# Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs

**Georg Rehm, Stefanie Hegele**
DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
Corresponding author: georg.rehm@dfki.de

**Abstract**

We present the analysis of a large-scale survey titled "Language Technology for Multilingual Europe", conducted between May and June 2017. A total of 634 participants in 52 countries responded to the survey. Its main purpose was to collect input, feedback and ideas from the European Language Technology research and innovation community in order to assess the most prominent research areas, projects and applications, but, more importantly to identify the biggest challenges, obstacles and gaps Europe is currently facing with regard to its multilingual setup and technological solutions. Participants were encouraged to share concrete suggestions and recommendations on how present challenges can be turned into opportunities in the context of a potential long-term, large-scale, Europe-wide research, development and innovation funding programme, currently titled Human Language Project.

**Keywords:** Multilinguality, LR National/International Projects, Infrastructural/Policy Issues, LR Infrastructures and Architectures

## 1. Introduction

Europe is a multilingual society with 24 official Member State languages and many additional unofficial and regional languages as well as languages of minorities, immigrants and important trade partners. Nevertheless, day in and day out, language barriers keep severely hampering the free flow of information, thought, ideas, goods and products through the continent. Powerful multilingual as well as cross-lingual and monolingual language technologies, making use of the latest Artificial Intelligence algorithms in combination with ever-growing data sets, have the potential of helping to overcome language barriers.

The recent study "Language Equality in the Digital Age – Towards a Human Language Project", commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), recommends, to the European Union, to initiate a new, large-scale European Language Technology research, development and innovation flagship programme, called, in the study, the *Human Language Project* (HLP) (STOA, 2017). It is foreseen to be a long-term European collaborative programme between research, innovation, industry, academia, administrations and citizens with the goal of achieving the next scientific breakthroughs for the automatic processing and generation of written or spoken natural language. In addition to basic research, the Human Language Project[1] is foreseen to include applied research as well as innovation and commercialisation activities. Important research themes are, among others, (1) Crosslingual Big Data Language Analytics, (2) High-Quality Machine Translation, (3) Meaning, Semantics and Knowledge as well as (4) Conversational Technologies.

A key goal of our survey was to get an overview of the current situation of Language Technology research activities throughout Europe and to determine where important gaps and obstacles exist. More concrete details on the uniqueness of the HLP, which needs to be specifically designed for Europe's demands, are discussed in Section 5.

## 2. Recent Developments

The principle that all 24 official languages share an equal status and are supported on the same level is perpetuated in the EU Charter (Article 22) as well as in the Treaty on the European Union (Art. 3(3) TEU). The META-NET White Paper Series, however, has revealed that there is a steadily increasing and rather severe threat of digital extinction for at least 21 European languages (Rehm and Uszkoreit, 2012; Rehm et al., 2014).

To address this threat and recognise Europe's opportunities, among others, in the fostering of a truly Digital Single Market, META-NET[2] (a Network of Excellence consisting of more than 60 research centers in 34 European countries) has been committed to support work on multilingual technologies and to provide strategic guidance since 2010 (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; Rehm et al., 2016a). Selected META-NET activities were recently funded through the EU project CRACKER (2015-2017).[3] CRACKER's objectives encompass, among others, preparing and publishing research and innovation agendas (Rehm, 2015; Rehm, 2016; Rehm, 2017). It has also established the Cracking the Language Barrier[4] federation which acts as an umbrella initiative for European projects and organisations working on technologies for multilingual Europe.

Europe has a long-standing research, development and innovation tradition with several hundred universities and research centers performing excellent, highly visible and internationally recognised research on all European and many non-European languages. Especially in the field of Machine Translation most of the basic research has happened in Eu-

[1] While identical in name, the "Human Language Project" only bears a marginal relationship to previous initiatives bearing the same name. Among others, Abney and Bird (Abney and Bird, 2010) called their "universal corpus of the world's languages" the "Human Language Project". In 2012 TAUS launched their concept for an open language resources and tools platform, which TAUS called "Human Language Project", (see:https://www.taus.net/knowledgebase/index.php/Human_Language_Project). The initiative specified in the STOA report has a much broader scope and set of objectives than the two initiatives mentioned above.

[2] http://www.meta-net.eu

[3] http://www.cracker-project.eu

[4] http://www.cracking-the-language-barrier.eu

ropean research projects. Moses (Koehn et al., 2007), until 2016 the state of the art phrase-based statistical MT system, and recent European NMT results, especially those of the European research project QT21, are just two examples for excellence and world class research (Bojar et al., 2017). Nonetheless, challenges are omnipresent and must be addressed by the EU, the Member States as well as stakeholders from academia and industry.

## 3. Method

The survey contains a total of 29 questions (see Appendix A) of which 16 are open questions with free text answers. The remaining ones are a mixture of multiple choice and yes/no questions. The findings of this survey served as an important contribution to the final Strategic Research and Innovation Agenda[5] on Language Technologies for Multilingual Europe which was presented and discussed at the META-FORUM 2017 conference[6] on November 13/14 and published in its final version in December 2017.

The survey is divided into three main parts covering (1) background, research interests and projects of the participants, (2) visions for a large-scale European Language Technology research and development programme and (3) ideas on talent generation and retention in Europe. This division allowed to capture an overview of current and on-going research activities and developments in the field in the first part, reaching early-stage as well as more senior community members. The second part was intended to gather more expert knowledge with regard to visions and concrete plans for future work, in particular steps and prerequisites needed for initializing a large-scale Human Language Technology Project tailored especially to Europe's demands and current opportunities. The third and final part addresses the current challenge of the brain drain the European LT (and also AI) community is experiencing.

Participants were not obliged to answer all questions, but encouraged to fill in the ones they feel comfortable with. The survey was designed and set up using the service Typeform[7], a software for building online forms (see Figure 1).

## 4. Analysis

The survey was launched on 16 May and closed on 4 July 2017. As an incentive to maximise the number of answers, those who submitted the survey had the chance to win a tablet computer. After testing and making sure that the questions were phrased the right way, the survey was shared within a smaller circle (mainly members of META-NET, META, CRACKER as well as members of the Cracking the Language Barrier federation) with an appeal to share the survey within their own respective networks and also through social media. In a second round a wider audience of more than 4000 people was targeted, including participants of former META-FORUM and other conferences as well as respondents of the META-NET Open Letter campaign, conducted in 2015 (Rehm et al., 2016a).[8] We also announced the survey on the major mailing lists relevant for the field.

[5]http://cracker-project.eu/sria/
[6]http://www.meta-net.eu/events/meta-forum-2017/
[7]https://www.typeform.com
[8]http://multilingualeurope.eu

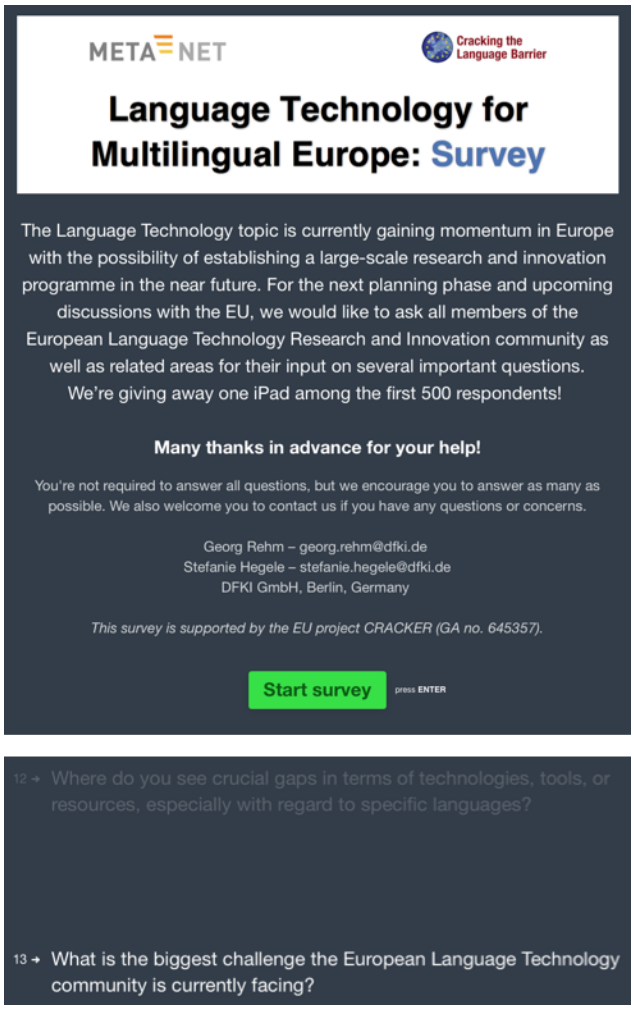


Figure 1: Welcome screen and example of survey questions

The survey created a total of 634 responses and, considering the number of questions, a surprisingly high completion rate of 27%. The average time needed for completing the survey was 35,48 minutes (see Figure 2). Both the completion rate and the average time indicate that the respondents are very passionate about the language topic and Europe's multilingual challenge. One major goal of this survey was to bring the European LT community together and gather responses from a wide and demographically distributed audience.

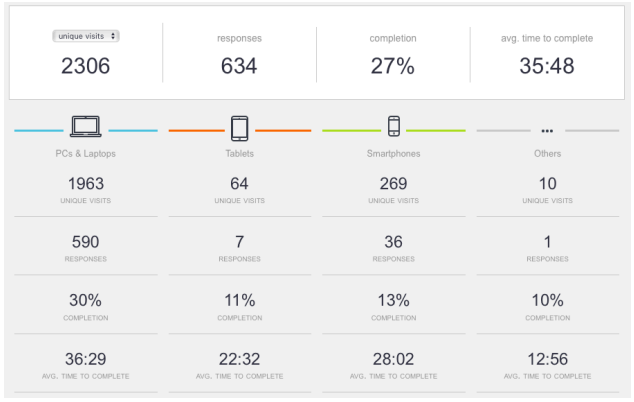


Figure 2: Survey completion rates on different devices

## 5. Analysis

The Human Language Project, as initially suggested in the STOA study (STOA, 2017), is to provide a sizable amount of funding in order for the field to reach a set of strategic research objectives. All European research, development and innovation projects that fall under the umbrella of the HLP are to be coordinated in a systematic way. The key scientific goal and also challenge of the HLP is an ambitious goal: Deep Natural Language Understanding by 2030 (including Generation). In the survey we asked the respondents multiple questions regarding key characteristics of a potential Human Language Project to get a better idea of where the needs, gaps and demands are. In the following we give a brief overview of the challenges, we discuss a possible setup of the HLP, provide details regarding important research areas mentioned in the responses as well as the economic impact (as suggested and discussed in the survey). Since the scope of this paper does not allow to analyse all 29 questions in detail, we focus on the ones we consider most insightful and provide relevant quantitative and qualitative statistics and findings (see Figure 5 for some of the key insights). We refer to specific questions, listed in Appendix A, using abbreviations in the form of Q1, Q2 etc. The analysis follows the survey's original tripartition.

### 5.1. Part 1: Background and Research Interests

The first part of the survey consisted of 14 questions aiming to collect background information of participants' organisations and their size (and also revenue if applicable) as well as the type of role and day-to-day responsibilities. Further, participants were asked to define the research fields, areas and sub-areas, methods and applications they work on. Particularly important for this survey was to assess in which economic sectors developed applications can be used. Finally, two open answer questions tackle the problems on current gaps (especially with regard to particular languages) and challenges within the European LT community.

Below we present the demographic details, the current challenges and gaps in terms of technology as well as their economic impact, especially with regard to the Digital Single Market (DSM).

#### 5.1.1. Respondents Demographics

Access statistics of the survey web page and Google Analytics reveal that the survey was opened by potential respondents in 67 different countries with most views from 1) Germany, 2) Spain, 3) United Kingdom and 4) Italy. Completed surveys were collected from 52 countries (see Figure 4). Among the represented countries were 37 European countries including 27 EU Member States.

As for socioeconomic statistics, the distribution shows that a large majority of participants hold senior roles at their respective organisations (such as professor, senior researcher, group leader etc.). This information about the roles seen in context with the seniority level (53% have more than 20 years of work experience and another 27% more than 10 years) and the participation from 52 countries clearly portrays a wide and diverse range of the European Language Technology research and innovation community (and even beyond). This expertise and long experience are accordingly reflected in the high quality of answers collected (see Figure 3 for more statistics). The most commonly represented research fields include Language Technology (64%), Computational Linguistics (56%), General Linguistics (42%), Artificial Intelligence (39%) and Computer Science (31%).

The majority of participants is based at universities and research centers. The most frequently mentioned organisations were: Charles University in Prague, Vilnius University, University of Copenhagen and DFKI. However, the survey also reached a substantial group of participants from industry, 33 (corresponding to 5% of all respondents) from large enterprises with more than 10000 employees such as Microsoft, IBM, Intel and Nuance and 68 (11%) from SMEs. When it comes to day-to-day responsibilities 71% of all participants state an involvement in research, closely followed by 52% naming project management and 43% project execution as their most crucial tasks. This variety of engagement and responsibilities allows to get insightful input on concrete research topics (for basic and applied research as well as innovation topics), methods and best practices as asked in the second part of the survey. In addition, the vast expertise in management and project acquisition indicates competence for answering questions related to strategic planning as well as questions requiring a wider perspective on the field such as the impact Language Technology could have on the Digital Single Market.

#### 5.1.2. Technological Gaps and Challenges

Regarding crucial gaps in terms of technologies for specific languages (Q13), almost 40% of all respondents highlight that there is insufficient research being done for minority languages and dialects, directly resulting in a shortage of available resources. This lack becomes most evident in Machine Translation applications for smaller European languages as well as other standard NLP tools and systems (according to approx. 19%). Further gaps mentioned are imposed by limited funding for low-resourced languages and copyright restrictions for certain data sets. Further, interoperability and standardisation need to be intensified.

When asked about the biggest challenge the European Language Technology field is facing at the moment (Q14) around 16% of all provided survey answers stress that the neglect of smaller languages is a severe threat, which is leading to a fragmented rather than a united and multilingual Europe. Assessing the languages most widely used in research (Q9 and Q10), around 90% state that they work with English (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. Still frequently used, though not even half as popular, are the big European languages: Spanish (49%), German (41%), French (37%) and Italian (23%).

Other challenges include the insufficient amount of data resources (approx. 13%), an unwillingness of collaboration within the community (approx. 8%) and, as already indicated above, a lack of funding (approx. 8%).

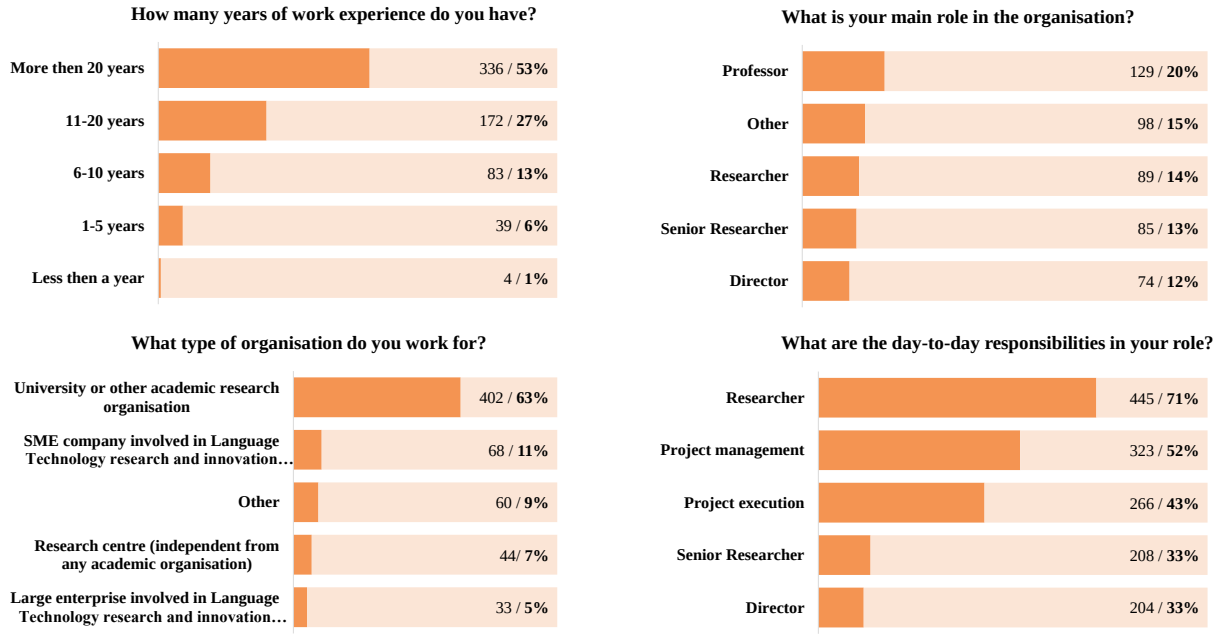


Figure 3: Overview of survey's socioeconomic statistics

### 5.1.3. Economic Impact and the DSM

We also asked the respondents questions (Q11, Q12) regarding the economic impact of language technologies, especially in the context of the Digital Single Market (DSM). Identified as the sectors to have the highest potential contribution to commercial growth are Education (71%), Information & Communication Technologies (64%) as well as Human Health & Social Work (45%). Specific services and applications that could benefit the Multilingual Digital Single Market comprise better Language Resources and Technologies (73%), Translation Services (46%), Multilingual Solutions for E-Learning (41%) and E-Health (38%). In the context of industries, sectors and verticals the necessity of an on-going knowledge transfer and effective collaboration between academia and industry is highlighted. The Health sector is unequivocally the most significant one, Education comes in second, closely followed by Tourism and Travel, Law and Justice, Translation, E-Commerce, Entertainment (incl. arts, creativity, culture and cultural heritage), Media, Business (incl. various services and business intelligence applications), Security, Public services and Administration, Government and Finance. Socio-economic opportunities are brought by guaranteeing better access to multilingual data and services for all people. This establishes a solid basis for the inclusiveness of minorities and people with special needs. Thus, in a wider context multilingualism helps remove barriers, fosters collaboration and creates more cultural awareness.

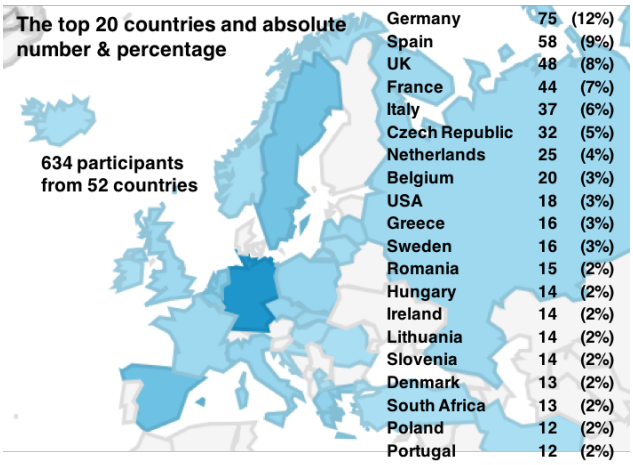


Figure 4: Number of collected responses sorted by country

## 5.2. Part 2: Visions for a Future Large-Scale Language Technology Programme

The second part entailed a total of 11 questions assessing the general support for a joint European Language Technology Project tackling the challenge of Deep Natural Language Understanding. In the following we analyse the questions on the organisational set up, strategic guidelines and governance of a potential Human Language Project, the most important research areas as well as applications and services that should be components of a HLP.

### 5.2.1. Support for a Human Language Project

The overall suggestion to initiate a large-scale Human Language Project (HLP) received substantial support from the group of respondents with 97% stating that they are in favour of establishing such a funding programme. Only a very small number of participants (3%) does not agree; their main arguments are unsuccessful previous attempts of similar programmes which did not achieve their targeted goals because of bureaucratic hurdles and a lack of focus. Further-

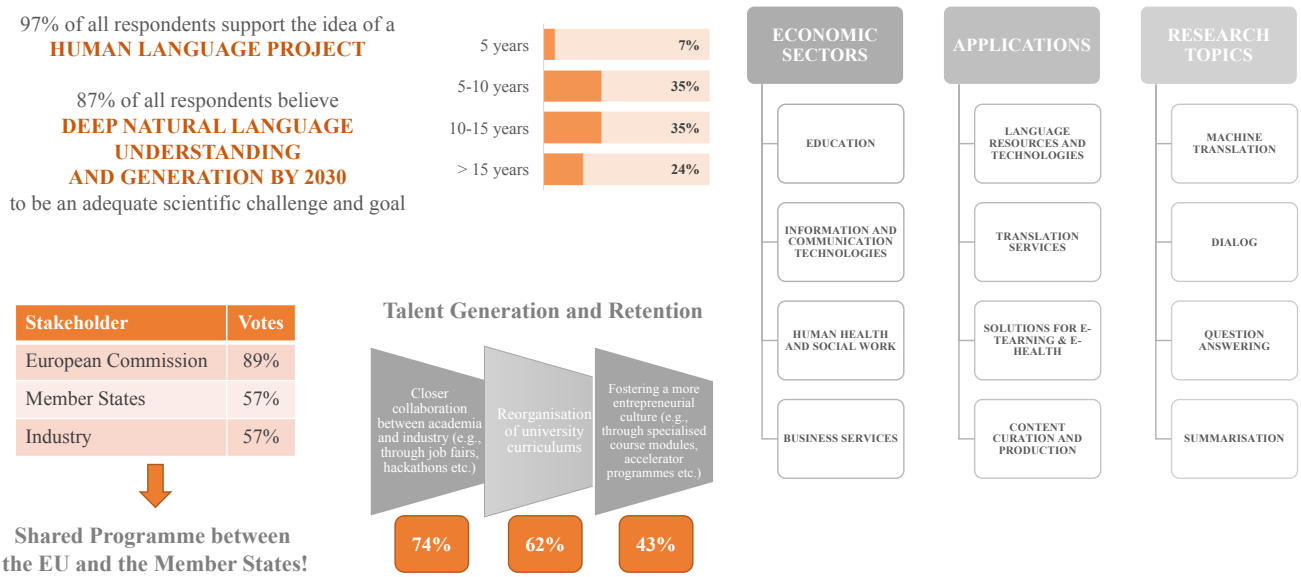


Figure 5: Overview of survey's key input on a Human Language Project

more, 87% consider the survey's suggested key strategic vision – to achieve Deep Natural Language Understanding and Generation by 2030 – as realistic and therefore an adequate scientific challenge. An appropriate timeframe would be likely to fall in the range of 10-15 years (7% believe that 5 years is a sufficient period, 35% opt for 5-10 years and another 35% for 10-15 years).

### 5.2.2. Organisational Frame and Governance

As far as funding is concerned a shared responsibility between the European Union, industry and member states was envisioned with the EU as the stakeholder that should be "naturally" responsible. The distribution of votes for stakeholder involvement looks as follows: European Commission (89%), Industry (57%) and Member states (57%).

When it comes to strategic guidance what can be derived from the survey responses is the strong suggestion to concentrate funding on smaller scale projects, starting bottom-up with smaller goals, and also to avoid heavy bureaucracy.

Regarding the governance of a potential HLP, one shared suggestion is the wish to put democratic organisation processes in place, e. g. with shifting presidents and elected committee and board members among institutions and countries. Also highlighted was the need to reposition the strategy of EU research with a focus on scientific breakthroughs in order to diversify from the US and large corporation paradigms. This involves fostering strong collaborations between stakeholders, better school and especially university education with more incentives for young researchers (see Section 5.3.), integration of user and customer experience as well as feedback processes, following market-driven approaches to ensure industrial growth.

### 5.2.3. Key Research Areas

In terms of research, the Human Language Project aims to tightly intertwine basic research, applied research, innovation and commercialisation (Q20).

As far as basic research is concerned a majority mentioned the further development of existing resources (incl. corpora, ontologies, dictionaries etc.) and improvement of data annotations (approx. 9%). In this context, effective legal frameworks for better accessibility are also necessary. Besides, basic research should be centered around deep learning and neural networks (approx. 7%) as well as Natural Language Understanding (approx. 7%). A majority also highlighted the need to further work on existing NLP tasks and tools such as Question Answering, Summarisation, Information Extraction and Sentiment Analysis (approx. 6%).

Applied research should strongly focus on MT according to around 13% of all respondents. Seen as crucial is thereby, again, the improvement of multilingual resources, data sets and terminology repositories, allowing for standardisation and interoperability (approx. 10%). In addition, there is a demand for improved open-source platforms with a wide range of available systems and applications and truly open and unencumbered data and code repositories (approx. 4%), which are further discussed in Section 5.2.4.

When it comes to innovation the inclusion of all languages and fostering of inter-cultural systems is regarded as a top priority (9%). This also presupposes better and stronger relations between academia and industry (7%). Also stressed is the need to bring together knowledge and methods developed for different fields and domains, e. g., e-health, e-government and e-justice (5%). In addition, there is an interest for more advanced visualisations and interfaces, new innovative tools incorporating NLU and seamless human-computer as well as human-robot interactions (5%).

#### 5.2.4. Applications and Platforms

As for the most important topics, applications and platforms to be integrated (Q24), Machine Translation is uncontroversially the most important one according to approx. 14% of respondents. Considered as almost equally important are the availability of download services for multilingual resources including ontologies, lexicons, dictionaries etc. (approx. 10%). As for further applications a more in-depth development of already existing NLP tools is encouraged, especially speech applications (approx. 10%). Other listed applications include information extraction and retrieval, summarisation, search systems and intelligent assistants.

Among the topics and domains most relevant for the development of future applications and services are education, health, e-participation and e-government (10%).

Regarding the setup of a European Language Technology Data and Service platform and the collaboration between respective stakeholders (Q25), about 30% of all survey answers emphasise the importance of easy accessibility and open licensing for available tools and data. Commonly agreed upon exchange formats and standards also need to be set up. Almost 11% see an involvement of all stakeholders, i. e., data providers, LT providers and LT consumers, as necessary. Effective communication requires a unified, high-level, transparent and user-friendly approach with common goals (approx. 11%). Other recommendations submitted are to facilitate administrative processes on EU level, to adopt best practices from initiatives such as CLARIN[9] and META-NET, to enforce project evaluation processes and to establish business models and commercialisation plans to raise awareness for the ongoing work and the field of Language Technology in general.

### 5.3. Part 3: Talent Generation and Retention

The last part of the survey addresses another challenge Europe's LT community is currently facing, i. e., the constantly increasing brain drain (STOA, 2017). Q26 and Q27 assess the incentives needed for early stage researchers to stay in Europe as well as the skills that are mostly demanded in Language Technology and related fields. In order to best address the skill gap, 74% out of all respondents envision closer collaboration between academia and industry (e. g., through job fairs and hackathons). A large percentage of 62% also sees opportunities in the reorganisation of university curriculums, 43% emphasise the importance of fostering a more enterpreneurial culture through specialised course modules, accelerator programmes etc. Regarded as relevant skills are advanced linguistic knowledge and programming skills (approx. 21%). Linguistic expertise encompasses hereby all disciplines including semantics, syntax, phonetics, formal linguistics, corpus linguistics etc. The most popular programming language among the respondents is Python, closely followed by Java. Considered as the most essential soft skills are collaboration, team work and networking as well as innovative thinking, creativity and proactivity.

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

## 6. Conclusions

The survey has shown that there is a profound common interest and passion not only with regard to Multilingual Europe but also in making the ambitious idea of a large-scale, long-term Human Language Project a reality. A HLP should foster the creation of new approaches, algorithms, data sets and resources which can be employed across modalities, platforms and cultures. With regard to opportunities for research and technology development the three most prominent areas to focus on in the near future Natural Language Understanding, Machine Translation, Educational and Language Learning technologies as well as Deep Learning. Further, the answers emphasise that raising awareness for the Language Technology potential in Europe on a political level is more important now than ever before. The upcoming Brexit and trend of highly qualified researchers emigrating to the US leaves the European Language Technology community in a place where change is needed in order to compete with innovative systems and technologies built and research results produced in the US and elsewhere. Investing into the HLP would secure Europe's place in the pole position in this field for many years to come, solve the threat of Digital Language Extinction and create a truly multilingual Digital Single Market. In addition, it would open up a new direction in the education of young researchers, create attractive jobs for high potentials and foster innovation especially when it comes to new companies. Europe is in the position to shape and claim this topic as its own.

On top of that, the survey inspired plenty of positive comments, for example:

- *"This inspired my brains a lot. Thanks for good questions. I think this is the BEST questionnaire I have ever filled! Good luck with your work! Do not hesitate to contact me if you like to ask or discuss more. I would enjoy continuing in this kind of way, it makes me excited!"*
- *"Human Language Project is an excellent initiative."*
- *"Congratulations for the initiative and the option to include as many answers as possible."*
- *"Best wishes to the survey – this is one of the most important topics for Europe at the present time."*

## Acknowledgements

CRACKER has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357).

## Bibliographical References

Abney, S. and Bird, S. (2010). The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 88–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Georg Rehm et al., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. http://www.meta-net.eu/whitepapers.

Rehm, G. and Uszkoreit, H. (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer Publishing Company, Incorporated.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, Iceland, May.

Rehm, G., Hajic, J., van Genabith, J., and Vasiļjevs, A. (2016a). Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 1586–1592, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., Branco, A., Budin, G., Calzolari, N., Daelemans, W., Garabík, R., Grobelnik, M., García-Mateo, C., van Genabith, J., Hajič, J., Hernáez, I., Judge, J., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Mariani, J., McNaught, J., Melero, M., Monachini, M., Moreno, A., Odjik, J., Ogrodniczuk, M., Pęzik, P., Piperidis, S., Przepiórkowski, A., Rögnvaldsson, E., Rosner, M., Pedersen, B. S., Skadiņa, I., Smedt, K. D., Tadić, M., Thompson, P., Tufiş, D., Váradi, T., Vasiļjevs, A., Vider, K., and Zabarskaite, J. (2016b). The Strategic Impact of META-NET on the Regional, National and International Level. *Language Resources and Evaluation*. 10.1007/s10579-015-9333-4.

Rehm, G. (2015). Strategic Agenda for the Multilingual Digital Single Market – Technologies for Overcoming Language Barriers towards a truly integrated European Online Market, April. Version 0.5. April 22, 2015. Prepared by the EU-funded projects CRACKER and LT_Observatory.

Rehm, G. (2016). Language as a Data Type and Key Challenge for Big Data. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content, July. Version 0.9. July 04, 2016. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded projects CRACKER and LT_Observatory.

Rehm, G. (2017). Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda, November. Version 1.0. Unveiled at META-FORUM 2017 in Brussels, Belgium, on November 13/14, 2017. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded project CRACKER.

STOA. (2017). Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. http://www.europarl.europa.eu/stoa/.

## A Survey Questions

Below we list all 29 survey questions, divided into three main blocks as well as two closing questions.

### A1. Background, Research Interests, Projects

The first 14 questions focus on the demographic background, research interests and projects of the respondents.

- Q1: Personal details
- Q2: What is the name of the organization you work for?
- Q3: What type of organisation do you work for?
- Q4: What is your company's estimated annual revenue in Euro?
- Q5: What is the size of the organisation (total number of employees)?
- Q6: What is your main role in the organisation?
- Q7: What are the day-to-day responsibilities in your role?
- Q8: What are the key research fields, areas and subareas, methods and applications you work on?
- Q9: Which languages do you mainly work with in your research or offer in your products or services?
- Q10: Which languages would you like to include in your research, products or services in addition - but cannot due to a lack of technologies, tools or resources?

- Q11: In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?
- Q12: Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?
- Q13: Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?
- Q14: What is the biggest challenge the European Language Technology community is currently facing?

### A2. Visions for a Future Large-Scale Language Technology Programme

Questions 15-25 focus on the vision of a Language Technology Programme (Human Language Project) in the context of Europe's multilingual challenges and gaps.

- Q15: Do you support the idea of setting up a large-scale Human Language Project?
- Q16: Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.
- Q17: Do you think Deep Natural Language Understanding by 2030 is the right vision and an adequate scientific challenge?
- Q18: Which strategic vision would you suggest instead?
- Q19: How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact?
- Q20: In the context of a HLP, what are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to: a) Basic research, b) Applied research, c) Innovation, d) Industries/Sectors/Verticals
- Q21: Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?
- Q22: Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance
- Q23: How should the Human Language Project be funded?
- Q24: What are, in your opinion, the five key topics,applications, services that must be included in such a platform?
- Q25: Do you have any additional recommendations regarding the setup of the European Language Technology Data andS ervice Platform? For example, regarding the collaboration between data providers, LT providers and LT consumers?

### A3. Part 3: Talent Generation and Retention

Questions 26 and 27 focus on concepts for talent generation and retention in Europe.

- Q26: Which technical or soft skills do you personally consider most important for your specific area/projects?
- Q27: How can the skill gap best be addressed?

### A4. Last but not least

Questions 28 and 29 focus on survey dissemination statistics and final comments.

- Q28: How did you find out about this survey?
- Q29: If you have any additional comments, concerns or suggestions please do not hesitate to share them.