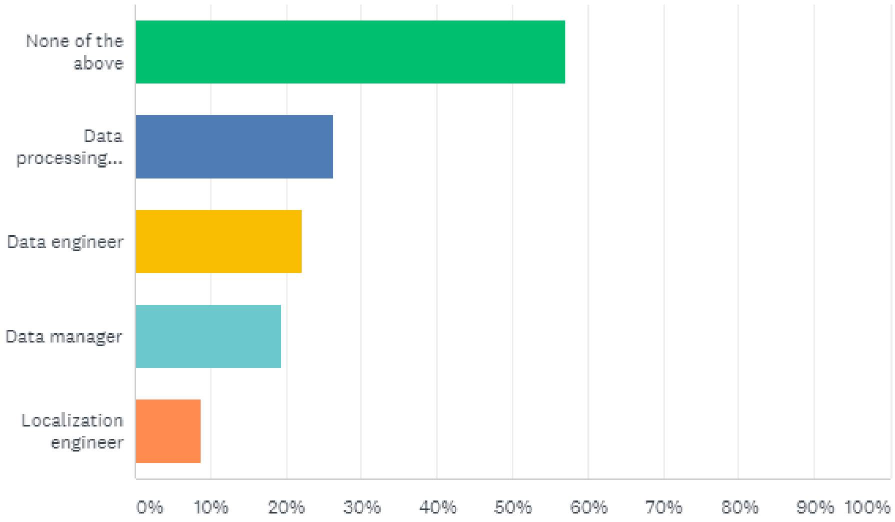

Figure 15: Data specialists employed on staff at respondents'' organizations

In order to raise the level of data management skills at organization and resolve the problems found with LT solutions – quality issues and language coverage issues – more data specialists are needed on staff at organizations in the language industry.

What is promising, though, is that most respondents seem to acknowledge these facts – that is, to understand the importance of data. This is clearly indicated in the final two questions of the Language Technology Survey. When asked "to what extent do you think language data is impacting your business," 58% indicated "very heavily." This strong statement by the majority of respondents implies that members of the language community are aware of the current impact of data on their business. 27% of respondents answered "somewhat," while only a small fraction (4%) answered "not at all."

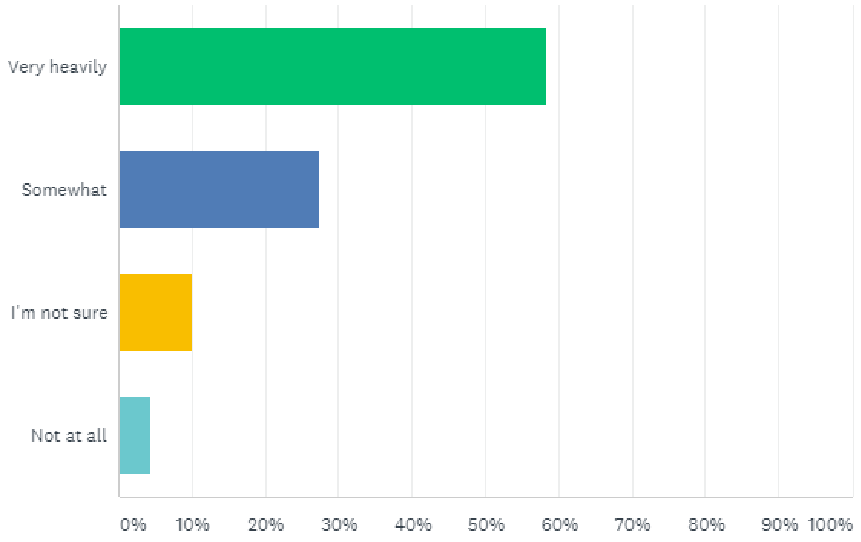

Figure 16: Responses to question: "To what extent do you think language data is impacting your business?"

Respondents therefore demonstrate that language data is impacting their business now. They also demonstrate, in their responses to the final question, that they feel this impacting will only increase in the future. When asked: "To what extent do you think language data will impact your business in the future?" the percentage of respondents who replied "very heavily" rises by ten points, to 68% (see Figure 17). To mirror this shift, the percentage of respondents who replied "somewhat" goes down by ten points, to 18%. The so-called naysayers, however, remain relatively unchanged: 3% still think that language data will impact their business "not at all" in the future. This indicates that a small subset of respondents – i.e., 3-4% – have no faith in the importance of language data either now or in the future.

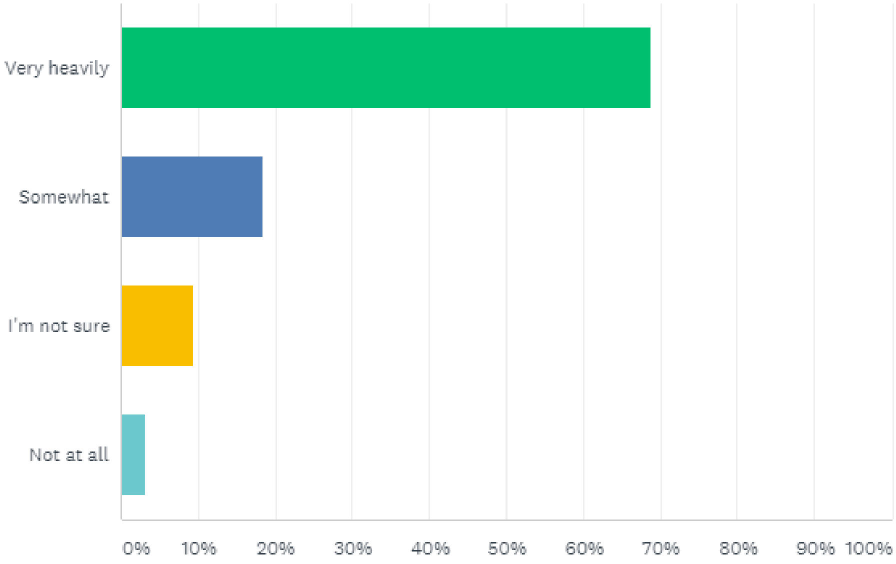

Figure 17: Responses to question: "To what extent do you think language data will impact your business in the future?"

The findings of the Language Technology Survey provided a large number of key findings that provide the basis for putting forth a number of major findings later in this report. However, the findings of the Language Technology Survey come into strongest focus when compared and contrasted with the findings of the Language Resources Survey, which will be analyzed in the next section.

Language Resources Survey

This survey was intended as a highly specific, targeted survey on language resources for specialists and experts in language resource and data processing, most of whom work for language technology companies, universities, and research centers in the EU.

Methodology

To gather respondents, researchers conceived and prepared a multi-question survey (questions in Annex 2) intended for a specific audience of language resource specialists. Following an extensive validation process, the survey was sent out to language specialists who are active in the LR sector and who have subscribed to a set of specific language technology and machine translation mailing lists:

- corpora@uib.no
- Mt-list@eamt.org
- nodali@helsinki.fi
- meta-net-all@meta-net.eu

These lists include the members of META-NET, a Network of Excellence consisting of 60 research centres from 34 countries.

Respondents

In total, 67 individuals responded to the survey.

Key findings

The very first responses already reveal a troubling pattern regarding the insufficiency of language resources, echoing the problems with language resource coverage found in the Language Technology Survey. When asked to assess the overall volume of openly available language resources, half of respondents (50%) answered that this was “insufficient” for meeting their needs. 34% responded that this was “somewhat sufficient” for meeting their needs. Only 11% thought that the overall volume of openly available resources was “sufficient” for meeting their needs.

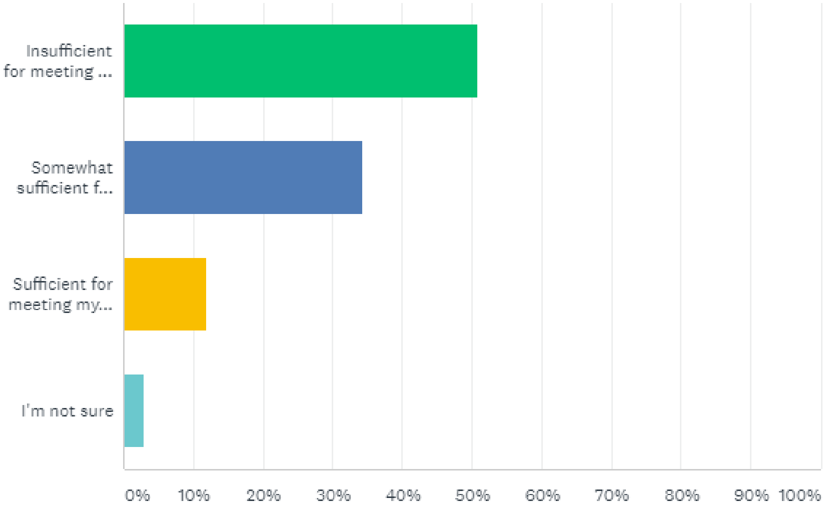

Figure 18: Assessment of overall volume of openly available language resources

Underscoring this finding, the responses to the second question – “In your work, have you encountered problems with language resource availability” – also demonstrate clearly that language resources are lacking. The vast majority of respondents, 88%, indicated that they had encountered problems with language resource availability in their work.

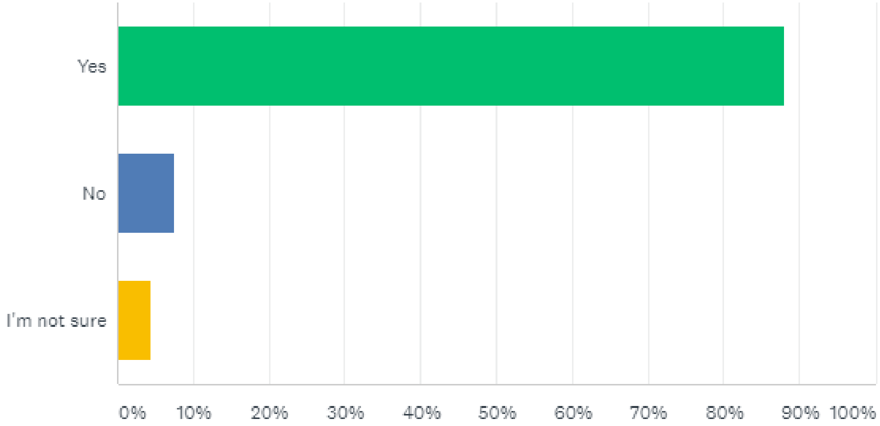

Figure 19: Responses to question: "In your work, have you encountered problems with language resource availability?"

Furthermore, when asked whether they had encountered problems with language resource quality – i.e., data noisiness, errors, etc. – the majority indicated that they had. 78% responded yes, they had encountered problems with language resource quality, and only 10% responded no, they had not (see Figure 20).

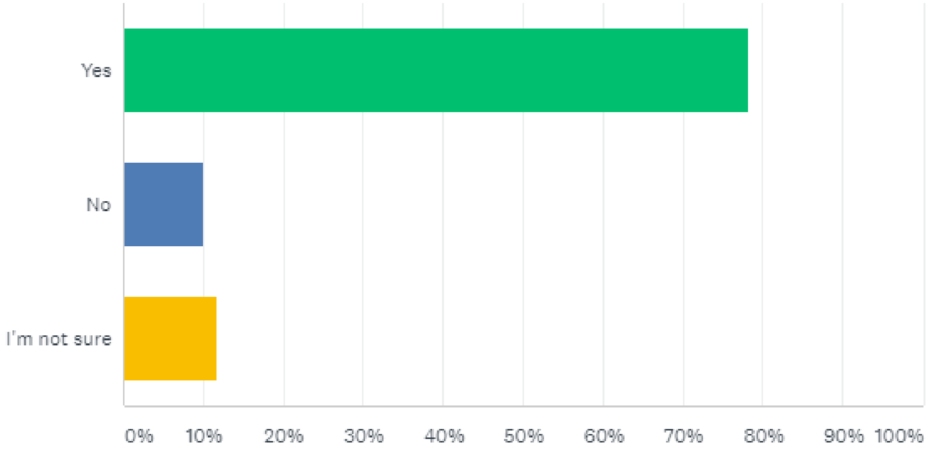

Figure 20: Responses to question: "have you encountered problems with language resource quality?"

As has been indicated multiple times in this report, problems with language resource availability and quality are strongly correlated with problems in language solution quality, as language solutions such as MT are built with language resources as input material. This would begin to explain the findings of the Language Technology Survey, wherein 90% of respondents indicated that "quality issues" and 51% indicated that "language coverage issues" were a problem with online machine translation services (see Figure 11).

In the next questions, experts were prompted to pinpoint the precise types of language resources issues they had encountered. The results (see Figure 21) are very strong and encompass a range of issues, the top three being:

- Data availability issues (89%)
- Openness of data issues (70%)
- Intellectual Property Rights (IPR) issues (67%)

Once again, we see that data availability is the top issue with language resources – simply put, not enough data being available for developing LT solutions. The other two top issues, however – openness of data and IPR issues – indicate that data may actually be available, but can't be accessed due to open data issues and IPR clearance. This is a troubling trend, which will be explored later in more detail and will form the basis for our policy recommendations later in the report.

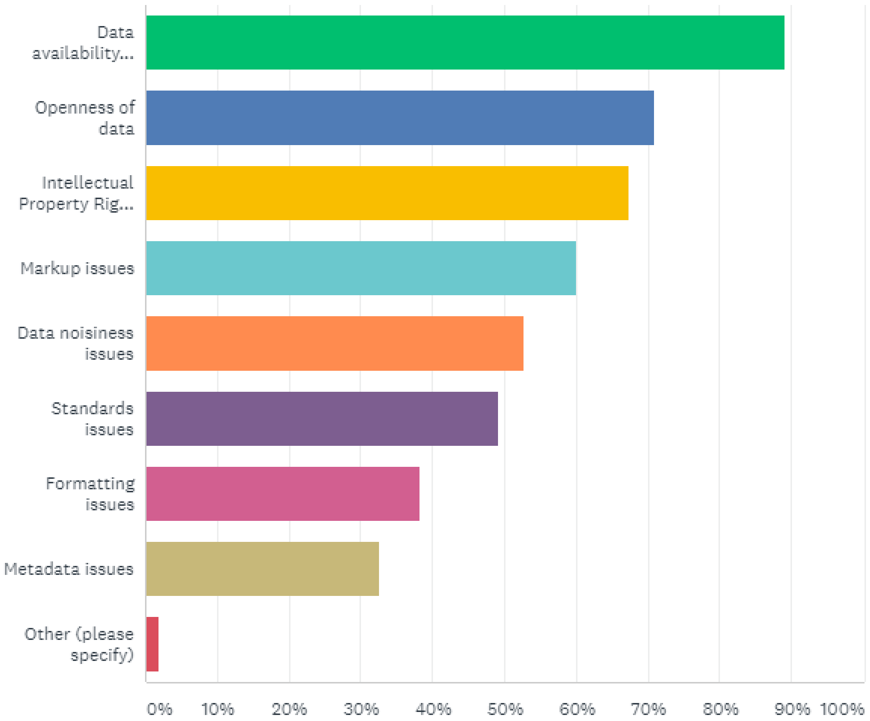

Figure 21: Issues encountered with language resources

To deal with these and other issues, respondents to the Language Resources Survey (highly specialized in the field of resource processing) indicated that extensive language resource processing was required. The survey found that almost all respondents to the Language Resources Survey (96%) had needed to perform processing of language resources to use them (e.g., data annotation, formatting, IPR clearance, etc.).

However, in the Language Technology Survey, which was sent to a wider swath of the language community, encompassing non-technical specialists in localization, the majority of respondents asserted that they would characterize the language data management processes in their organization as "ad hoc" (27%) and "loosely structured." Furthermore, when asked which of the following language processing specialists they employed on staff – data managers, data engineers, data processing specialists, and localization engineers – the majority (57%) answered “none of the above” (see the section: Data skills in the LT community). This discrepancy will be explored later in the major findings section of the report.

Though language resource specialists are the ones equipped to deal with issues like language resource quality, markup issues, and data noisiness issues, the tools to do this have also been lacking. When asked if they had encountered a lack of natural language tools (i.e., text processing tools, speech processing tools, and semantic analysis tools), the majority (80%) responded “yes.”

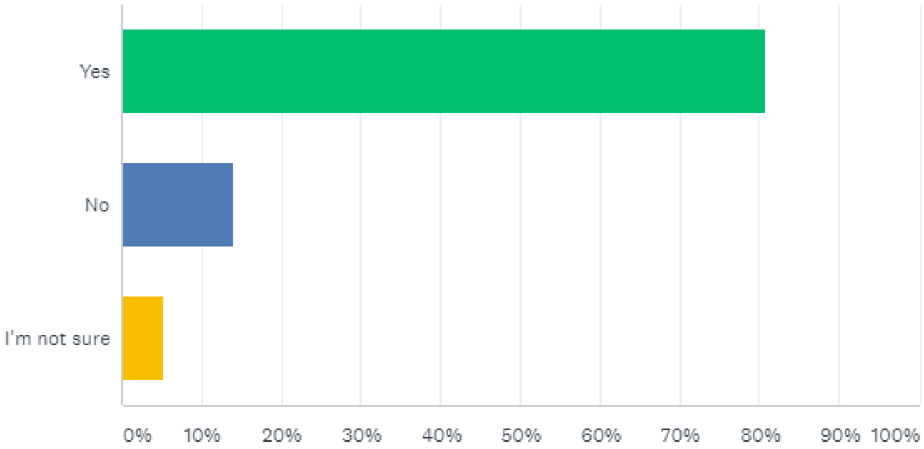

Figure 22: Response to question: "Have you encountered a lack of language processing tools?"

What this points to is essentially a long string of deficiencies and unavailability in the language space: specialists lack the tools to properly process language resources (and often lack the resources themselves), therefore poor-quality language resources are produced, leading to poor quality language solutions. These deficiencies will also be explored later in the report.

In the Language Resources Survey, specialists also point to the exact tools that they have found lacking. The top three were:

- Semantic analysis tools (80%)
- Text processing tools (e.g., tokenizers, parsers, part-of-speech taggers) (57%)
- Terminology tools (44%)

Interestingly, semantic analysis tools took the top spot. As semantic analysis represents the “next step” in language processing – adding a layer of meaning to language – the lack of these tools serves is a potential warning sign to problems in the further development of language technology solutions.

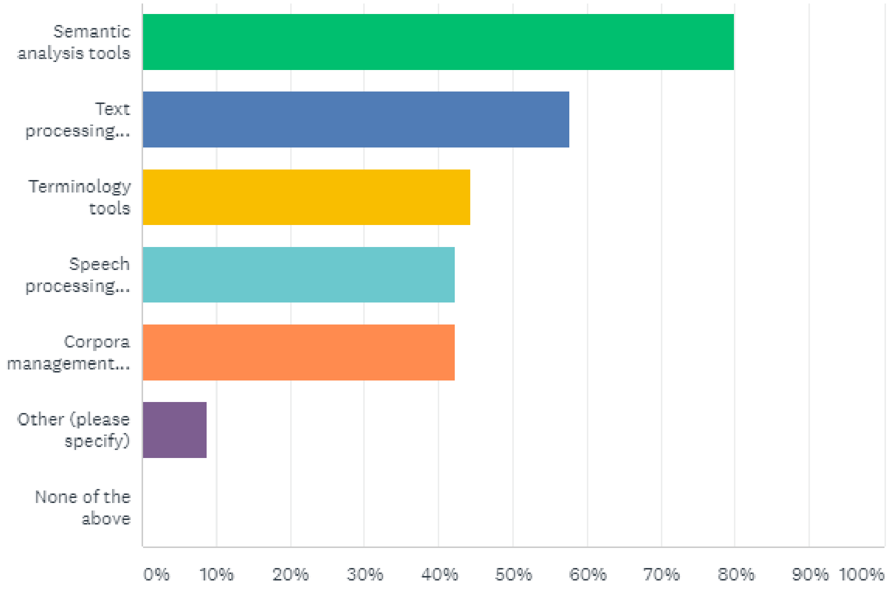

Figure 23: Language processing tools that have been lacking

The Language Resource Survey also uncovered valuable data points regarding the exact domains wherein language resources were well-covered as well as where they were lacking. When posed with the question of which domains had the best coverage of openly available language resources, specialists answered with the following three:

- News (64%)
- Legal (28%)
- IT (26%)

These figures, of course, fall in line with the overall availability of translated content – large volumes of online news articles, large volumes of online legal texts (particularly open corpora from the European Commission and the United Nations), and large volumes of texts translated by and for the large global tech companies (e.g., Microsoft, Google, Oracle, IBM, etc.).

These domains are also in line with the large volumes of public language resources available in some of the largest public corpora, including the European Commission's DGT-TM, or Directorate-General for Translation's Translation Memory, available in the Open Data Portal,[6] and the OPUS corpus, currently maintained at the University of Helsinki.[7] Both of these corpora contain a majority of legal data, while the OPUS corpus also contains news data.

But were problems begin to emerge is with domains that were identified as not having enough language resources – i.e., the worst coverage of openly available language resources. In the survey, specialists identified the following three domains as poorly covered:

- Medical (30%)

[6] http://data.europa.eu/euodp/en/data/dataset/dgt-translation-memory

[7] http://opus.nlpl.eu/

- Education (22%)
- Tourism (20%)

Of course, these domains are incredibly crucial for the EU economy, wherein both the pharmaceutical and tourism industries make up a large portion of GDP. Not having enough data for building language technology solutions for the medical, education, and tourism sectors risks upholding language barriers in these important sectors and seriously hinders cross-cultural dialogues. What is more, all of these sectors rely on human-to-human communication, wherein language barriers are inherent; the availability of powerful language solutions for these sectors could greatly promote cross-cultural and cross-lingual understanding.

These results will be analyzed in more depth in the major findings section of the report.

Further underscoring the importance of language resource availability, the majority of respondents (79%) also indicated that they had encountered resources that can't be used or are not available for use for their purposes. The major reasons why they couldn't be used were as follows:

- Data availability issues (68%)
- Openness of data (68%)
- Intellectual Property Rights (IPR) issues (57%)

What is interesting here is that data availability was significantly higher than an issue like cost in determining why a resource couldn't be used. Likewise, both openness of data and IPR issues were high on the list.

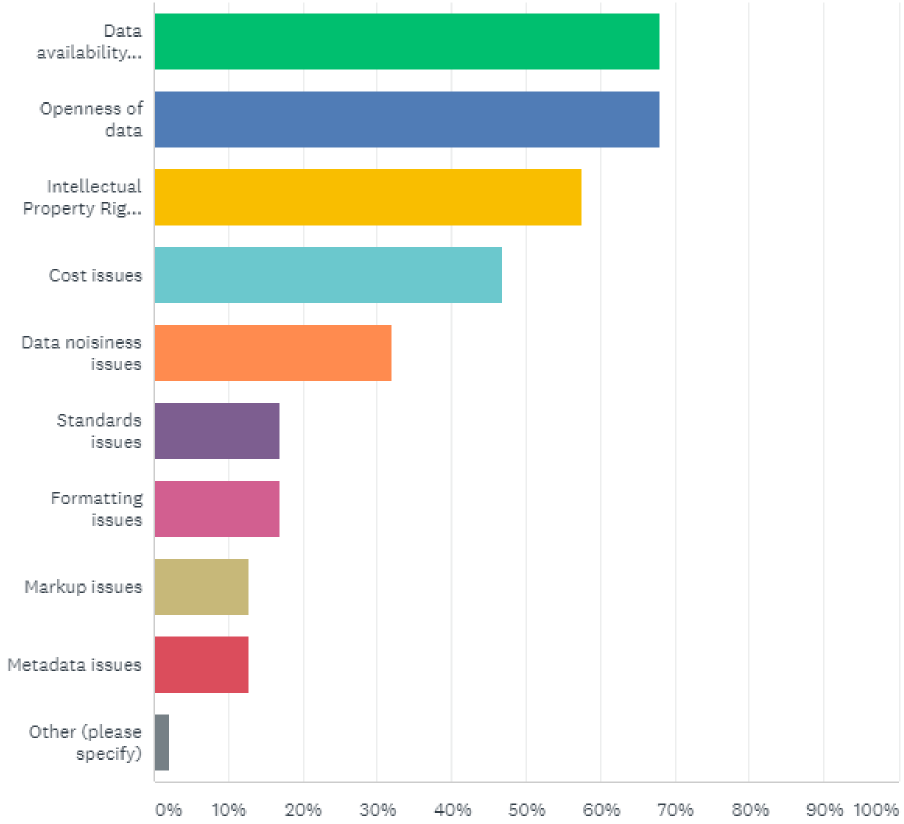

Figure 24: Reasons why language resources couldn't be used by respondents for their purposes.

Faced with these issues, the majority of respondents (51%) answered that, in their work, they had had to forego or turn down a project, research study, or other opportunity due to language resource issues. This means that, once again language resource issues are having an impact on the growth of the EU economy, as specialists are being forced opportunities due to problems with data. This is almost certainly the case in the greater Big Data industry, where the lack of data resources are preventing many projects from coming to fruition.

Respondents to the KPLEX Language Resources Survey also pinpointed the exact language resource issues that had caused them to forego a project, specifically:

- Availability of language resources (87%)
- Nonexistence of language resources (58%)
- Cost of language resources (58%)
- Intellectual Property Rights (IPR) issues (45%)

Here once again we find that the availability of language resources is the biggest factor in preventing specialists from engaging in projects, followed by the nonexistence of resources, which garnered just as many votes as the cost of resources as a factor forcing specialists to forego projects.

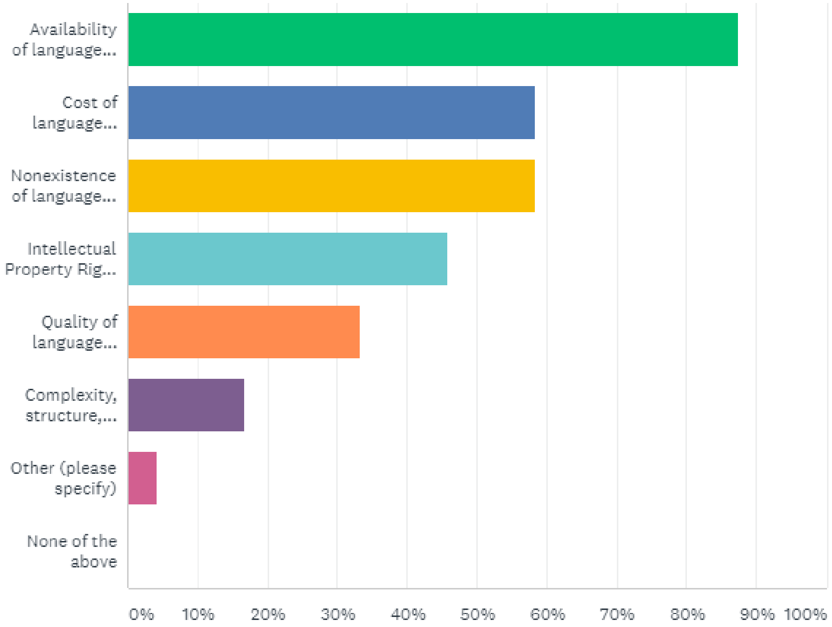

Figure 25: Issues that caused respondents to forego a project.

However, respondents also gave us an indication of what would help to resolve these problems. Specialists were asked: “To what extent would the following solutions to frequently encountered language resource issues be important and helpful to you in your work?” Respondents were then asked to rate the degree of importance of the following solutions: more freely available public data; easier IPR clearance processes; better language coverage; higher volumes of data overall; higher quality of data.

In response, the queried specialists indicated the following possible solutions to language resource issues as “essential”, presented below in their order of preference:

- More freely available public data (67%)
- Easier IPR clearance processes (52%)
- Better language coverage (42%)
- Higher volumes of data overall (32%)
- Higher quality of data (30%)
- Expanded domain coverage (25%)
- Better structured data (11%)

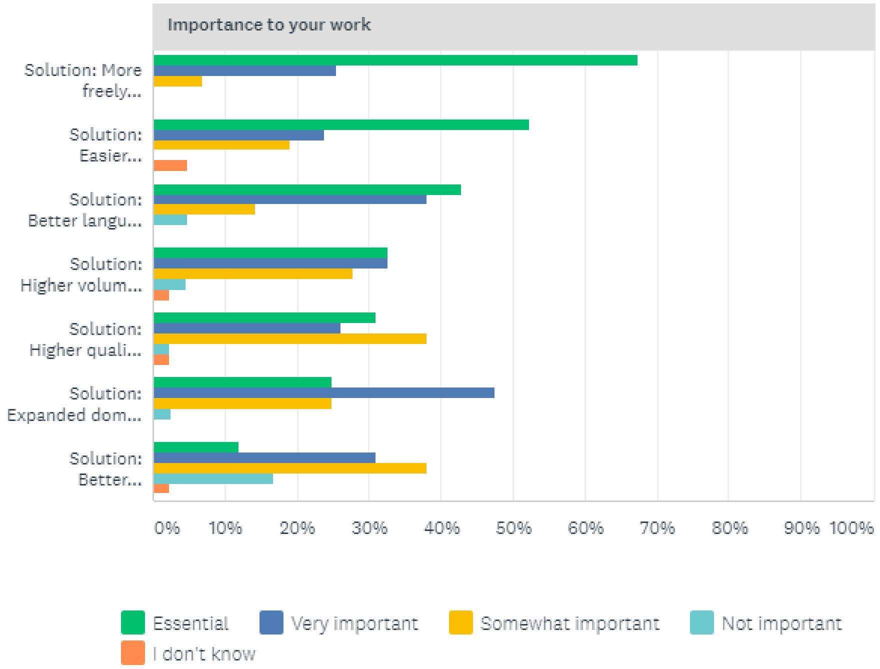

Figure 26: Importance of solutions to language resource issues

Herein we found an extremely interesting data point: that researchers find the provision of more freely available public data to be the most essential solution to language resource issues. This is well within the grasp of EU policymakers to enable, therefore it will serve as a basis for our policy recommendations later in the report.

Likewise, easier IPR clearance procedures and processes – the second most essential solution mentioned by respondents – can also be addressed by EU policymakers, helping to promote the development of this possible solution as a reality.

Of course, the solutions “better language coverage” and “higher volumes of data overall” also fall within the scope of enabling the creation and sharing of more language data for all EU languages – a central tenet of this report.

Fortunately, respondents themselves proved, through their response to the penultimate question, that their own approach to data sharing was in line with practices that can increase the supply of language data and coverage for EU languages. When asked, “Are you willing to share the language data available to your organization?” 48% responded “yes” and 37% said “maybe.” The heartening result is that absolutely nobody answered “no.”

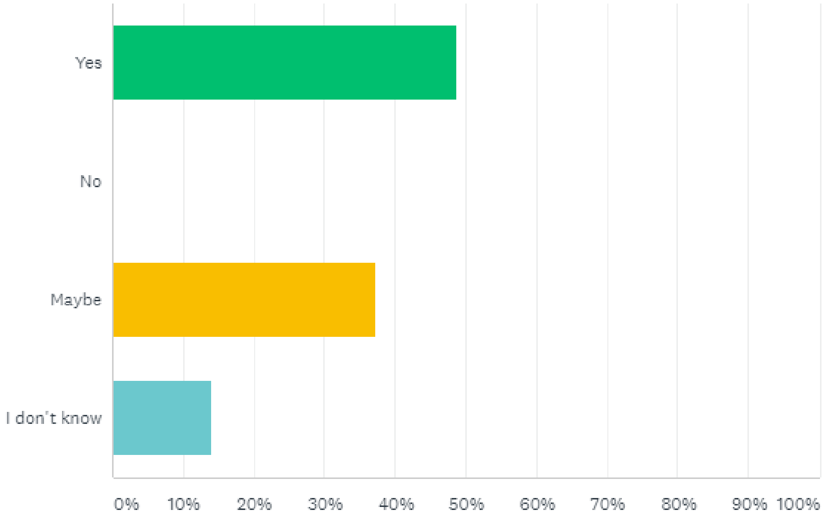

Figure 27: Respondents' answers to question: "Are you willing to share the language data available to your organization?"

For those 37% of respondents that answered “maybe” to the previous question, the final question of the Language Resources Survey teased out the reasons why respondents would have concerns about sharing their data. When asked “what are your main concerns in sharing the language data available to your organization?” respondents answered as follows:

- IPR clearance issues (56%)
- Data may contain personal information (41%)
- Lack of personnel required to share data (41%)
- Confidentiality issues (39%)
- Loss of competitive advantage in relation to competitors (21%)
- Quality of data (19%)
- Lack of motivation to share data (17%)

Here again we find IPR clearance as the top obstacles to sharing data – fortunately, one that can be deal with by policymakers. We also find a concern for personal information, which can also be dealt with by specialized anonymization tools. Obviously these tools must be made more widely available for data specialists, in addition to training on how to use them. Lack of personnel also points to a problem found in the Language Technology Survey: a decided lack of qualified data experts and specialists – which, again, will be presented later in this report.

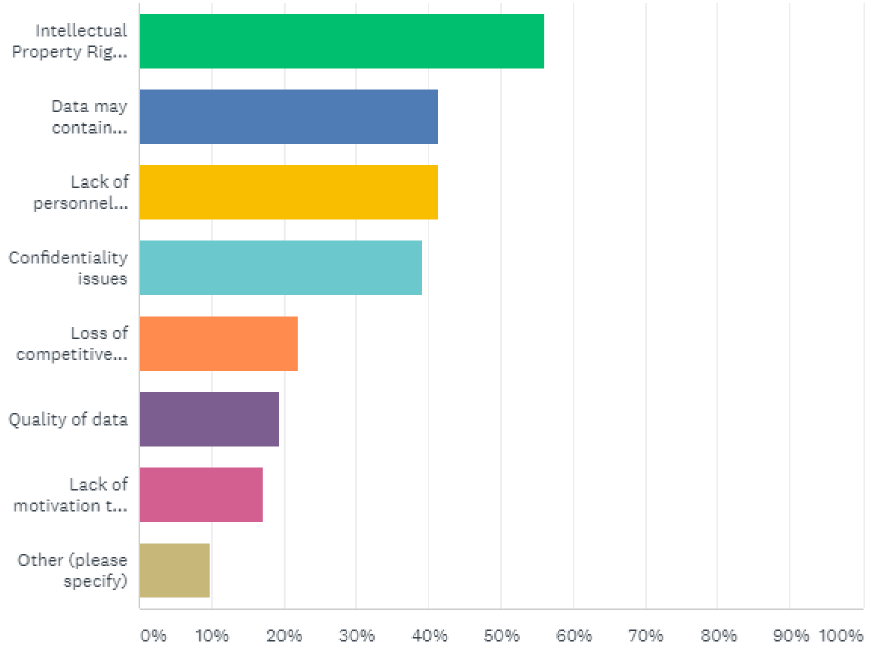

Figure 28: Main concerns of respondents in sharing their language data

Major findings of KPLEX Surveys

Both of the KPLEX surveys, the Language Technology Survey and the Language Resources Survey, present fascinating insight into the state of language data in the EU.

Language Technology Survey

The Language Technology Survey presents several major findings:

- MT systems are most widely used in generic online translation services, for English and the FIGS languages (French, Italian, German, and Spanish), in order to translate words and short texts for the purposes of reading and analysis
- Users' main concerns with online MT are:
    - (1) quality issues (90%)
    - (2) language coverage issues (51%)
- Though MT is used by a large number of users for their native language, other LT solutions like chatbots and automated speech recognition are used by small number of individuals for their native language, with ASR used by only 47% for their native language and chatbots used by only 25% of individuals for their native language
- More than half of users are only partially satisfied with LT support for their native language – also reflected clearly in the low percentage of users who apply chatbots and ASR for their native language – and 25% are not satisfied at all
- Language data management processes at organizations are mostly "loosely structured" or "ad hoc"

- More than half of organizations do not employee data engineers, data specialists, or data managers on staff

Language Resources Survey

The Language Resources Survey presents several major findings:

- Half of language resource specialists felt that the overall volume of language resources is insufficient to meet their needs
- The most frequently encountered issues with language resources are:
    - (1) data availability issues (89%)
    - (2) openness of data issues (70%)
    - (3) Intellectual Property Rights (IPR) issues (67%)
- The majority of language resource specialists (88%) have found problems with language resource availability and the majority (78%) have found problems with language resource quality in their work
- Almost all language resource specialists (96%) have had to perform processing of language resources (e.g., data annotation, formatting, IPR clearance, etc.) to use them, though the majority of specialists (80%) have encountered a lack of natural language tools to do so (e.g., text processing tools, speech processing tools, and semantic analysis tools)
- The tools that language resource specialists most frequently found lacking were:
    - (1) semantic analysis tools (80%);
    - (2) text processing tools (e.g., tokenizers, parsers, part-of-speech taggers) (57%)
    - (3) terminology tools (44%)
- Domains that specialists found had the <u>most</u> available data were as follows:
    - (1) news
    - (2) legal
    - (3) IT
- Domains that specialists found had the <u>least</u> available data:
    - (1) medical,
    - (2) education
    - (3) tourism
- The majority of language resource specialists (79%) have encountered resources that can't be used or are not available for use for their purposes, due to the following reasons:
    - (1) data availability issues
    - (2) openness of data
    - (3) IPR issues
- The majority of language resource specialists (51%) have had to forego or turn down a project, research study, or other opportunity due to language resource issues, specifically:
    - (1) availability of language resources
    - (2) nonexistence of language resources
    - (3) cost of language resources
    - (4) IPR issues

- Solutions that language resource specialists considered “essential” in helping to overcome language resource issues:
    - (1) more freely available public data
    - (2) easier IPR clearance processes
    - (3) better language coverage
- No language resource specialists answered that they weren’t at all willing to share the language data available to their organization
- Leading factors that were a concern to language resource specialists in considering whether to share date included:
    - (1) IPR clearance issues
    - (2) data may contain personal information
    - (3) lack of personnel required to share data
    - (4) confidentiality issues

Conclusions

In its two surveys of the language community, KPLEX found that MT systems are most widely used in generic online translation services for the world’s largest languages: English, French, Italian, German, and Spanish. However, according to the surveys, more than half of users said they were only partially satisfied with LT support for their native language, and 25% were not satisfied at all. Crucial LT solutions like chatbots and automated speech recognition (ASR) are used by small number of individuals for their native language, with chatbots used by only 25% of individuals for their native language.

The only remedy to this situation – wherein users are dissatisfied with LT support for their native language and therefore underuse crucial LT solutions like chatbots and ASR – is higher quality LT solutions, which of course means more LR, i.e., data. However, KPLEX also found this to be a serious problem. The surveys found that half of LR specialists felt that the overall volume of LR is insufficient to meet their needs, with the least amount of data available in the following crucial domains: medical, education, and tourism. The most frequently encountered issues with LR are data availability, openness of data, and Intellectual Property Rights (IPR) issues.

A lack of data processing tools for all EU languages, as well as the availability and openness of data, were also identified as a problem. The majority of LR specialists (80%) have encountered a lack of natural language tools (i.e., text processing tools, speech processing tools, and semantic analysis tools) to process data –which they also acknowledged to be a crucial step in developing LT solutions. Moreover, the vast majority of LR specialists (79%) have encountered LR that can’t be used for their purposes – due to data availability, openness of data, and IPR issues – and the majority of LR specialists (51%) have had to forego or turn down a project, research study, or other opportunity due to LR issues, most frequently on account of the availability or non-existence of LR.

What this shows is that access to and availability of LR and language processing tools for all EU languages continues to be a major issue, and almost certainly has led to the dissatisfaction of LT users in solutions for their native language. Survey respondents asserted that they considered the following solutions “essential” in helping to overcome language resource issues: more freely available public data, easier IPR clearance processes, and better language coverage. Helping to make more data available for LT developers in all EU languages – particularly public data – as well as easing IPR clearance procedures, is therefore a crucial step in providing better LT solutions for users.

Promoting the spread of better data management processes, and the proliferation of data management skills at organizations, can also lead to better LT solutions. Though survey respondents acknowledged that language data was impacting their business very heavily, and would continue to do so in the future, respondents admitted that language data management processes at organizations are mostly loosely structured or ad hoc.

Moreover, more than half of organizations do not employee data engineers, data specialists, or data managers on staff. Therefore promoting the spread of better data management processes, for example, by making available educational materials on data management and data literacy, and promoting educational opportunities for technically skilled employees to requalify as data managers, could help to resolve this serious lack of data skills at organizations.

In conclusion, KPLEX research in WP5 has illuminated a strong link between user dissatisfaction with LT solutions for their languages and an overall lack of LR and data tools for developing these systems. KPLEX has also found a lack of not only tools for processing LR, but also a lack of data management skills. The existence of these issues in the LT industry also portends their emergence in the Big Data industry as a whole. Therefore, the European Union must make policy decisions on promoting data availability and skills before these issues pose a serious crisis for Big Data in the near future.

Policy recommendations for the European Commission

In this report, the KPLEX researchers utilized the LT industry as a test case for examining issues surrounding data and its availability, as well as the impact of data on technology, infrastructure, and employment. That being said, many of the issues in LT can also be applied to the Big Data industry as a whole, encompassing Therefore many of these issues threaten to pose a political problem for the European Union.

To mitigate these potential issues, KPLEX researchers propose several policy actions for the European Commission:

- Promote the availability of language resources as openly available data (e.g., in the EU Open Data Portal) for all EU languages
- Promote the availability of language resources in crucially under-resourced domains: medical, education, and tourism
- Promote the spread of better data management processes, so that organizations are empowered to process language data and build their own powerful language technology solutions
- Ease IPR clearance processes for publicly available language resources, so that language specialists do not face obstacles in utilizing data for building valuable language technology solutions for all EU languages

Availability of language resources for EU languages

The KPLEX surveys clearly indicate that language resources are lacking for many EU languages. The surveys found that half of language resource specialists felt that the overall volume of language resources is insufficient to meet their needs. As a result, more than half of users are only partially satisfied with LT support for their native language.

To mitigate this problem, KPLEX researchers suggest a policy that seeks to promote the sharing of open data by EU institutions, particularly the sharing of Translation Memories accumulated by these institutions in localization contracts.

Availability of language resources for under-resourced domains

The KPLEX surveys found a lack of language resources in the following domains: medical, education, and tourism. When establishing a policy that promotes the sharing of open data by EU institutions, the EU should also place a particular emphasis on promoting data sharing in these domains.

The EU can also achieve a boost in language resources for these domains by targeting medical organizations, tourism boards, and universities, in particularly, to share their data in open data portals.

Spread of data management practices

The KPLEX surveys found that language data management processes at organizations are mostly “loosely structured” or “ad hoc,” and that more than half of organizations do not employee data engineers, data specialists, or data managers on staff. This threatens to seriously inhibit the ability of organizations to manage language data.

To mitigate this problem, KPLEX suggests implementing policies that would promote the spread of better data management processes, so that organizations are empowered to process language data and build their own powerful language technology solutions. One example of how this could be achieved is by making available educational materials on data management and data literacy, and by promoting educational opportunities for technically skilled employees to requalify as data managers, could help to resolve this serious lack of data skills at organizations.

Easing of IPR clearance processes

The KPLEX surveys found that IPR issues were the third most frequently encountered issue with language resources , as well as the top three reason why the majority of LR specialists have encountered resources that can’t be used or are not available for use for their purposes. In addition, IPR issues were the fourth biggest reason why the majority of LR specialists have had to forego or turn down a project, research study, or other opportunity due to LR issues, as well as the number one leading factor that was a concern to LR specialists in considering whether to share data.

For this reason, KPLEX strongly suggests implementing policies that would ease IPR clearance issues for language resources. This position was also strongly supported by respondents in the KPLEX surveys: “easier IPR clearance processes” was the number two solution that language resource specialists considered “essential” in helping to overcome language resource issues. Therefore this should send a clear signal to the EU that easing IPR clearance processes will go a long way toward ensuring that more language resources are not only available to specialists, but also easy to use in building strong language solutions for overcoming language barriers in the EU.

References

Language Technologies for Multilingual Europe – Towards a Human Language Project, prepared by the Cracking the Language Barrier federation, http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V1.0-final.pdf

Strategic Research Agenda for Multilingual Europe, presented by the META Technology Council in 2013, https://link.springer.com/content/pdf/10.1007%2F978-3-642-36349-8.pdf

Language Equality in the Digital Age, a Science and Technology Options Assessment prepared for the European Parliamentary Research Service, http://www.europarl.europa.eu/RegData/etudes/STUD/2017/598621/EPRS_STU(2017)598621_EN.pdf

Annex 1: Language Technology Survey questions

Q1. What type of organization do you work for?

- Language technology provider
- Localization/translation services vendor
- Corporate localization department
- Public sector institution
- University/research institution
- Self-employed
- Other (please specify)

Q2. What is your occupation?

- Translator/linguist
- Editor
- Terminologist
- Researcher
- Developer
- Manager
- Government official
- Student
- Teacher/professor
- Other (please specify)

Q3. How often do you use machine translation?

- Regularly, every day
- Often, a few times per week
- Rarely, a couple times per month
- Never
- Other (please specify)

Q4. Which languages do you use machine translation for?

Q5. On a scale of 1-5, to what extent is the machine translation you use “fit for purpose,” i.e., does it serve your needs?

Q6. Where do you use machine translation? (click all that apply)

- Online translation sites (e.g., Google Translate)
- Mobile translation apps
- Translation feature in web browsers
- Social media (i.e., translation feature in Twitter, Facebook)
- Email systems
- Computer-Assisted Translation (CAT) tools
- Translation Management Systems (TMS)
- Corporate intranet service
- Other (please specify)

Q7. What kinds of texts do you translate with machine translation? (click all that apply)

- Words and short texts
- Full documents (e.g., Word files)
- Emails
- Text messages/chats
- Professional translation files (e.g., TMX, XLIFF, etc.)
- Social media posts
- Websites
- Online news articles
- Other (please specify)

Q8. What is your primary purpose for using machine translation in your work?

- Reading and analysis of information
- Preparation of written texts in other languages
- Communication with speakers of other languages
- Language learning
- Professional translation
- Other (please specify)

Q9. How much do you trust machine translation in helping you read a text and make a decision about its content?

- Very much -- I rely on and trust machine translated results
- Only somewhat -- I usually double-check machine translation results with other sources
- Not at all -- I am very cautious in using machine translation results
- Other (please specify)

Q10. What is your native language?

Q11. Have you used these language technology solutions for your native language?

- Machine translation
- Automated speech recognition

- Chatbots/virtual assistants

How would you rate their quality?

- Excellent
- Very good
- Satisfactory
- Poor
- Extremely poor
- I'm not sure

Q12. On a scale of 1-5, how important do you think it is to ensure language technology support for your native language?

Q13. To what extent do you agree with the following statement: “My native language has full support from language technology services like machine translation, virtual assistant chatbots, and voice controlled devices”?

- Fully
- Partially
- Not at all
- I'm not sure

Q14. What are your main concerns with online machine translation services? (click all that apply)

- Quality issues
- Cost issues
- Language coverage issues
- Security issues
- Integration issues
- No concerns
- Other (please specify)

Q15. How would you assess the level of language data management* skills in your organization?

**Language data includes parallel texts, documents, audio files, terminology, translation memories etc. Data management encompasses the collection, processing, administration, and workflow administration associated with language data*

- Nonexistent
- Poor
- Sufficient
- Strong
- I'm not sure

Q16. How would you characterize the language data management processes in your organization?

- Nonexistent
- Ad hoc/case by case
- Loosely structured

- Highly structured and regulated
- I'm not sure
- Other (please specify)

Q17. Which of the following professionals do you employ on staff?

- Data manager
- Data engineer
- Data processing specialist
- Localization engineer
- None of the above

Q18. To what extent do you think language data is impacting your business?

- Not at all
- Somewhat
- Very heavily
- I'm not sure

Q19. To what extent do you think language data will impact your business in the future?

- Not at all
- Somewhat
- Very heavily
- I'm not sure

Annex 2: Language Resource Survey questions

Q1. How would you assess the overall volume of openly available language resources?

- Sufficient for meeting my needs
- Somewhat sufficient for meeting my needs
- Insufficient for meeting my needs
- I'm not sure

Q2. In your work, have you encountered problems with language resource availability?

Q3. In your work, which languages have not provided sufficient volumes of resources for meeting your needs? (click all that apply)

Q4. In your work, have you encountered problems with language resource quality?

Q5. In your work, which languages have displayed prominent issues with resource quality, e.g., noisiness, formatting, etc.? (click all that apply)

Q6. In your work, what kinds of issues have you encountered with language resources? (click all that apply)

- Data availability issues
- Data noisiness issues
- Standards issues
- Markup issues
- Formatting issues

- Intellectual Property Rights (IPR) issues
- Openness of data
- Metadata issues
- Other (please specify)

Q7. In your work, have you needed to perform processing of language resources (e.g., data annotation, formatting, IPR clearance, etc.) to use them?

Q8. In your work, have you encountered a lack of natural language tools (e.g., text processing tools, speech processing tools, semantic analysis tools)?

Q9. In your encounters, which languages have had insufficient coverage from natural language tools?

Q10. In your encounters, which tools have been lacking?

- Text processing tools (e.g., tokenizers, parsers, part-of-speech taggers)
- Speech processing tools (e.g., recognition, synthesis)
- Semantic analysis tools
- Corpora management tools (e.g., building, cleaning, searching, indexing)
- Terminology tools
- None of the above
- Other (please specify)

Q11. In your work, has the metadata quality of openly available language resources met your needs?

- Yes, I have found metadata quality to be sufficient for my needs.
- Sometimes metadata was not sufficient, but I found a way to deal with these issues and they were not an obstacle.
- No, metadata quality did not allow me to fully utilize resources for my purposes.

Q12. In your work, which three (3) domains have had the best coverage of openly available language resources?

- Legal
- IT
- Automotive
- Pharmaceutical
- Medical
- Financial
- News
- Patents
- Industrial manufacturing
- Education
- Environment
- Agricultural
- Tourism
- None of the above
- Other (please specify)

Q13. In your work, which three (3) domains have had the worst coverage of openly available language resources?

- Legal
- IT
- Automotive
- Pharmaceutical
- Medical
- Financial
- News
- Patents
- Industrial manufacturing
- Education
- Environment
- Agricultural
- Tourism
- None of the above
- Other (please specify)

Q14. In your work, have you encountered resources that can't be used or are not available for use for your purposes?

Q15. In your encounters, why couldn't these language resources be used for your purposes?

- Data availability issues
- Data noisiness issues
- Standards issues
- Markup issues
- Formatting issues
- Intellectual Property Rights (IPR) issues
- Openness of data
- Cost issues
- Metadata issues
- Other (please specify)

Q16. Have you ever had to forego or turn down a project, research study, or other opportunity due to language resource issues?

Q17. If you have had to forego a project due to language resource issues, what were the issues?

- Availability of language resources
- Cost of language resources
- Nonexistence of language resources
- Quality of language resources
- Complexity, structure, format of language resources
- Intellectual Property Rights (IPR) issues
- None of the above
- Other (please specify)

Q18. To what extent would the following solutions to frequently encountered language resource issues be important and helpful to you in your work? Please rate the degree of importance of each.

- Solution: Better language coverage
- Solution: Higher volumes of data overall
- Solution: Expanded domain coverage
- Solution: Higher quality of data
- Solution: Better structured data
- Solution: More freely available public data
- Solution: Easier Intellectual Property Rights (IPR) clearance processes

Q19. Are you willing to share the language data available to your organization?

Q20. What are your main concerns in sharing the language data available to your organization?

- Confidentiality issues
- Data may contain personal information
- Intellectual Property Rights (IPR) clearance issues
- Loss of competitive advantage in relation to competitors
- Quality of data
- Lack of personnel required to share data
- Lack of motivation to share data
- Other (please specify)