values from the CLARIN licencing framework[27], META-SHARE licences, and the ELRA licence wizard[28]. For all other LRs, a thorough analysis of over 300 licences (all licences in the SPDX list[29]) was done by our legal team who went through the different conditions of use such as the intellectual property rights granted by the licences, the requirements on redistribution imposed by the licence, the requirements on use of the data and, finally, the requirements imposed on users (Rigault et al. 2022b).

7 Language Resources and Data Management

ELG is a platform for commercial and non-commercial Language Technologies, both functional (running services and tools) and non-functional (datasets, resources, models). In order to achieve this, the consortium in charge of the ELG platform has enacted several priorities that include the processing of massive amounts of data and of different types. These large amounts of data derive from partners' contributions, external providers willing to share their datasets through ELG, our harvesting of other repositories as well as different kinds of resource and repository identification work. As can be expected, such a data intensive project requires clear data management policies, in particular considering GDPR constraints. For that purpose, we implemented a Data Management Plan (DMP) as a concrete necessity for organisational, technical and legal management of all data types processed in the course of the project (Rigault et al. 2022a). The DMP documents the variety of data types collected, received and/or processed in the course of the project and reports on how the data is going to be managed with regard to technical, organisational and legal aspects. The DMP also complies with best practices and, in particular, with the requirements of Horizon 2020 as well as GDPR obligations. It defines useful practices to enhance compatibility with the FAIR principles (see Section 7 in Chapter 2 and Wilkinson et al. 2016)[30], as endorsed and specified for Horizon 2020. Moreover, the DMP provides advice in terms of best practices for language resource creation in all steps of an LR life cycle (Choukri and Arranz 2012; Rehm 2016).

8 Conclusions

We integrated more than 10,000 metadata records for datasets, models and other classes of language resources into the ELG platform. These LRTs have been carefully described so as to ease their findability (following the FAIR principles) and to

[27] See https://www.clarin.eu/content/licenses-and-clarin-categories#res and https://www.clarin.eu/content/clarin-license-category-calculator

[28] http://wizard.elra.info/principal.php

[29] https://spdx.org/licenses/

[30] https://www.go-fair.org

ensure compliance with the ELG metadata schema while advocating for interoperability. A series of steps and best practices has been followed with the objective of establishing procedures for resource identification, description and ingestion. The work carried out during the ELG project has allowed us to consider expertise and lessons learned to improve protocols and principles. This has been the reason for updating the integration approach of some repositories (e. g., ELRC-SHARE and Zenodo). The strategy behind the choice of repositories has also been planned carefully, following technical and strategic priorities, as well as evolutionary needs and demands. ELG users can now either access thousands of resources or contribute resources through the different means provided. Legal issues have also been considered with a special focus on licensing. Moreover, a Data Management Plan has been conceived to address the handling of all types of data (including sensitive data) within ELG as well as guiding the production and life cycle aspects of LRs.

References

Arranz, Victoria, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jan Hajic, Ondrej Kosarko, Cristian Berrio, Andrés Garcia-Silva, Rémi Calizzano, Nils Feldhus, Miltos Deligiannis, Penny Labropoulou, Stelios Piperidis, and Ulrich Germann (2021). *Deliverable D5.2 Data Sets, Identified Gaps, Produced Resources and Models (Version 2)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.2-final.pdf.

Arranz, Victoria, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Penny Labropoulou, Miltos Deligiannis, Leon Voukoutis, Stelios Piperidis, and Ulrich Germann (2022). *Deliverable D5.3 Data Sets, Models, Identified Gaps, Produced Resources and their Exploitation within ELG (Version 3)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.3-final.pdf.

Choukri, Khalid and Victoria Arranz (2012). "An Analytical Model of Language Resource Sustainability". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 1395–1402. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/846_Paper.pdf.

Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli (2012). "The META-SHARE Metadata Schema for the Description of Language Resources". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 1090–1097. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3421–3430. URL: https://www.aclweb.org/anthology/2020.lrec-1.420/.

Lösch, Andrea, Valérie Mapelli, Khalid Choukri, Maria Giagkou, Stelios Piperidis, Prokopis Prokopidis, Vassilis Papavassiliou, Miltos Deligiannis, Aivars Berzins, Andrejs Vasiljevs, Eileen Schnur, Thierry Declerck, and Josef van Genabith (2021). "Collection and Curation of Language Data within the European Language Resource Coordination (ELRC)". In: *Proceedings of the Conference on Digital Curation Technologies (QURATOR 2021)*. Ed. by Adrian Paschke, Georg Rehm, Jamal Al Qundus, Clemens Neudecker, and Lydia Pintscher. Vol. 2836. CEUR Workshop Proceedings. Berlin, Germany: CEUR-WS.org. URL: http://ceur-ws.org/Vol-2836/qurator2021_paper_6.pdf.

Mapelli, Valérie, Victoria Arranz, Hélène Mazo, and Khalid Choukri (2022). "Language Resources to Support Language Diversity – the ELRA Achievements". In: *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 551–558. URL: http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.58.pdf.

Piperidis, Stelios, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi (2014). "META-SHARE: One year after". In: *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: ELRA, pp. 1532–1538. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/786_Paper.pdf.

Rehm, Georg (2016). "The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources". In: *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: ELRA, pp. 2450–2454. URL: https://aclanthology.org/L16-1388.pdf.

Rehm, Georg and Katrin Marheinecke (2019). *Deliverable D7.2 National Competence Centres and Language Technology Council*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2021/02/ELG-Deliverable-D7.2-final.pdf.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rigault, Mickaël, Victoria Arranz, Khalid Choukri, Valérie Mapelli, Pawel Kamocki, and Lucille Blanchard (2022a). *Deliverable D5.6 Data Management Plan (Version 3)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D5.6-final.pdf.

Rigault, Mickaël, Victoria Arranz, Valérie Mapelli, Penny Labropoulou, and Stelios Piperidis (2022b). "Categorizing Legal Features in a Metadata-Oriented Task: Defining the Conditions of Use". In: *Proceedings of the Legal and Ethical Issues Workshop (LREC 2022)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 22–26.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18. URL: http://www.nature.com/articles/sdata201618.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2020). "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Yeganova, Lana, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de-Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes (2021). "Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set". In: *Proceedings of the Sixth Conference on Machine Translation*. ACL, pp. 664–683. URL: https://aclanthology.org/2021.wmt-1.70.

Chapter 9
Language Technology Companies, Research Organisations and Projects

Georg Rehm, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou

Abstract The European Language Grid is meant to develop into the primary platform of the European Language Technology community. In addition to LT tools and services (Chapter 7) and Language Resources (Chapter 8), ELG represents the actual members of this community, i. e., the companies and research organisations that develop language technologies and that are engaged in related activities. The goal of becoming the primary platform for LT in Europe implies that ELG should ideally represent *all* European companies and *all* European research organisations with corresponding metadata records in the ELG catalogue, which are interlinked with the respective LT tools and services as well as language resources they offer. This chapter describes the European stakeholders and user groups that are relevant for the ELG initiative, the composition of the community and the locations of the companies and research groups as currently listed in ELG. Furthermore, we describe a number of technical and organisational challenges involved in the preparation of our list of stakeholders, and outline the process of catalogue population.

1 Introduction

The European Language Grid is meant to develop into the primary platform of the European LT community. This is why, in addition to functional LT tools and services and more static Language Resources (LRs), ELG also represents the actual members of this community, i. e., the companies and research organisations that develop LTs and that are engaged in related activities such as the integration of LT into existing systems or support services such as data annotation at scale. This overall goal of eventually establishing ELG as the primary platform for LT in Europe implies that ELG should ideally represent *all* European companies and *all* European research

Georg Rehm · Katrin Marheinecke · Rémi Calizzano
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de, katrin.marheinecke@dfki.de, remi.calizzano@dfki.de

Penny Labropoulou
Institute for Language and Speech Processing, R. C. "Athena", Greece, penny@athenarc.gr

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_9

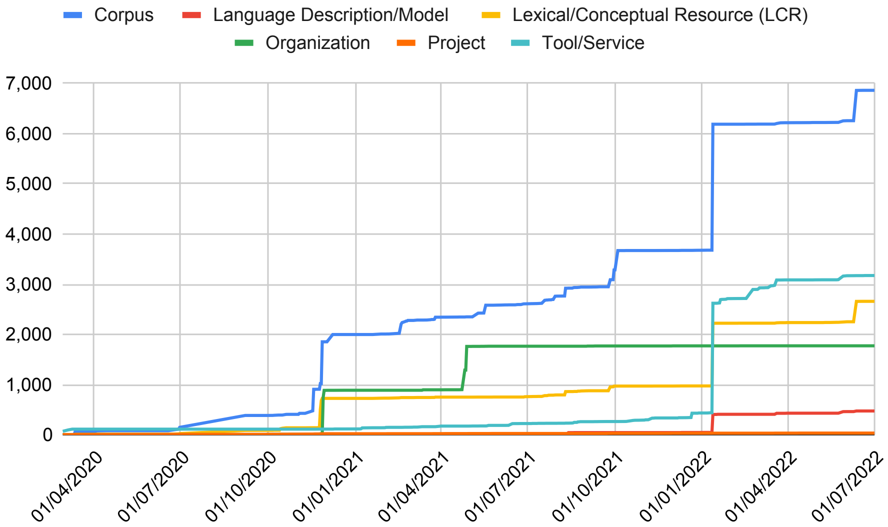


Fig. 1 Evolution of resources in ELG over time broken down by resource type

organisations in the ELG catalogue, which are interlinked with the respective LT tools and services as well as language resources these organisations offer on and through the European Language Grid. In other words, the European Language Grid also functions as the "yellow pages" of the European LT community, ideally listing and promoting *all* relevant members of this community, i. e., small and medium-sized companies as well as large enterprises, research centers, universities and other academic institutions that develop LT but also organisations in the periphery of this core, e. g., integrators and annotation service providers (Rehm et al. 2020, 2021).[1]

In addition to serving as the central directory for members of the European LT community, ELG also includes information about relevant projects in the area.[2] The reasoning behind this is the way many LTs are typically developed, i. e., through publicly funded project consortia in which academic or commercial organisations participate. Such projects often result in concrete tools and technologies as well as language resources, which can then be made available, among others, through ELG, which allows representing and interlinking these project artefacts (LTs, LRs), the projects that helped create these artefacts and the members of the respective project consortia. Technically, project consortia can provide relevant metadata to create and later edit and update their own project pages in ELG ensuring more visibility as well as an additional dissemination channel for their projects' outputs.

In the second half of the ELG project's runtime, corresponding activities in terms of populating the ELG catalogue with information about companies, academic organisations and projects have been drastically increased so that, towards the end of the project, ELG now includes convincing figures in terms of community members,

[1] https://live.european-language-grid.eu/catalogue/?entity_type__term=Organization

[2] https://live.european-language-grid.eu/catalogue/?entity_type__term=Project

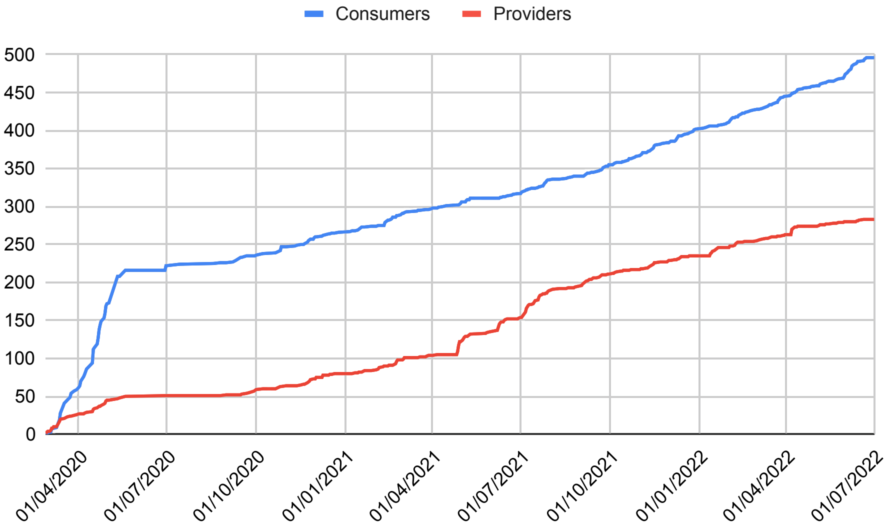


Fig. 2 Number of ELG consumer and provider accounts over time

projects and also active users of the platform. At the time of writing, ELG lists more than 13,000 metadata records on tools and services, resources, organisations and projects. Figure 1 shows the corresponding development of the ELG catalogue and its population over time, differentiated by type of entry.

Not only the number of resources and organisations listed in ELG is constantly growing. In addition, the number of users is rising continuously. The number of ELG users of the consumer category who have a registered a user account went up significantly at the end of April 2020, after the first official release to the public, and has grown further ever since. The number of ELG users of the provider category, i. e., users with the right to integrate metadata, tools and resources in ELG, is also increasing continuously, albeit more slowly, as can be expected (see Figure 2).

As encouraging as this development is, ELG is still at the beginning. The platform has been designed in such a way that it can be actively used by the community and that it can grow. To achieve this goal of a true one-stop shop for the whole European LT community, it is necessary to steadily expand the consumer and provider base and monitor as well as reflect all changes and new developments in the European LT landscape. Only with this momentum will the desired snowball effect be generated eventually, which ultimately helps ELG to achieve sustainable success from which all stakeholders can benefit.

2 The European Language Technology Landscape

One key characteristic of the European Language Technology landscape is its extreme fragmentation, which has been mentioned repeatedly throughout the years, as, for example, in the META-NET White Paper Series (Rehm and Uszkoreit 2012), in the META-NET Strategic Research Agenda (Rehm and Uszkoreit 2013; Rehm et al. 2016), in the *Final study report on CEF automated translation value proposition in the context of the European LT market/ecosystem* (Vasiljevs et al. 2019) or in the various reports of the European Language Equality project (especially see Aldabe et al. 2022). In fact, this extreme fragmentation is one of the main reasons why the ELG platform has been developed in the first place because the fragmentation is generally perceived as one of the main reasons why the European LT community has been unable to unleash its full potential.

The analysis in the CEF LT Market study (Vasiljevs et al. 2019) shows that European LT vendors are often SMEs with local or regional, often highly specialised solutions. In the study, 473 companies were collected that are active in EU member states in the domain of LT and that fully qualify as LT vendors. According to the research, the total size of the LT industry within the EU member states (plus Iceland and Norway) was estimated at approx. 800M€ in the year 2017. In the study sample investigated, only 14% of the LT vendors had a revenue of more than €10M, whereas almost half of them (48%) had a revenue below €1M. In terms of size, 52% of the companies had between 10 and 99 employees, and 26% had less than 10 employees, both combined representing nearly 80% of the 473 companies studied. Only 44% of the EU companies in this sample received external funding or venture capital.

Consequently, the global LT and NLP market continues to be dominated by large technology enterprises from the United States and Asia which establish “data-driven intellectual monopolies” (Rikap and Lundvall 2020) – in that regard, large companies are the exception in Europe. However, these big non-European LT providers have certain deficiencies regarding under-resourced languages, customisation needs, as well as security and privacy requirements which is a frequently expressed demand from corporate clients and European administrations (Overton 2017).

Despite the fact that the LT market is relatively small when compared to the general IT market at large, it is a market with strong competition, which is one of the reasons why many LT developing companies tend to focus on highly specialised niche markets with less intense competition. This, however, affects profitability, which is, on average, rather low and margins are compressed. On the other hand, LT can also be considered a growing market: today, (potential) customers have more awareness of the benefits of LT, which is also due to marketing activities of large international players. From a local vendors’ point of view, the large technology enterprises help create a market awareness that simply did not exist ten years ago. Nevertheless, these companies are also the toughest competition of the European LT community as they tend to offer high-quality LT software free of charge or for very low prices, which European SMEs usually cannot afford to do.

The STOA study *Language equality in the digital age – Towards a Human Language Project* (STOA 2018), which examines the causes of language barriers in

Europe and formulates recommendations for policies to overcome these barriers, mentions among its 11 key recommendations the need for a pan-European LT Platform of resources and services and ELG has stepped up to solve this problem (also see European Parliament 2018). ELG not only brings together LT resources from all over Europe supporting almost all European languages (although ELG is not limited to European languages) but ELG also has the ambition to unite the European LT community behind these services, tools and resources using one shared umbrella platform to create a common access point and marketplace from which all languages and members of the community will eventually benefit (see Part III of this book).

At the time of writing, ELG contains approx. 1,800 organisations operating in the European LT sphere. One half of these organisations consists of companies, the other half of universities and research groups (Figure 3).[3]

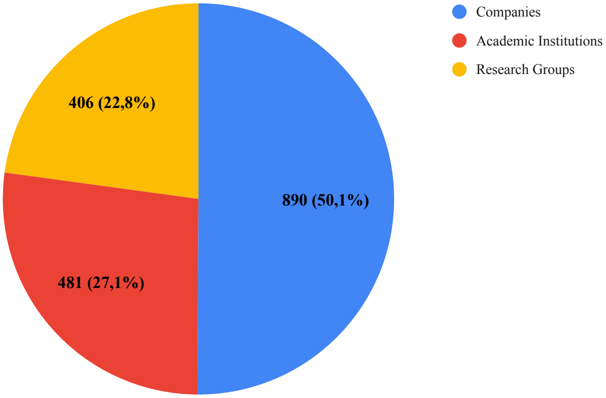


Fig. 3 Distribution of organisations listed in ELG per type

The quantitative distribution of LT developing organisations among the respective countries in Europe already hints at a strongly varying coverage of LT resources for their respective national and regional languages. Whereas countries like the UK, Germany or Spain are well or relatively well equipped with LT developing companies, smaller countries like Malta or Cyprus have only little representation in the European LT community (see Figure 4).[4] Figure 5 shows the geographical distribution in Europe of organisations listed in ELG.

[3] Companies are commercial organisations, academic institutions are universities and research centers, research groups are sub-groups of academic institutions, e. g., faculties or departments.

[4] In Figure 4, countries are ordered by decreasing number of organisations. The country with the head office of the respective organisation is used as the organisation's country.

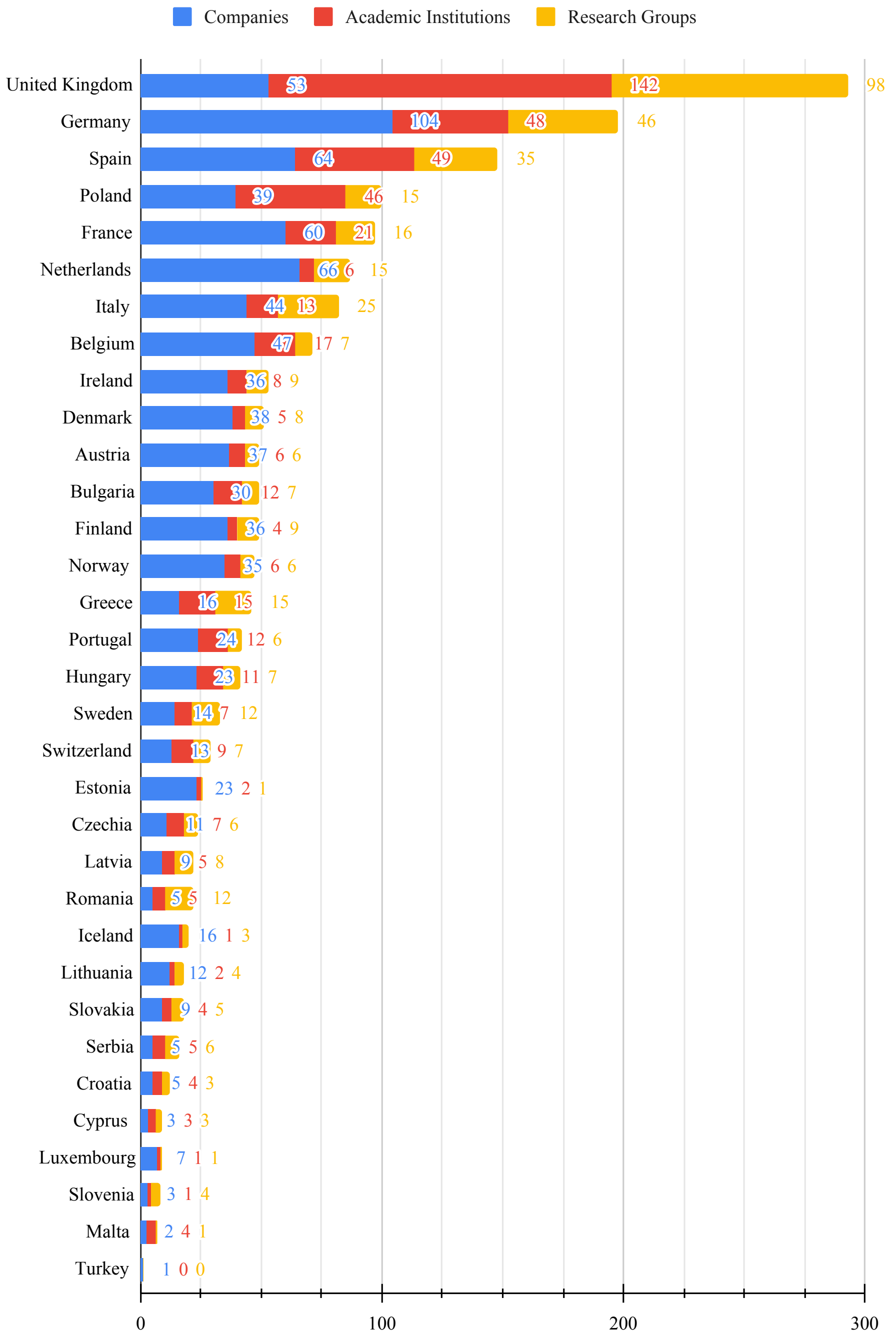


Fig. 4 Distribution of organisations listed in ELG per type and country

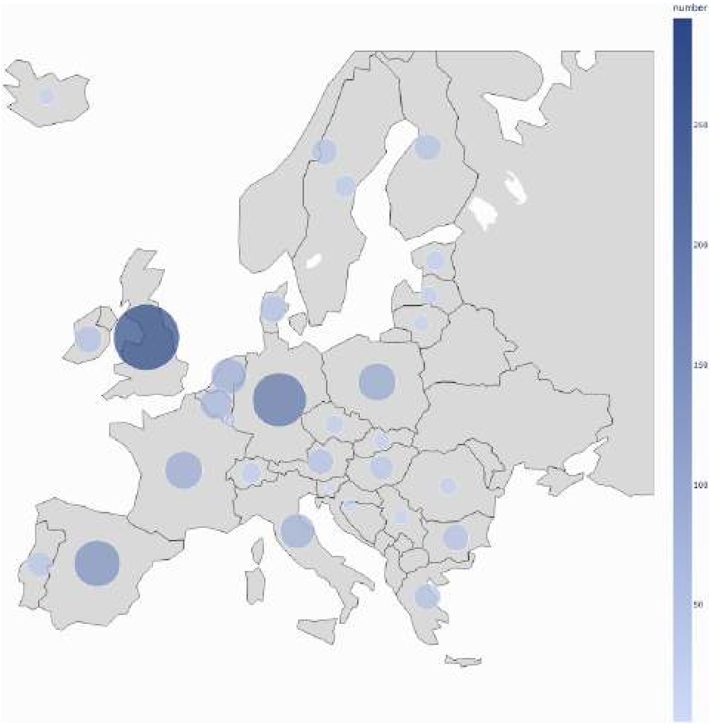


Fig. 5 Organisations listed in ELG per country

3 Organisations in the European Language Grid

To bootstrap the ELG catalogue with as many LT developing European companies and academic organisations as possible, we decided on the following procedure. First, together with the ELG National Competence Centres (see Chapter 11, p. 205 ff.), we collected LT developing organisations semi-automatically and in a decentralised way, i. e., on the national level (Section 3.1). Second, based on the results of this collection, metadata records were prepared that could be automatically ingested into the ELG catalogue (Section 3.2). This resulted in the ELG catalogue being populated with approx. 1,800 metadata records, i. e., pages, each of which describes one LT developing organisation with a basic profile. These organisation profiles can then be claimed by the rightful owners (Section 3.3), i. e., an organisation described in such an ELG page can take over the maintenance of its own page and enrich it with additional information, e. g., upload a logo, associate resources with their organisation etc. (Section 3.4). This bootstrapping procedure enables members of the European LT community to participate actively in ELG with their own organisation within minutes. As a positive side effect, it enabled ELG – including its sister project ELE – to produce a fairly detailed picture of the European LT landscape.

3.1 Collecting the Members of the European LT Community

In order to populate ELG with organisations, we used our own databases, carried out desk research and, most importantly, we involved the 32 National Competence Centres (NCCs) to tap into their detailed knowledge of their respective countries' LT communities. Our general goal was to identify and to record, in a machine-readable format, as many national and regional members of the European LT community as possible so that ELG can eventually provide as complete and up to date a picture as possible. In September 2020, this data collection task was conducted with NCC Leads representing their countries and regions to ideally identify all companies and academic organisations in the European LT community to be listed in ELG.

To streamline the process, based on data gathered in various workshops, conferences and other events over the last ten years, the ELG project team created lists of organisations involved in LT activities in all European countries. Each entry in the list contained, among others, the following information: organisation name, department name, website, address (region, ZIP code, city, country) and LT areas in which they are active. Each NCC Lead received the data records for their country, along with detailed guidelines, and they were asked to check the data included in the list, to correct the data if necessary (e. g., remove duplicates with similar names, correct wrong names of organisations) and to complete them where possible, i. e., to fill in blanks. Furthermore, the NCCs were asked to do their own research and provide new, unlisted organisations. The goal was to find all relevant organisations of each country that develop, market or sell LT in their countries. This way, the ELG consortium wanted to ensure that in addition to well-known orgnaisations also start-ups andyoung research groups are included in ELG.

The feedback received from the NCCs was submitted to a comprehensive internal quality review by the ELG team, which resulted in the final dataset that reflects a fairly complete representation of the relevant stakeholders and providers of Language Technology and language-centric AI in Europe.[5]

3.2 Preparation and Integration of Metadata Records

The efforts of the NCCs and the ELG team for the collection of data regarding LT organisations relevant for ELG resulted in two spreadsheets per country containing companies and research groups respectively. All entries were automatically converted into XML files that are compliant with the ELG metadata schema as described in Chapter 2. Furthermore, for columns corresponding to metadata elements that take values from controlled vocabularies (e. g., LT area), we mapped the input to the values in the controlled vocabulary. This process also served as a sanity check during

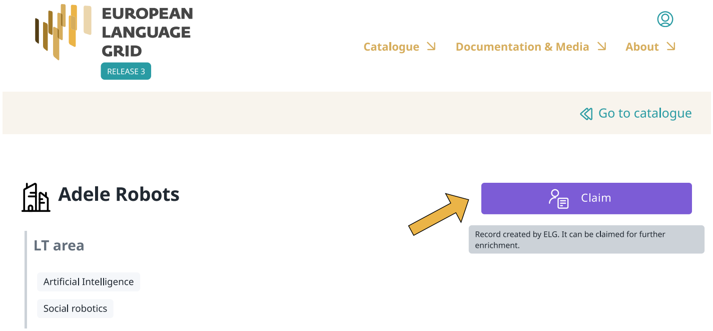
[5] In this procedure, the regulations of the Data Protection Act were adhered to at any time and no personal data have been published without the consent of the data owners.

which errors were identified and resolved. The procedure resulted in 1,740 XML files, 867 for companies and 873 for research groups.

The ELG life-cycle for the publication of individual resources includes a validation process aiming to ensure the quality of the metadata published in ELG (see Chapter 2). For the import of the organisation-related XML files, we applied a different procedure that involved their bulk import with the assignment of the tag "imported by ELG". Metadata records marked as such do not go through a validation process and are immediately published on ELG.

3.3 Claiming and Enriching Organisation Pages

Once the population of ELG with these entries was completed, a campaign was launched inviting (via email) legitimate owners to claim, edit and curate the entries of their own organisations. Since the pages created by the ELG team contained only minimal information, the representatives of the organisations were invited to enrich these pages with reliable and accurate content and also to start providing tools, services and resources. In several email campaigns, we reached out to contact persons identified by the NCCs and we informed them about the existence of their organisations' pages on ELG, also inviting them to take over the pages. To do so, the legitimate owner can "claim" their organisation's page as their own by clicking the "Claim" button on the page (see Figure 6).



Record created by ELG. It can be claimed for further enrichment.

Fig. 6 Imported organisation page with a "Claim" button

The claiming process can only be triggered by persons signed in with an ELG account (with provider role). This step serves as a security mechanism ensuring correct and rightful authorisation of eligible persons. Once a request is made, the ELG team checks its validity, which also includes checking the email address used to register the ELG account, making sure that it belongs to the organisation, the page of which is being claimed. Approval of the request entails that the entry is assigned to the claimant and returns to a status that it can be edited. The claiming person is prompted by email that they can now start editing the metadata entry and ELG page. Once edited, the page needs to be submitted to publication and the usual ELG validation process starts, i. e., the changes made to the resource are reviewed by the ELG team and the entry is made publicly available again.

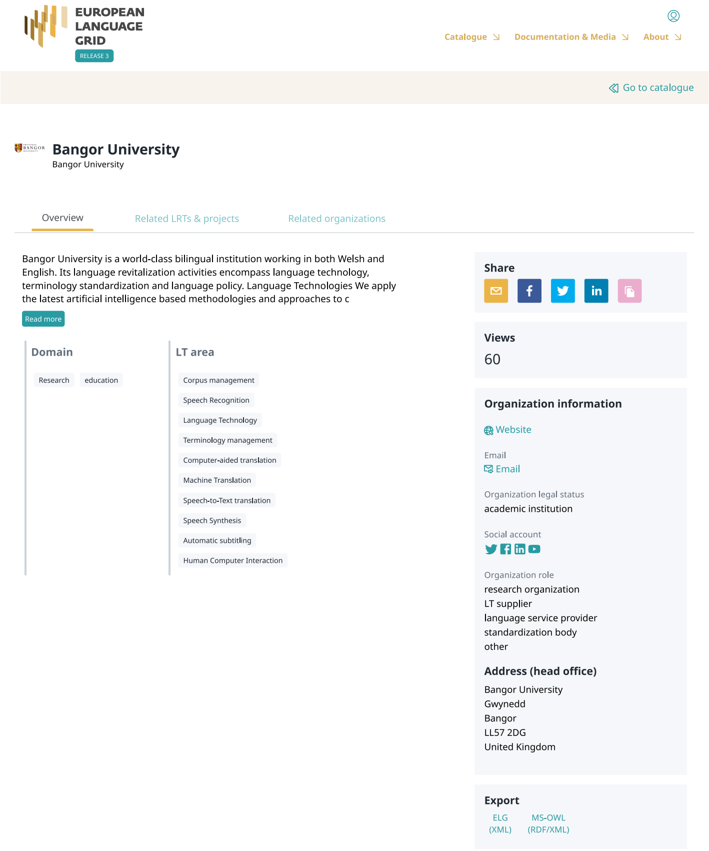
3.4 Organisation Pages in the European Language Grid

Organisation pages can include different tabs. The "Overview" tab includes a description of the organisation as well as an info box on the right with data such as postal address and contact email as well as a link to the organisation's own website. This tab can also include keywords that describe the general domain and LT areas an organisation addresses. ELG pages can also be exported in XML format. The tab "Related LRTs & projects" lists all resources and technologies the respective organisation has made available on ELG and the projects they are involved in. This helps companies to promote their tools and resources and to show connections between companies or research organisations and their research projects and corresponding results. The "Related organisations" tab is especially important for academic institutions and universities to reflect their relationship to other departments, faculties or the umbrella organisation (usually the university). Figure 7 provides an example for a page of an academic organisation. Figure 8 (p. 182) shows a company page.

4 Projects in the European Language Grid

ELG is also able to represent research projects, especially for the purpose of acknowledging the funding that made the development of a technology or resource possible and also to interlink projects with organisations and resources.[6] ELG project pages are structured in a similar way, but they are especially adapted to the characteristics and metadata of a typical research project. In addition to information regarding the start and end of the project, the info box also contains details on the funding agency, the funding country, the type of project and the amount of funding provided. Besides the project description and keywords, the "Overview" tab contains the list of consortium partners, that are linked to their respective ELG pages if they exist. Again, the tab "Related LRTs" lists all technologies and resources associated with or resulting from the project. Two examples are shown in Figures 9 (p. 183) and 10 (p. 184).

[6] At the time of writing, we are preparing a list with more than 500 projects that will be imported into the ELG catalogue in the second half of 2022; this list was put together in a similar manner as the list of organisations described in Section 3.1.



Bangor University
Bangor University

Bangor University is a world-class bilingual institution working in both Welsh and English. Its language revitalization activities encompass language technology, terminology standardization and language policy. Language Technologies We apply the latest artificial intelligence based methodologies and approaches to c

Domain
Research
education

LT area
Corpus management
Speech Recognition
Language Technology
Terminology management
Computer-aided translation
Machine Translation
Speech-to-Text translation
Speech Synthesis
Automatic subtitling
Human Computer Interaction

Views
60

Organization information
Website
Email
Email
Organization legal status
academic institution
Social account
Organization role
research organization
LT supplier
language service provider
standardization body
other

Address (head office)
Bangor University
Gwynedd
Bangor
LL57 2DG
United Kingdom

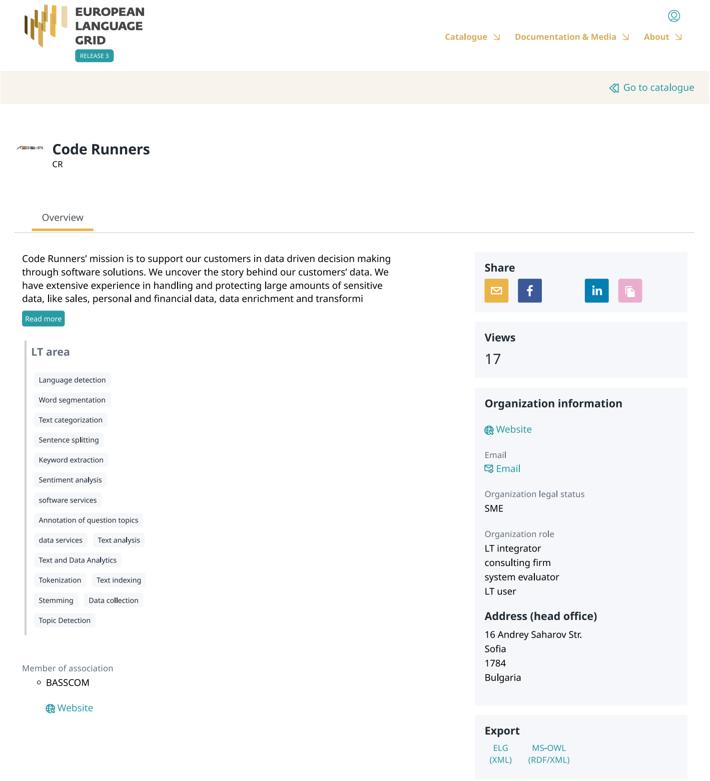
Export
ELG (XML)
MS-OWL (RDF/XML)

Fig. 7 Example ELG organisation page: Bangor University

5 Conclusions

The European Language Grid is meant to develop into the primary platform of the European LT community. In addition to the technical resources, ELG also represents the actual members of this community: companies and research organisations that develop LTs and related organisations. Our ambition is for ELG eventually to represent *all* companies and *all* research organisations active in the European LT community. In order for ELG to function as a marketplace for European LT, it also needs to provide core information about the European LT community (i. e., “yellow pages” functionalities).

ELG currently contains approx. 1,800 organisations active in the European LT community. Like every similar repository or digital catalogue with certain artefacts, one of the key challenges is the maintenance of the records and metadata entries, i. e., keeping the entries up to date and also making sure that the community is fully



Code Runners
CR

Code Runners' mission is to support our customers in data driven decision making through software solutions. We uncover the story behind our customers' data. We have extensive experience in handling and protecting large amounts of sensitive data, like sales, personal and financial data, data enrichment and transformi

LT area

Language detection
Word segmentation
Text categorization
Sentence splitting
Keyword extraction
Sentiment analysis
software services
Annotation of question topics
data services
Text analysis
Text and Data Analytics
Tokenization
Text indexing
Stemming
Data collection
Topic Detection

Member of association

- BASSCOM

Website

Views
17

Organization information

Website

Email
Email

Organization legal status
SME

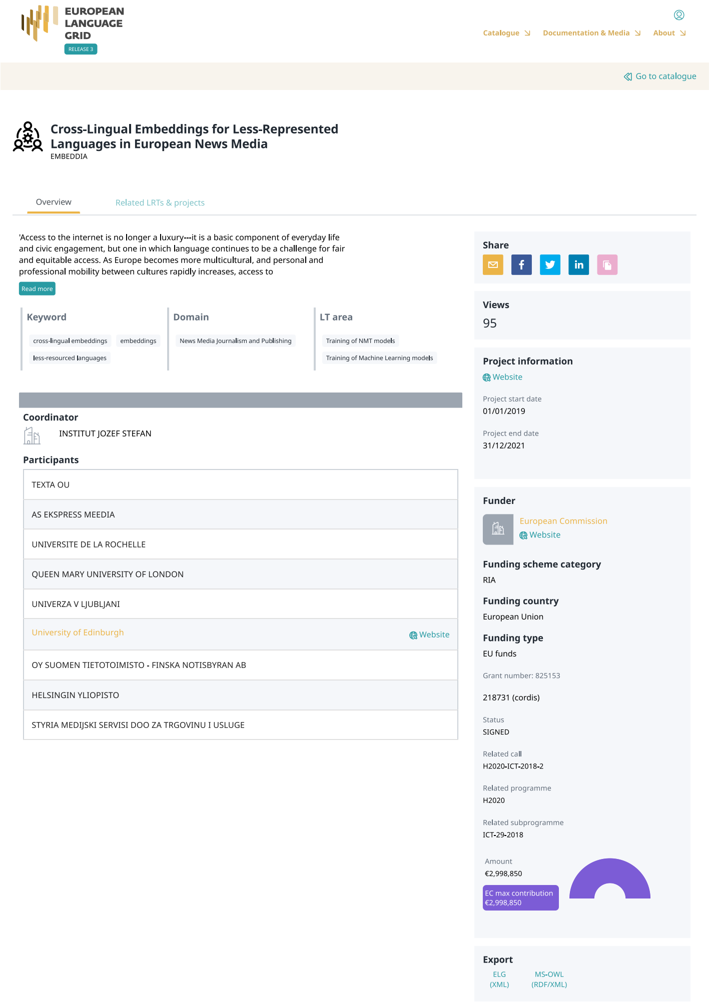
Organization role
LT integrator
consulting firm
system evaluator
LT user

Address (head office)
16 Andrey Saharov Str.
Sofia
1784
Bulgaria

ELG (XML)
MS-OWL (RDF/XML)

Fig. 8 Example ELG organisation page: Code Runners

represented. Our long-term vision for ELG is to become the primary platform of the European LT community, which entails that *all* members of the European LT community, both commercial and academic, immediately recognise the value, importance and relevance of ELG and, thus, actively want to participate in ELG, keepting their pages up to date, sharing technologies and resources, benefiting from this European marketplace. Until this intended snowball effect is fully in place, i. e., all stakeholders recognise the benefit ELG brings about and participate actively, we will perform, even if time-consuming and logistically challenging, manual updates of the ELG catalogue, we will continue to convert as many members of the community as possible into active users and also active providers of ELG and we will increase our our outreach activities, encouraging more organisations to claim their ELG pages. As soon as the snowball effect is in place and ELG is accepted as the primary platform of the European LT community, all participating organisations will have a sufficient amount of intrinsic motivation to maintain their ELG pages and to keep their information, technologies and resources up to date. At this time, ELG strives to be an established player, which is known throughout the community so



Cross-Lingual Embeddings for Less-Represented Languages in European News Media
EMBEDDIA

'Access to the internet is no longer a luxury---it is a basic component of everyday life and civic engagement, but one in which language continues to be a challenge for fair and equitable access. As Europe becomes more multicultural, and personal and professional mobility between cultures rapidly increases, access to

Keyword	Domain	LT area
cross-lingual embeddings embeddings less-resourced languages	News Media Journalism and Publishing	Training of NMT models Training of Machine Learning models

Coordinator
INSTITUT JOZEF STEFAN

Participants

TEXTA OU	
AS EKSPRESS MEEDIA	
UNIVERSITE DE LA ROCHELLE	
QUEEN MARY UNIVERSITY OF LONDON	
UNIVERZA V LJUBLJANI	
University of Edinburgh	Website
OY SUOMEN TIETOTOIMISTO - FINSKA NOTISBYRAN AB	
HELSINGIN YLIOPISTO	
STYRIA MEDIJSKI SERVISI DOO ZA TRGOVINU I USLUGE	

Views
95

Project information
Website

Project start date
01/01/2019

Project end date
31/12/2021

Funder
European Commission
Website

Funding scheme category
RIA

Funding country
European Union

Funding type
EU funds

Grant number: 825153

218731 (cordis)

Status
SIGNED

Related call
H2020-ICT-2018-2

Related programme
H2020

Related subprogramme
ICT-29-2018

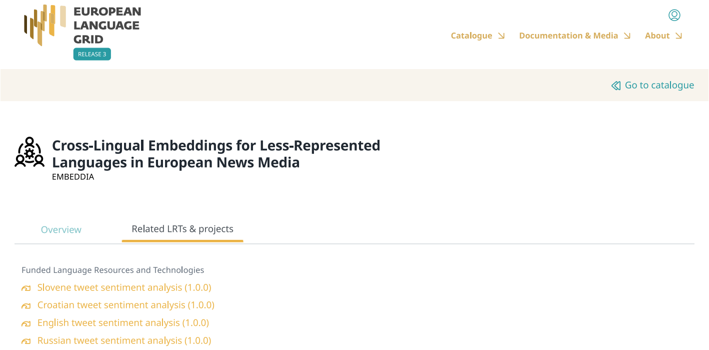
Amount
€2,998,850

EC max contribution
€2,998,850

Export
ELG (XML) MS-OWL (RDF/XML)

Fig. 9 Example ELG project page: EMBEDDIA (Overview)

that also new companies are attracted by and to ELG. In addition to simplifying the claim process, the attractiveness of ELG will be further enhanced through increased community-related promotions, new features and improved offerings.



Cross-Lingual Embeddings for Less-Represented Languages in European News Media
EMBEDDIA

Funded Language Resources and Technologies

- Slovene tweet sentiment analysis (1.0.0)
- Croatian tweet sentiment analysis (1.0.0)
- English tweet sentiment analysis (1.0.0)
- Russian tweet sentiment analysis (1.0.0)

Fig. 10 Example ELG project page: EMBEDDIA (Related LRTs)

References

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE. URL: https://european-language-equality.eu/wp-content/uploads/2022/06/ELE___Deliverable_D3_1__second_revision_2.pdf.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. URL: http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Overton, David (2017). *Next Generation Internet Initiative – Consultation*. URL: https://ec.europa.eu/futurium/en/system/files/ged/ec_ngi_final_report_1.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings* *of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg, New York, Dordrecht, London: Springer. URL: http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: *Language Resources and Evaluation* 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4. URL: http://link.springer.com/article/10.1007/s10579-015-9333-4.

Rikap, Cecilia and Bengt-Åke Lundvall (2020). "Big Tech, Knowledge Predation and the Implications for Development". In: *Innovation and Development*, pp. 1–28. DOI: 10.1080/2157930X.2020.1855825.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. URL: https://data.europa.eu/doi/10.2861/136527.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. URL: https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Part III
ELG Community and Initiative

Chapter 10
European Language Technology Landscape: Communication and Collaborations

Georg Rehm, Katrin Marheinecke, and Jens-Peter Kückens

Abstract The European Language Technology community is a diverse group of stakeholders that is characterised by severe fragmentation. This chapter provides an overview of the stakeholders that are relevant for the European Language Grid. We also briefly describe our communication channels and strategies with regard to the promotion of ELG. Furthermore, we highlight a few of the current projects and initiatives and their relationship to and relevance for ELG, especially with regard to collaborations. The overall goal of the target group-specific communication strategy we developed is to create more and more uptake of ELG in the European LT community, eventually creating a snowball effect.

1 Introduction

A key challenge to which ELG aims to respond is the ubiquitous fragmentation of the European LT landscape. ELG addresses this problem by bringing together all European stakeholders under one umbrella platform (European Parliament 2018). While Chapter 9 (p. 171 ff.) provides a high-level description of the LT companies, research organisations and projects registered in ELG at the time of writing (including statistics etc.), the present chapter focuses upon the stakeholder groups themselves. The challenge of severe fragmentation (STOA 2018) has been taken up in ELG from the very beginning on different levels by implementing various communication and cooperation activities. Their aim has been to make ELG known in all relevant communities within a short time in such a way that companies and research organisations develop an active interest in ELG: the more providers offer high quality and attractive services and datasets, the faster ELG will become a central marketplace, which in turn will benefit providers and users alike. This is why the ELG consortium pursued a strategy through which the communication activities in combination with the high quality of the platform and its services and resources, as well as fast and reliable

Georg Rehm · Katrin Marheinecke · Jens-Peter Kückens
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de, katrin.marheinecke@dfki.de, jens_peter.kueckens@dfki.de

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_10

support services, produce this desired snowball effect. Some of the communication and cooperation areas and activities were:

National Competence Centres (NCCs) Establish a network of 32 carefully selected National Competence Centres (see Chapter 11, p. 205 ff.).

ICT-29b) Projects Cooperate with the six EU projects funded through the call ICT-29b), i. e., Bergamot[1], Comprise[2], ELITR[3], Embeddia[4], GoURMET[5], Prêt-à-LLOD[6] and their consortia and networks.

Major European Initiatives Collaborate with all relevant major European initiatives including, among others, the European AI on Demand Platform[7], CLAIRE[8], HumanE AI Net[9], CLARIN[10] and others (see Chapter 2, Section 8, p. 27 ff., as well as Chapter 6, p. 107 ff.). These collaborations are described in more detail in Section 4 (p. 199 ff.) of the present chapter.

Events Organise local, regional and national events together with the ELG National Competence Centres (see Chapter 11, p. 205 ff.).

Talks and Presentations Give talks and presentations, especially at networking and outreach events, to decision-makers and multipliers, both in the industrial sector but also in scientific European conferences.

Open Calls and Pilot Projects Selected 15 pilot projects, with which we also collaborated in terms of communication activities on their respective regional and local levels (see Part IV, p. 256 ff., of this book).

Next up, Section 2 describes the European Language Technology community in more detail, focusing upon the different stakeholder groups. A key driver of success of the ELG initiative is this support and buy-in from the stakeholder community including the uptake of the platform. In addition to these networking activities in the project, several public communication channels have been established. Under the umbrella brand "European Language Technology", ELG and its sister project European Language Equality (ELE, see Rehm and Way 2023) have started addressing the stakeholders and initiatives listed above, community members outside these networks and the wider public in order to provide them with news about relevant project developments, events and updates on ELG features, among others. For this purpose, social media profiles on Twitter and LinkedIn were established. We also set up an email newsletter, which was initially published on a monthly basis and later on changed to a biweekly schedule. These communication channels, their purpose, effectiveness and the content shared through them is further detailed in Section 3.

[1] https://browser.mt

[2] https://www.compriseh2020.eu

[3] https://elitr.eu

[4] http://embeddia.eu

[5] https://gourmet-project.eu

[6] https://pret-a-llod.github.io

[7] https://www.ai4europe.eu

[8] https://claire-ai.org

[9] https://www.humane-ai.eu/workpackages/

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

2 Stakeholders of the European Language Grid

For our main groups of stakeholders we defined their specific relationships with ELG and how we will communicate and engage with them in terms of communication channels but also in terms of messages, considering our overall communication goals. Most target groups also interact with ELG in one way or another, which is why they are, in most cases, not just passive audiences but also immediate stakeholders within the ELG community (Rehm et al. 2020c, 2021). In the following, all stakeholder groups are defined including aspects relating to communication.

2.1 Language Technology Providers

The interests of LT *providers* (see Chapter 9 for specific numbers) are different from those of LT *users*, which is why specific communication formats need to be applied. Typically, commercial providers of LT (also see Vasiljevs et al. 2019) want to showcase their products and promote their solutions and services or their company and – on a more abstract level – they look for an appropriate marketplace in which they can participate. In contrast to other target groups, their interactions with ELG are active and direct. In order to upload or offer a service or tool via ELG, they need specific technical information and an understanding of how ELG works. This demand is met through various forms of communication, including a technical documentation with clear and in-depth explanations of ELG's functionalities, based on which video tutorials were prepared. These videos are shared through all communication channels. Furthermore, blog articles explain specific ELG features to LT providers and short announcements of new features are included in the newsletter.

The more ELG meets business requirements, the more likely LT providers are to actively use and promote it and to exploit it as an additional or sales channel or even as their preferred marketplace. Our communication activities addressing LT providers uses a marketing tonality and promotes the advantages of the ELG initiative. We have also reached out repeatedly to LT providers, inviting them to send in their questions or feedback with regard to their experience with ELG, including missing features or suggestions for improvement.

In many cases research centres and universities are also LT providers, but their interest is usually not a monetary but a research-driven one. This stakeholder group provides larger or smaller datasets or perhaps tools or rudimentary, experimental services that have evolved from research projects rather than robust, production-ready services that can be directly monetised. For researchers, sharing their results, the further development of tools and the exchange with other researchers is the main driver to use ELG. Finding datasets and tools bundled in one place, they can test functionalities in the development phase and provide feedback. Ideally, they spread the word about ELG in scientific articles or in academic conferences, and they can be approached most easily through these channels. Public communication about the usefulness of an easy-to-use platform for hosting, sharing and making available LT services has proven an effective measure to attract researchers and initiate direct communication about ELG.

2.1.1 Participants in the Open Calls – Pilot Projects

ELG tested the platform and demonstrated its usefulness with the help of 15 pilot projects that it supported financially (see Part IV for more details). After their completion, the results were fed back into the ELG platform and community. To attract companies or research centres to submit proposals and to make the selected pilot projects known, communication activities were necessary. The open calls were published and advertised through email campaigns, through the ELG website and on multiple events. META-FORUM 2019 was the first public occasion where the open calls have been publicly presented. This target group had a high demand for information, therefore different channels like online trainings, videos, fact sheets and news articles were implemented. The pilot projects were an important measure and instrument to make ELG known to a wider audience, communication in this area had to be especially effective. The overwhelming response with more than 200 project proposals in total proves that this strategy has worked out and the successful completion of all 15 selected pilot projects is evidence of successful communication (and a thorough evaluation of the proposals). The results of the pilot projects were also presented in the virtual project expos at META-FORUM 2020, 2021 and 2022 and also in a number of sessions and presentations.[11]. Several pilot projects were showcased on the ELG blog, presenting their activities but also the greater implementations of making use of a pan-European LT platform, while the promotional videos created for META-FORUM 2021 were featured in the newsletter and on social media.

2.2 Language Technology Users

The users of Language Technology are the most diverse and also, by a margin, the biggest target group. Users include almost everyone – from students doing research for a paper to job seekers in the LT field, to companies looking for a machine translation solution for the corporate website, just to mention a few examples. Members of this group can look for information, try to find certain LT services or datasets or they can be potential buyers or integrators of LT. This enormous group interacts with ELG in the form of a user, consumer or potential customer (Rehm et al. 2021). This stakeholder group is addressed by a communication strategy that treats this highly heterogeneous audience as a homogeneous entity. The strategy involves focusing on what is common in terms of customer needs and preferences instead of focusing upon the differences of individual subgroups. Communication-wise, messages promoting ELG are designed to have a general appeal, transmitting communication

[11] https://www.european-language-grid.eu/meta-forum-2021/project-expo/

primarily concentrated on the common needs such as information (ELG as an information hub), match-making (a digital marketplace where supply meets demand) and elimination of language barriers. The main communication channels include email campaigns, social media posts, regular newsletter editions and the ELG website, but also presentations and booths at industry events and conferences. For the target group of LT users, we emphasised the possibilities of modern LT and its various application areas. With this aim at stressing the importance of LT, for instance in terms of cross-language communication, information access and automation in fields such as research and the information industry, ELG intends to include both experienced and potential LT users and informs about the important role LT plays in the digital world.

2.2.1 Public Administrations and NGOs

As an EU-funded project, ELG can also provide technologies to public administrations, e. g., to the European institutions or national or regional administrations. For example, ELG offers the language resources provided by ELRC, which were collected and prepared to serve the needs of public services and administrations across the EU, Norway and Iceland. At the same time, ELG wants to offer solutions to non-governmental organisations that often have to pursue their goals with limited financial resources. They can benefit from ELG as users of LT because they typically do not have the funding or technological know-how to find LT services or tools that would suit their needs. Apart from more general forms of communication like email campaigns or press releases, representatives of public administrations as well as NGOs were invited to conferences like META-FORUM, where traditionally one of the keynotes or opening addresses is given by a representative of the EU.

2.2.2 European Citizens – Members of the European Language Communities

This stakeholder group also includes the members of the European language communities, i. e., all citizens of Europe, speaking and representing the official EU languages, regional or minority languages or any of the other languages spoken in Europe. Communication, networking and surveying activities have primarily taken place in the EU project European Language Equality (ELE). Through the tight collaboration between ELG and ELE we have been able to identify and exploit a number of synergies, such as, among others, the EU Citizen Survey, through which we have been able to learn more about how Europe’s citizens perceive Language Technology and what kind of preconceptions and demands they have.

2.3 Additional Horizon 2020 EU Projects

The projects supported through the Horizon 2020 call ICT-29-2018 (see Section 1) are a special stakeholder group, as their consortia consist of research centres and universities as well as several industry partners. All projects dealt with domain-specific, challenge-oriented LT and provided services, tools and datasets which are also showcased in ELG. As the projects were especially featured, they benefited from a higher level of promotion (Rehm et al. 2021). Furthermore, they could make use of the various features as well as of the vast community connected with ELG. Due to their outreach into industry and academia, they functioned as excellent multipliers on multiple occasions. This target group proved to be very dynamic. We were engaged in active, bi-directional communication with all consortia, e. g., via online meetings, mutual invitations to each other's events, or by advertising our projects on our websites. Communication activities with this group had started in early 2019 and turned out to be successful and vivid.

2.4 Major European Projects and Initiatives

An overarching platform like ELG can only be successful if it is recognised in and used by the whole LT community. To establish ELG within the LT scene and to avoid silo-thinking, we communicated closely and in a targeted manner with other major projects and initiatives in the field including neighbouring areas, in an attempt to establish collaborations to create synergies and to share best practices. The ELG consortium has cooperated directly with projects active in similar areas, with a similar scope or working on similar topics, for example, the European AI on Demand Platform (i. e., the AI4EU EU project), CLAIRE, CLARIN and various other projects and initiatives. In addition to meetings, conferences like META-FORUM are an appropriate format to share information and knowledge about each other's activities. At META-FORUM 2019, 2020, 2021 and 2022, many relevant projects and initiatives showcased their plans and missions with the help of (virtual) expo booths, presentations or panel discussions. Members of the ELG consortium took every opportunity to present ELG at conferences and public events to make the ELG concept and approach known in different sectors and industries. Existing networks like ELRC (European Language Resource Coordination) and META-NET were tapped regularly with regards to knowledge transfer and information exchange. Section 4 presents these collaborations in more detail.

2.5 National Competence Centres

The National Competence Centres (NCC) played a crucial role for ELG's communication and promotion activities (see Chapter 11). This stakeholder group also functioned as an abstract communication channel (Rehm et al. 2021). The NCCs were and still are an important target group included in our communication channels, they also served as multipliers of the ELG mission in their own regions and networks, through mailings, social media posts, newsletter features, face-to-face meetings, conferences, tutorials, training sessions and promotion events.

2.6 Public at Large

ELG is a public and inclusive platform that also attempts to address citizens interested in Language Technology. Members of civil society who browse the web and visit ELG with no specific intention, also need to be addressed adequately. ELG wants to promote the purpose and usability of LT beyond the borders of tech-savvy stakeholders. Our communication activities aim not only at experts, but also at the public at large. Appropriate communication channels are news and blog posts on the website or videos on platforms such as YouTube. Social media channels, especially Twitter, are used to communicate updates and project results in a style that intends to make them interesting and comprehensible to audiences beyond the core LT community. Of interest are especially those ELG features that have broader social implications due to related topics in the news, which are perceived positively by followers and readers with diverse professional and personal backgrounds.

3 Communication and Outreach Activities

As a project with several objectives, addressing various gaps in the European Language Technology landscape and serving as a marketplace for research and industry, ELG depends on the reputation and brand it has established. In addition to the platform's functionality and positive experience of users and providers interacting with ELG, another relevant aspect is the ease of access with regard to content and information served by the platform. This refers to the information architecture of the website, structure and quality of the technical documentation, responses to requests directed at the ELG technical team as well as the overall communication strategy.

3.1 Communication Strategy

A communication strategy enables effective communication, in the case of ELG, this relates to informing specific target audiences and the broad public about the project and its results, gaining users and providers for ELG and representing ELG as a brand for pan-European, multilingual and all-encompassing LT. The key elements of the communication strategy are the stakeholders, the overall goals of the communication, the messages to communicate, the communication channels and the timing.

We have two main communication goals that are closely aligned with our Unique Selling Propositions (USP), which are the key differentiators from existing platforms and offerings on the market. The success of the project and the ELG legal entity depend on these two USPs to be widely known by all relevant stakeholders. This is why the USPs became central messages for communication related to the uptake and popularity of ELG directed at potential users, participating organisations or stakeholders to be won over.

ELG is the primary platform for Language Technologies in Europe.

ELG strives to become the most important and most relevant marketplace for Language Technology in Europe – a one-stop LT shop in which all kinds of stakeholders can find what they are looking for in terms of services, tools or resources provided by research or industry. ELG is not only a directory of companies, universities and research centres, but also contains a repository of thousands of datasets as well as hundreds of functional tools and services. To make ELG useful and efficient for its users, visibility and completeness are crucial. Moreover, to include as many relevant players as possible, one of the main objectives is wide outreach.

ELG provides Language Technology for *Europe built* in *Europe.*

The second USP relates to the fact that LT from other continents or large global technology corporations do not have intimate linguistic knowledge of Europe's languages including their varieties (i. e., European developers of LT can serve European demand in a better and more adequate way) and that legal aspects such as copyright law, the General Data Protection Regulation (GDPR) and other policies are well considered by European players. The same goes for core European values like privacy, confidentiality and trustworthiness. Users of ELG do not have to fear their data being sold to third parties when using or offering services or resources on the platform.

3.2 Communication Campaign

The ELG communication campaign was developed and operationalised with communication experts and continuously revised and expanded to meet the changing conditions in the project and initiative. The initial situation was thoroughly analysed and then appropriate marketing measures were planned using various communication channels including social media.

3.2.1 Communication Objectives

In addition to communicating the overall USPs of ELG to the relevant stakeholders, all ELG communication activities are also geared towards supporting and realising

ELG's overall objectives. We distilled the overall objectives into three main messages, which are the underlying drivers in all ELG communication activities:

- *Support the Multilingual Digital Single Market* by providing technologies for all European languages, which can be used by LT user stakeholders in all European countries to provide digital offerings, products and solutions that support all European languages relevant for the respective LT user stakeholder.
- *Establish and grow a vibrant community and help coordinate all European LT activities* by becoming the primary platform for LT in Europe.
- *Develop and offer a powerful and scalable LT platform* through a novel technological approach, which enables innovations and synergies between commercial and non-commercial LT providers, buyers and users.

3.2.2 Communication Channels

For ELG, we selected four main areas of communication as the most relevant ones for informing the main stakeholders and for marketing the project and the platform. These four areas include the ELG website itself, the annual ELG conference (and other events), the ELG social media channels and the ELG newsletter. While the ELG website and the representation of the project at conferences and events was primarily connected to the ELG brand, a more flexible approach was chosen for social media and the newsletter.

For the duration of the project, we maintained, in addition to the actual European Language Grid, a separate ELG website for information, promotion and marketing purposes. This website served as the face to the public with all relevant information on the project itself and its wider setup, including, among others, the ELG architecture, NCCs, annual conferences, newsletter and many other topics. It also included a news section and a blog. This stand-alone website has been merged with the European Language Grid proper in the summer of 2022 so that all the relevant information and the European Language Grid itself are now available at the same address.[12]

From 2019 to 2022, ELG organised an annual conference (in 2022 in collaboration with the EU project ELE). At these conferences, all relevant aspects of ELG have been presented and discussed with relevant stakeholders. In addition, ELG participated in many other conferences, workshops, industry events and expos. For more details see Chapter 11, Section 3 (p. 210 ff.).

In terms of social media channels, ELG uses Twitter and LinkedIn, their main advantages are the potential to create a very wide reach and large number of followers, thus enabling the project to address exactly the right stakeholders. Instead of establishing dedicated channels for ELG, we decided to create one slightly more general online identity, namely the umbrella-brand "European Language Technology" (ELT), which serves as the name of the social media channels on Twitter[13]

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

[13] https://twitter.com/EuroLangTech

and LinkedIn[14]. This brand serves as the outlet and interaction channel for ELG and also for its sister project, European Language Equality. The ELT brand solves the problem of communicating about two related but different projects through a single channel, while tackling the topic of European Language Technology from a technological (ELG) and from a strategic perspective (ELE). The approach has proven successful, as the ELT channels quickly gathered several hundred followers each. Table 1 shows some key statistics on both platforms.

Channel	Twitter	LinkedIn
Followers (total)	666	818
Posts (total)	316	150
Posts per month (example: March 2022)	27	19
Followers gained per month (example: March 2022)	63	75
Profile visits per month (example: March 2022)	5,944	198
Impressions per month (example: March 2022)	40,300	9,248

Table 1 European Language Technology: social media statistics (July 2022)

The differences in the statistics of the two channels can be attributed to the fact that while Twitter generally sees more activity in interaction and content reception, LinkedIn follows more professional conventions and goals. Its user base has a slightly bigger overlap with the main target groups of ELG. This is why the LinkedIn channel gained more followers even though there was less activity in comparison to Twitter. Both channels are used for communicating a variety of contents in specified formats: 1. new ELG platform features and quotes from reports are shared in specifically designed images (known as shareables); 2. new blog articles are promoted through links and quotes from the text; 3. upcoming events are promoted using, e. g., summaries of the programme and links to the event website; 4. related news from other sources are shared through links or the retweet/sharing function, ideally with a comment regarding the relevance for ELG.

Following the concept of the ELT brand, a newsletter was established under the same name, sharing information from and about ELG and ELE with a total of approx. 4,000 subscribers as of July 2022.[15] We invited many of our existing contacts to subscribe to the newsletter, we invite visitors of the website to subscribe to the newsletter and we also share the newsletter on a regular basis through our other communication channels. At first the newsletter was published on a monthly, later on a bi-weekly basis. Each issue of the newsletter includes a general introduction to the latest edition, including a list of highlights from social media and an overview of press articles in relation to ELT, followed by dedicated sections on ELG and ELE. The ELG section contains general news from and about ELG, a summary of the latest ELG blog article, a few of the latest tools or services added to ELG and the latest organisation that joined ELG (short profile and link to their ELG entry).

[14] https://www.linkedin.com/company/74073406

[15] https://www.european-language-technology.eu/elt-newsletter-archive/

4 Collaborations with other Projects and Initiatives

ELG is a technology platform for the whole European LT community, which is why collaboration played and plays an important role for the success and uptake of the ELG initiative (Rehm et al. 2020c). While we are unable to list all projects and organisations we collaborated with during the ELG project's runtime, below we attempt to list the major ones (see Chapter 2, Section 8, p. 27 ff., as well as Chapter 6, p. 107 ff.).

European Language Equality ELG and ELE[16] worked together on many different topics. ELE collected more than 6,000 LT and LR records, which were ingested in ELG, resulting in a substantial increase of the total number of available resources (Giagkou et al. 2022). The Digital Language Equality metric, developed by ELE (Gaspari et al. 2022; Grützner-Zahn and Rehm 2022), is based on the contents of the ELG catalogue and can be accessed through a dashboard developed by ELE and available on ELG.[17] While ELE prepares the strategic agenda and roadmap towards digital language equality in Europe, ELG offers the appropriate platform for sharing and deploying these Language Technologies. The synergies between the projects were communicated through blog articles and our shared social media channels as well as our shared newsletter.

Open Calls and Pilot Projects ELG collaborated with the organisations behind the 15 selected pilot projects in terms of technical aspects and communication activities on their respective regional and local levels (see Part IV, p. 256 ff.).

ICT-29b) Projects ELG collaborated with the six EU projects funded through the Horizon 2020 call ICT-29b), i. e., Bergamot[18], Comprise[19], ELITR[20], Embeddia[21], GoURMET[22], Prêt-à-LLOD[23] and their consortia and networks, especially with regard to outreach and communication, coordination and making project results available through ELG.

European AI on Demand Platform ELG cooperated with the European AI on Demand Platform through the EU project AI4EU.[24] Topics include strategic and coordination aspects, the technical interoperability between both platforms (Rehm et al. 2020b), the preparation of an AI ontology and participation in outreach and promotion events.

HumanE AI Net This EU network of excellence[25], which also belongs to the European AI on Demand Platform, aims at facilitating a European brand of trustwor-

[16] https://european-language-equality.eu

[17] https://live.european-language-grid.eu/catalogue/dashboard

[18] https://browser.mt

[19] https://www.compriseh2020.eu

[20] https://elitr.eu

[21] http://embeddia.eu

[22] https://gourmet-project.eu

[23] https://pret-a-llod.github.io

[24] https://www.ai4europe.eu

[25] https://www.humane-ai.eu/workpackages/

thy, ethical AI that enhances human capabilities and empowers citizens and society to effectively deal with the challenges of an interconnected globalised world. ELG supports this initiative as language is a core topic in human-oriented AI. Many organisations involved in ELG are also active in HumanE AI Net through specific microprojects that focus on certain research questions, funded by the initiative. HumanE AI Net and ELG collaborated with regard to joint outreach and promotion activities.

CLAIRE ELG and the Confederation of Laboratories for AI Research in Europe[26], the world's largest network for AI research, collaborated with regard to strategic and coordination topics. ELG, representing the language-centric AI landscape, serves as a link between the LT and the AI communities. We also participated in various joint events.

CLARIN ELG and the Common Language Resources and Technology Infrastructure[27] (Eskevich et al. 2020) collaborated with regard to strategic and technical aspects such as metadata harvesting (see Chapter 6) and events.

Microservices at your Service This CEF-supported EU project collects and develops a larger number of functional services, develops ELG-compatible containers and makes these available through ELG.[28] Additionally, the two projects collaborated by participating in relevant outreach and training events.

NTEU and MAPA The two CEF-supported EU projects Neural Translation for the EU (NTEU)[29] and Multilingual Anonymisation for Public Administrations (MAPA)[30] have contributed a large number of tools and services to ELG (García-Martínez et al. 2021). NTEU alone has provided hundreds of high quality machine translation models, which are now available through ELG.

WeVerify This EU project develops tools and technologies for the identification and verification of various types of news and media (Marinova et al. 2020).[31] Internally, the WeVerify tools make use of several ELG services.

ELRC The CEF-supported EU initiative European Language Resource Coordination (ELRC)[32] supports multilingual Europe, among others, by collecting publicly available language data from national public administrations and making them available to the European Union through the repository ELRC-SHARE (Lösch et al. 2018). ELG automatically harvests ELRC-SHARE, enabling the discovery of these resources through ELG. ELRC and ELG also collaborated in terms of joint communication and dissemination activities.

QURATOR The German project QURATOR has developed a technology platform and large number of tools, services and resources for several digital con-

[26] https://claire-ai.org

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

[28] https://www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry

[29] https://nteu.eu

[30] https://mapa-project.eu

[31] https://weverify.eu

[32] https://www.lr-coordination.eu

tent curation use cases (Rehm et al. 2020a).[33] Both projects, QURATOR and ELG worked together closely from the very beginning in terms of platform development, communication and dissemination, among others, through the annual QURATOR conferences. Many tools and resources created by QURATOR are available through ELG.

PANQURA This sister project of QURATOR focuses upon the application of QURATOR technologies to the COVID-19 pandemic, striving for more transparency in times of a global crisis.[34] Among others, PANQURA has developed tools for the automated assessment of the credibility of online content, which are now available through ELG (Schulz et al. 2022).

OpenGPT-X and Gaia-X The German project OpenGPT-X develops large language models for the German language.[35] The project is part of a group of AI projects that will test and deploy their project results through the emerging Gaia-X infrastructure.[36] In Gaia-X, representatives from business, politics, and science are working together to create a federated and secure data infrastructure for Europe, addressing the topic of data sovereignty in Europe. OpenGPT-X will not only make use of various resources available in and through ELG, the project will also extend ELG so that the platform is compatible with Gaia-X, i. e., OpenGPT-X will integrate the ELG platform into the emerging Gaia-X infrastructure.

NFDI4DS The project NFDI for Data Science and AI[37] is part of the German NFDI initiative, which develops, with a total of approx. 20-25 projects, the national German research data infrastructure.[38] NFDI4DS will support all steps of the research data life cycle, including collecting or creating, processing, analysing, publishing, archiving, and reusing resources in Data Science and AI. In NFDI4DS, ELG will be integrated into the emerging NFDI infrastructure.

DataBri-X The EU project Data Process and Technological Bricks for expanding digital value creation in European Data Spaces (DataBri-X), which will start in October 2022, will develop a toolbox for data processing, data handling and data curation. The ELG platform will be used and also extended as one technical infrastructure in this project.

SciLake The EU project Democratising and Making Sense out of Heterogeneous Scholarly Content (SciLake), which will start in January 2023, will build upon the OpenAIRE ecosystem and European Open Science Cloud (EOSC) services to facilitate, among others, the creation, interlinking and maintenance of research-oriented knowledge graphs. In SciLake we will establish technical bridges between the ELG platform and EOSC.

[33] https://qurator.ai

[34] https://qurator.ai/panqura/

[35] https://opengpt-x.de

[36] https://gaia-x.eu

[37] https://www.nfdi4datascience.de

[38] https://www.nfdi.de

5 Conclusions

As a community platform and initiative, ELG does not operate in a vacuum without contact to other projects, groups or initiatives. On the contrary, it is of fundamental importance that ELG is tightly integrated into the community with active use of the ELG platform by many members of the community. To achieve this, ELG has defined its target groups and cooperates closely with a number of relevant projects to exploit existing synergies. These networking and collaboration efforts will be continued after the runtime of the ELG EU project, i. e., when the ELG legal entity is established and operational. This approach is based on a clear communication strategy with transparent goals that are pursued jointly with other key stakeholders.

While we have been able to establish a shared platform for the European LT community during the 42 months of the ELG project, we now need to concentrate on engaging with more and more stakeholders so that ELG is also utilised and expanded by more and more active users, resulting in a European Language Grid *from* the European LT community *for* the European LT community.

References

Eskevich, Maria, Franciska de Jong, Alexander König, Darja Fišer, Dieter Van Uytvanck, Tero Aalto, Lars Borin, Olga Gerassimenko, Jan Hajic, Henk van den Heuvel, Neeme Kahusk, Krista Liin, Martin Matthiesen, Stelios Piperidis, and Kadri Vider (2020). "CLARIN: Distributed Language Resources and Technology in a European Infrastructure". In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France: ELRA, pp. 28–34. URL: https://aclanthology.org/2020.iwltp-1.5.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. URL: http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

García-Martínez, Mercedes, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis (2021). "Neural Translation for European Union (NTEU)". In: *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*. Association for Machine Translation in the Americas, pp. 316–334. URL: https://aclanthology.org/2021.mtsummit-up.23.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022). "Introducing the Digital Language Equality Metric: Technological Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Giagkou, Maria, Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, Athanasia Kolovou, and Leon Voukoutis (2022). "Collaborative Metadata Aggregation and Curation in Support of Digital Language Equality Monitoring". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 27–35. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.3.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). “Introducing the Digital Language Equality Metric: Contextual Factors”. In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Lösch, Andrea, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith (2018). “European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management”. In: *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA. URL: https://aclanthology.org/L18-1213.

Marinova, Zlatina, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva (2020). “Weverify: Wider and Enhanced Verification for You Project Overview and Tools”. In: *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–4. DOI: 10.1109/ICMEW46912.2020.9106056.

Rehm, Georg, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezanezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova, and Franziska Heine (2020a). “QURATOR: Innovative Technologies for Content and Data Curation”. In: *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*. Ed. by Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, and Lydia Pintscher. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020. Berlin, Germany. URL: http://ceur-ws.org/Vol-2535/paper_17.pdf.

Rehm, Georg, Dimitrios Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julián Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John P. McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš (2020b). “Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability”. In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France, pp. 96–107. URL: https://www.aclweb.org/anthology/2020.iwltp-1.15.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020c). “The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe”. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg and Andy Way, eds. (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Forthcoming. Springer.

Schulz, Konstantin, Jens Rauenbusch, Jan Fillies, Lisa Rutenburg, Dimitrios Karvelas, and Georg Rehm (2022). "User Experience Design for Automatic Credibility Assessment of News Content About COVID-19". In: *Proceedings of HCI International 2022 – Late Breaking Papers*. Accepted for publication. 26 June-01 July 2022.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. URL: https://data.europa.eu/doi/10.2861/136527.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. URL: https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Chapter 11
ELG National Competence Centres and Events

Katrin Marheinecke, Annika Grützner-Zahn, and Georg Rehm

Abstract The National Competence Centres (NCCs) in ELG are an international network of 32 regional and national networks, lead by one regional/national representative. The 32 NCCs play a crucial role in ELG, they support the project by bringing in their corresponding regional and national perspective and stakeholders, organising ELG workshops and functioning as regional/national representatives. The chapter explains why, despite a considerable coordination effort, it was worth putting this network together. One important task carried out by the NCCs was to conduct regional/national dissemination events and to participate in relevant regional/national events and also in the annual META-FORUM conferences, organised by ELG.

1 Introduction

The diverse *Multilingual Europe* community, consisting of multiple stakeholder groups, is an important component of our concept for the ELG (Rehm et al. 2020). This heterogeneous set of stakeholder groups includes LT provider companies, LT user/buyer companies, research centres and universities involved in LT research, development and innovation activities, language communities, politics and public administrations, national funding agencies, language service providers and translators as well as the European citizen at large (Rehm et al. 2021).

In this chapter we focus upon one specific part of the wider group of stakeholders involved in the ELG initiative, i. e., the National Competence Centres (NCCs). The ELG NCCs are an international network of 32 regional and national networks. Section 2 describes the NCCs as well as the activities carried out together with the NCCs. We also touch upon the setup procedure and the involvement of the NCCs. Conferences, workshops and other events play a crucial role in disseminating the mission and idea of the ELG initiative, as well as the platform itself. We involved the NCCs to help spread the word about ELG on the regional and national levels. A major part

Katrin Marheinecke · Annika Grützner-Zahn · Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany,
katrin.marheinecke@dfki.de, annika.gruetzner-zahn@dfki.de, georg.rehm@dfki.de

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_11

of their involvement was, thus, devoted to the organisation of and participation in conferences and events.

Section 3 provides a brief overview of the events and conferences ELG organised or participated in. We focus upon the four editions of the annual META-FORUM conference series, which were organised by the ELG project (2019 until 2022). Due to the impact of the COVID-19 pandemic, more than two thirds of all events planned under the umbrella of the project had to be organised as virtual events.

2 National Competence Centres

The ELG National Competence Centres comprises 32 colleagues from all over Europe who all have their own strong regional and national networks, which comprise both industry and also research. For the setup of the NCCs, we benefited from structures and instruments that have been set up by partners of the ELG consortium starting in 2010 and that have been in active use since then, including META-NET[1], META-SHARE (Piperidis et al. 2014)[2], CRACKER (Cracking the Language Barrier, Rehm 2017)[3], EFNIL (European Federation of National Institutions for Language)[4], ELRC (European Language Resource Coordination)[5] and the META-FORUM conference series (Rehm et al. 2016, 2020).

In ELG, we made use of this large set of collaborators, established infrastructures and communication instruments. The involvement in different projects and initiatives made it possible to set up a strong and representative network of National Competence Centres with broad reach into regional and national networks already during the ELG proposal preparation phase, i. e., before the project had actually started. We invited more than 30 experts from the field that met a number of criteria (participation in relevant initiatives, members of academic organisations, good connections to industry and research etc.) to participate in ELG as National Competence Centre Leads with a clearly defined set of tasks and responsibilities.

2.1 Tasks and Responsibilities

The NCCs support the ELG project and initiative in various ways. This international network of national networks not only significantly contributes to the population of the ELG cloud platform with services, resources and data sets, it also plays an important role for broadening the reach of the ELG project and initiative. Early in

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

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

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

[4] http://www.efnil.org

[5] https://lr-coordination.eu

the project, the NCCs were asked to provide information and share their knowledge, e. g., on national/regional information about services, datasets, resources, tools, technologies, research centres, experts, communities, companies, initiatives, projects etc. Additionally, the NCCs have been crucial as multipliers who spread the word about ELG and inform regional and national stakeholders and organisations about ELG and its benefits. The NCCs also fed local needs, ideas and demands back to the ELG to make sure that the ELG development takes the requirements of their constituency into account. Moreover, the NCCs helped with general outreach and dissemination activities, e. g., promoting events like the ELG conferences (Section 3) or the ELG open calls (see Part IV) through their own established channels and networks.

Whereas some activities could be performed by the NCCs with sending emails and providing quickly accessible information, there are a number of tasks that required more effort. These included:

- Organisation of a regional/national ELG workshop including agenda preparation, advertising and promotion (web, social media, emails), identification of speakers and participants etc.
- Participation in regional/national events (both scientific and industry conferences and workshops) on behalf of ELG to promote ELG and to interest relevant stakeholders from research and industry.
- Participation in each of the annual ELG conferences (META-FORUM) in order to strengthen the LT community, support dissemination activities related to ELG and to foster discussion on current LT-related topics and trends.
- Desk research and information gathering: Collection of relevant regional/national information regarding funding programmes, national language (technology) development plans, AI strategies etc. with the overall goal of putting together a comprehensive picture of the European LT landscape.[6]

These tasks corresponded to the priorities of the ELG project consortium, but were to be understood as recommendations rather than mandatory activities. The actual selection of tasks to be organised by an NCC depended on the situation in their country and was determined individually.

We organised meetings with all NCC Leads approximately twice a year; originally at least one annual meeting was meant to be held as a face-to-face meeting co-located with the annual ELG conference in order to minimise travel efforts.[7] Due to the COVID-19 pandemic, further face-to-face meetings have been impossible, which is why all follow-up meetings were held virtually. In the NCC meetings, all NCCs Leads were asked to report briefly on the situation in their countries; furthermore, planned activities and tasks foreseen were discussed. Contractual and organisational matters could also be addressed.

[6] With regard to these desk research activities, many synergies with the project European Language Equality (ELE), which started in January 2021 and which included almost all NCCs as consortium partners, have been identified and made use of, see https://european-language-equality.eu.

[7] The first and, so far, only face-to-face meeting of all National Competence Centres took place on 7 October 2019, as a pre-conference meeting of META-FORUM 2019 in Brussels (see Figure 1).

Fig. 1 National Comptence Centre meeting in Brussels, Belgium (7 October 2019)

2.2 Role and Structure

The rationale behind setting up this international network of national networks was to broaden the reach of the ELG consortium, to provide input with regard to the linguistic situation in the different countries and to fuel the knowledge transfer and sharing between national programmes and initiatives on the one hand and ELG on the other. Since the EU Member States and other European countries have quite diverse situations and individual language policies, a "one-size-fits-all" approach would not have worked. It was crucial for ELG to have access to dedicated experts in all countries to turn to and ask for input. Due to their vast personal connections, the NCCs were ideally suited to make the ELG initiative known in the local markets and in the research spheres of their home countries. It was a deliberate decision to move forward only with academic organisations as NCCs in order to guarantee independence from any commercial interests.

The network of NCCs was compiled based on participation in existing structures and initiatives (META-NET, ELRC NAPs, CLARIN etc.), taking into account scientific standing and existing connections to industry and research. Table 1 lists the NCC Leads, their country and affiliation. Figure 1 shows the NCC Leads at the NCC kick-off meeting in October 2019 in Brussels, Belgium.

2.3 Visibility and Promotion

The NCCs provided valuable insights and feedback to the ELG project and initiative. In return, the project consortium helped increase the visibility of the NCCs and their institutions, for example, by promoting the NCCs and their organisations on the ELG

Name and Country		Institution
Dagmar Gromann	AT	Zentrum für Translationswissenchaft, Universität Wien
Walter Daelemans	BE	Comp. Ling. and Psycholing. Res. Centre (CLiPS), Univ. of Antwerp
Svetla Koeva	BG	Institute for Bulgarian Language, Bulgarian Academy of Sciences
Marko Tadić	HR	Inst. of Ling., Faculty of Hum. and Social Science, Univ. of Zagreb
Dora Loizidou	CY	Department French and Modern Languages, University of Cyprus
Jan Hajič*	CZ	Inst. of Formal and Applied Linguistics, Charles University
Bolette S. Pedersen	DK	Centre for Lang. Tech., Dpt. of Nordic Research, Univ. of Copenhagen
Susanna Oja	EE	Competence Centre for NLP at the Institute of the Estonian Language
Krister Lindén	FI	Department of Digital Humanities, University of Helsinki
François Yvon	FR	Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS
Georg Rehm*	DE	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
Maria Gavriilidou*	EL	Institute for Language and Speech Processing (ILSP), R. C. “Athena”
Tamás Váradi	HU	Research Institute for Linguistics, Hungarian Academy of Sciences
Eiríkur Rögnvaldsson	IS	School of Humanities, University of Iceland
Andy Way	IE	ADAPT Centre, Dublin City University (DCU)
Bernardo Magnini	IT	Human Language Technology, Fondazione Bruno Kessler (FBK)
Inguna Skadina	LV	Institute of Mathematics and Computer Science, University of Latvia
Albina Auksoriūtė	LT	Institute of the Lithuanian Language
Dimitra Anastasiou	LU	Luxembourg Institute of Science and Technology (LIST)
Michael Rosner	MT	Department Intelligent Computer Systems, University of Malta
Vincent Vandeghinste	NL	Instituut voor de Nederlandse Taal (INT)
Kristine Eide	NO	Norwegian Language Council
Maciej Ogrodniczuk	PL	Institute of Computer Science, Polish Academy of Sciences
António Branco	PT	Department of Informatics, University of Lisbon
Dan Tufiş	RO	Research Institute for AI, Romanian Academy of Sciences
Cvetana Krstev	RS	Faculty of Mathematics, Belgrade University (UBG)
Radovan Garabík	SK	L’udovít Štúr Institute of Linguistics, Slovak Academy of Sciences
Simon Krek	SI	Jozef Stefan Institute (JSI)
Marta Villegas	ES	Barcelona Supercomputing Center (BSC)
Jens Edlund	SE	Royal Institute of Technology (KTH)
Hervé Bourlard	CH	Idiap Research Institute
Kalina Bontcheva*	UK	Department of Computer Science, University of Sheffield

** Person belongs to the ELG consortium*

Table 1 List of National Competence Centres

website.[8] At the ELG conferences, the organisers dedicated several sessions to the activities and concerns of selected NCCs and also addressed locally relevant aspects in the conference programme. Furthermore, the NCC meetings served as discussion platforms where the NCCs could promote their topics and exchange experience and knowledge with colleagues from other countries.

The fact that more than two thirds of the project’s runtime took place during the global COVID-19 pandemic thwarted our collective plans for almost all face-to-face events and workshops and severely affected our dissemination activities. However, the shift to virtual formats has allowed interested people to attend conferences or workshops who might not have attended otherwise because of the effort and expenses

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

involved. In addition, online events have made it easier for the ELG team to provide presentations and platform demos because there was no travel component involved.

In June 2022, a new format was introduced for META-FORUM 2022, as this final project conference was planned and conducted as a hybrid event, combining the benefits of face-to-face and online conferences.

2.4 Operational Aspects

Operationally, DFKI as the coordinating partner of the ELG project prepared subcontracts that specified the details of the cooperation between ELG and the NCCs. The NCCs agreed to take over tasks in the interest of disseminating and promoting the European Language Grid in their countries with the activities described. In return, the ELG project reimbursed costs incurred for activities like:

- Organisation of a regional or national ELG workshop.
- Participation in the annual ELG conferences 2019 and 2022 (including costs for travel and accommodation).
- Participation in regional or national conferences or other events to promote ELG (including costs for travel, accommodation and conference fees, if applicable).
- Desk research, participation in surveys or questionnaires, communication and participation in virtual meetings.

3 Conferences and Workshops

ELG organised four annual conferences (META-FORUM 2019, 2020, 2021, 2022) to present, discuss and widely disseminate the idea of a joint technology cloud platform under the umbrella of the wider *Multilingual Europe* topic. While these conferences are described in more detail in Section 3.1, the more focused ELG workshops and additional events are described in Sections 3.2 and 3.3, respectively.

3.1 META-FORUM Conference Series

META-FORUM is the annual international conference on Language Technologies in Europe, organised by ELG together with the META-NET Network of Excellence, dedicated to fostering the multilingual European information society. Previous META-FORUM editions were organised and financially supported through the EU projects META-NET (T4ME; 2010, 2011, 2012, 2013) and CRACKER (2015, 2016, 2017). For the four editions 2019-2022, ELG took over the organisation of META-FORUM, which at the same time serves as the annual ELG conference (Section 3.1.1 to 3.1.4). Table 2 shows all META-FORUM conferences so far.

The two main goals of META-FORUM are community building and outreach to the wider European Language Technology community including research and industry. The ELG editions also had the goal of promoting the ELG initiative and also ELG as the primary platform for Language Technology in Europe. The conferences featured presentations and project expos with a special collaboration focus in order to attract users and providers of LT. As the conference also functions as a dissemination and promotion platform, the ambition was to attract a large and varied number of participants so that all relevant stakeholder groups were adequately covered.

Year	Conference Motto	Location	Date
2010	Challenges for Multilingual Europe	Brussels, BE	Nov. 17/18
2011	Solutions for Multilingual Europe	Budapest, HU	June 27/28
2012	A Strategy for Multilingual Europe	Brussels, BE	June 20/21
2013	Connecting Europe for New Horizons	Berlin, DE	Sept. 19/20
2015	Technologies for the Multilingual Digital Single Market	Riga, LV	April 27
2016	Beyond Multilingual Europe	Lisbon, PT	July 04/05
2017	Towards a Human Language Project	Brussels, BE	Nov. 13/14
2019	Introducing the European Language Grid	Brussels, BE	Oct. 08/09
2020	Piloting the European Language Grid	*online*	Dec. 01-03
2021	Using the European Language Grid	*online*	Nov. 15-17
2022	Joining the European Language Grid	Brussels, BE	June 08/09

Table 2 META-FORUM conference series

3.1.1 META-FORUM 2019

META-FORUM 2019 took place in October 2019 in Brussels.[9] Its motto was "Introducing the European Language Grid". The first session was dedicated to a presentation of the overall ELG project including a very first prototype of the platform, which was demonstrated live on stage to the LT community and stakeholders from the EU institutions for the very first time. After presentations of the three project areas (ELG Platform, ELG Content, ELG Community), the open calls for pilot projects were announced including overall procedures and timeline. Another session focused on the six LT research projects – ELITR, COMPRISE, Bergamot, EMBEDDIA, Gourmet and Prêt-à-LLOD – funded under the Horizon 2020 call ICT-29b-2018 "A multilingual Next Generation Internet". Moreover, panel discussions and presentations on LT and AI, on LT and digital public services, on news from the language communities as well as discussions with stakeholders from industry were organised. An expo featured LT and relevant AI projects. Interest in the ELG platform was very high during and after the conference, as evidenced by a high number of relevant discussions during the sessions and in the breaks. These discussions provided valuable feedback for the further development of the platform. All in all, feedback regarding the event

[9] https://www.european-language-grid.eu/meta-forum-2019/

was overwhelmingly positive. Among others, stakeholders from minority languages expect ELG to make significant breakthroughs, because they hope to find datasets more easily. After the conference, we received several enquiries from companies highly interested in including their services in the ELG platform.

3.1.2 META-FORUM 2020

Due to the global COVID-19 pandemic, META-FORUM 2020 had to be organised as a virtual event, it was held in early December 2020.[10] The motto of the conference was "Piloting the European Language Grid" and it consisted of three half days of presentations and panel discussions. META-FORUM 2020 received a lot of interest with many fruitful conversations. Once again, a strong focus was on presenting the wider landscape of currently funded projects in the area of LT and language-centric AI but also the industry perspective was taken into account.

Holding a conference that is supposed to foster community building and networking as an online event, is a technical challenge. At the same time, the year 2020, with many cancelled events, made it even more necessary to provide room for open exchange among colleagues and (potential) collaborators. This is why we decided to organise a large project expo to enable in-depth discussions on different approaches in the various projects.[11] Like a face-to-face expo, not only the general idea of the respective project was presented but the virtual booths also allowed for technical demos, detailed explanations and profound exchange between visitors and the project representatives. The expo featured 35 projects, all of which had their own dedicated virtual meeting room. We also prepared web pages for each project with an abstract, project poster and other visual materials provided by the projects. Thus, visitors could study the material on the website or jump into the project meeting rooms (i. e., the virtual expo booths) and stay in the meetings as long as they liked. Apart from the first set of ten ELG pilot projects, the following projects participated in META-FORUM 2020 with project booths: AI4MEDIA[12], Bergamot[13], COMPRISE[14], CURLICAT[15], DSDE[16], Elexis[17], ELG[18], ELITR[19], ELRC[20], EMBED-

[10] https://www.european-language-grid.eu/meta-forum-2020

[11] https://www.european-language-grid.eu/meta-forum-2020/meta-forum-2020-project-expo/

[12] https://ai4media.eu

[13] https://browser.mt

[14] https://www.compriseh2020.eu

[15] http://clip.ipipan.waw.pl/CURLICAT

[16] https://www.cjvt.si/rsdo/en/project/

[17] https://elex.is

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

[19] https://elitr.eu

[20] http://www.lr-coordination.eu

DIA[21], EUCPT[22], FedTerm[23], Gourmet[24], Lynx[25], MAPA[26], MARCELL[27], Marian[28], MeMAD[29], MT4All[30], NexusLinguarum[31], NTEU[32], Prêt-à-LLOD[33], PRINCIPLE[34], PROVENANCE[35], QURATOR[36] and WeVerify[37]. In addition, members of the ELG consortium provided demos of the platform and discussed questions and ideas of (potential) users, providers and other interested parties.

Interest in the ELG platform and initiative was considerably stronger than in 2019, i. e., ELG was gaining more and more traction. META-FORUM has proven to be an effective marketing and information channel for ELG. Discussions that took place in the expo provided, again, a lot of valuable feedback and inspiration. This format worked also very well to advertise the work of the ELG pilot projects. Despite the challenging conditions, the conference was successful, while it is obvious that virtual events can only emulate certain parts of a face-to-face event while others – the often mentioned informal chats over coffee – are difficult to recreate in the virtual format. While not every participant attended each session, the online format made it possible for visitors to select only those sessions they are interested in and for which they had sufficient time capacities. The virtual format made it possible for all participants to attend including those with time and budget restrictions. A poll during the opening session showed that more than half of the participants attended META-FORUM for the first time in 2020. All META-FORUM 2020 sessions are available online.[38]

3.1.3 META-FORUM 2021

META-FORUM 2021 was the 10th edition of the conference series overall and the second to take place online, given the ongoing pandemic situation.[39] The motto of

[21] http://embeddia.eu

[22] https://www.presidencymt.eu

[23] https://www.eurotermbank.com

[24] https://gourmet-project.eu

[25] https://lynx-project.eu

[26] https://mapa-project.eu

[27] http://marcell-project.eu

[28] https://marian-project.eu

[29] https://memad.eu

[30] http://ixa2.si.ehu.eus/mt4all/

[31] https://nexuslinguarum.eu

[32] https://nteu.eu

[33] https://pret-a-llod.github.io

[34] https://principleproject.eu

[35] https://www.provenanceh2020.eu

[36] https://qurator.ai

[37] https://weverify.eu

[38] https://www.youtube.com/playlist?list=PLL1cFzaG0S5ghZz0HxO5TEUIdwrY7J8qJ

[39] https://www.european-language-grid.eu/meta-forum-2021/

the conference was “Using the European Language Grid” and it highlighted the first actual uses of the ELG platform. The setup of the conference was similar to the structure used in 2020. However, the project expo was organised using the virtual meeting space environment Gather.town to further stress the community aspect.[40]

As the motto implies, in 2021 using and benefiting from ELG was the main focus. To demonstrate this, five of the ten successfully finished pilot projects were featured with their results. Furthermore, representatives from the European LT industry took part in a panel to discuss their expectations towards and experiences with the ELG platform. In the more hands-on ELG integration tutorial, potential users who were considering to integrate their own tools and services into ELG had the chance to learn how. All META-FORUM 2021 sessions are available online.[41]

Overall interest in the conference was enormous and the number of participants significantly exceeded that of the previous year. The feedback collected in the session again proved to be a valuable source of information and was thoroughly evaluated after the conference in order to further enhance the platform development.

3.1.4 META-FORUM 2022

While the virtual editions of META-FORUM 2020 and 2021 were very successful, there are certain disadvantages of online-only events compared to face-to-face conferences. This is why META-FORUM 2022 was organised as a hybrid event, combining the advantages of flexibility and higher reach with the benefits of face-to-face discussions. The onsite conference in Brussels was held under appropriate COVID-19-safe conditions with approx. 100 participants from the European LT community and representatives of the European Institutions. Several hundred participants attended the conference online.

3.2 ELG Workshops

ELG is committed to community building and collaborating with relevant initiatives on the European level as well as forming its own network of networks (Section 2). The network of 32 NCCs acts as local and national bridges to the ELG initiative and cloud platform. Accordingly, dedicated workshops with and for the national LT communities have been a crucial task the NCCs were asked to fulfil.[42] These workshops were organised with the goal of making ELG known all over Europe.

Usually the workshops were organised as individual events by each NCC. In some cases, they were co-hosted by several NCCs together, e. g., the ELG workshop at SwissText 2020 (hosted by the ELG NCCs Austria, Switzerland and Germany) or

[40] https://www.european-language-grid.eu/meta-forum-2021/project-expo/

[41] https://www.youtube.com/playlist?list=PLL1cFzaG0S5iDaCg2SliyA-4axKY0LfiQ

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

National Competence Centre(s)	Location	Date
Switzerland, Austria, Germany	*online*	23 June 2020
Lithuania, Latvia, Estonia	Kaunas, LT	21 Sept. 2020
Poland	*online*	27 Oct. 2020
Finland	*online*	15 Dec. 2020
Germany	*online*	20 April 2021
Austria	*online*	11 May 2021
Switzerland, Austria, Germany	*online*	14 June 2021
Belgium, Luxembourg	*online*	08 July 2021
Spain	*online*	23 Sept. 2021
Czech Republic, Slovakia	*online*	18 Oct. 2021
Denmark	Copenhagen, DK	16 Nov. 2021
Netherlands	*online*	03 Dec. 2021
France	*online*	08 Feb. 2022
Bulgaria	*online*	11 Feb. 2022
Serbia	*online*	11 March 2022
Norway	Oslo, NO	16 March 2022
Romania	*online*	24 March 2022
Slovenia	*online*	27 May 2022
United Kingdom	*online*	17 June 2022

Table 3 Workshops organised by the National Competence Centres

the ELG workshop of the Baltic NCCs of Lithuania, Latvia and Estonia that was co-located with the Baltic HLT conference in 2020.

Since all workshops were held during the pandemic, almost all were online events that usually attracted between 25 and 100 participants. Depending on the country and target audience of the workshop, they either had a more informative or a more technical spin, or a combination of both. In an introductory talk by the project coordinator or a partner of the consortium, ELG and its history, its goals and current status was presented. In a separate presentation, the technical setup of the platform and its offerings were explained. After that, the NCCs either organised discussion panels or invited speakers from industry to emphasise the demands and expectations towards ELG. Especially these talks often spurred interesting and inspiring discussions and provided valuable feedback for the ELG consortium. In various workshops, a short hands-on tutorial session was included in which a member of the technical ELG team explained how to make available services or resources through ELG. Many of the ELG NCC workshops are available online.[43] Table 3 lists all NCC workshops.

3.3 Additional Conferences

Representatives of the ELG consortium took the opportunity to promote the platform and the initiative at numerous occasions throughout the run-time of the ELG project. In addition to local events, ELG was also present with talks and papers at

[43] https://www.youtube.com/channel/UCarEHmsWT2JslcvvWkbhL4A

more than 50 different European and international conferences, such as LT4ALL (2019), LREC 2020, AI Boost (2021), European Big Data Value Conference (2021), Fachtagung Maschinelle Verfahren in der Erschließung (Deutsche Nationalbibliothek, 2021) and Wales Academic Symposium on Language Technologies (2022).

4 Conclusions

The collaboration with the National Competence Centres was successful. The impact they have had in their countries to promote ELG cannot be overstated. Also, the NCCs' expert knowledge of language resources in their regions and their contacts to representatives from industry and research have been and continue to be extremely useful. Although the formal contracts with the NCCs will expire at the end of the project, we will make an effort to maintain good working relationships with these experts in the future and, if possible, to intensify the work again in future projects.

Under the umbrella of the ELG legal entity we will continue to organise events and workshops in the coming years to demonstrate new developments and to seek contact with the communities in the various European countries and regions in order to further promote networking. The annual META-FORUM conference is an established brand and will continue to be an important activity to bring stakeholders together and counteract the fragmentation of the European LT community. Experiences from the last years with different meeting formats have significantly extended the spectrum of what is possible.

References

Piperidis, Stelios, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi (2014). "META-SHARE: One year after". In: *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: ELRA, pp. 1532–1538. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/786_Paper.pdf.

Rehm, Georg, ed. (2017). *Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda*. CRACKER and Cracking the Language Barrier federation. URL: http://cracker-project.eu/sria/.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: *Language Resources and Evaluation* 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4. URL: http://link.springer.com/article/10.1007/s10579-015-9333-4.

Chapter 12
Innovation and Marketplace: A Vision for the European Language Grid

Katja Prinz and Gerhard Backfried

Abstract This chapter provides a comprehensive overview of innovation and the ELG marketplace as core elements for the generation of value and the creation of an active, attractive and vibrant community surrounding the European Language Grid. Innovation is an essential element in making ELG a credible and sustainable undertaking. However, it does not happen by itself nor materialise in a vacuum. Consequently, ELG provides a habitat for various kinds of innovation and a home for the necessary community to put innovation into action. The marketplace is essential for attracting participants supplying and demanding services, resources, components and technologies on a European scale. Innovation and marketplace – as well as the overall business model – are tightly connected and need to be developed and managed in a joint manner. Clearly, this is not a one-off activity, but rather needs to be carried out continuously and extend into the future. ELG is designed and created to promote the excellence and growth of the European LT market, creating new jobs and business opportunities and supporting European digital sovereignty. Encompassing a wide array of technologies and resources for many languages spoken across Europe and in neighbouring regions, it contributes to the Multilingual Digital Single Market as a cross-European driver for innovation.

1 Introduction

The ELG marketplace and the kinds of innovations it enables form central elements of ELG and its goal to become the *one-stop-shop for Language Technology in Europe*. These aspects are closely interlinked with a series of further topics concerning the business aspects of ELG in a wider sense, none of which can be viewed in isolation but rather need to be approached in a connected and holistic manner.

Artificial Intelligence (AI), Natural Language Processing (NLP) and Natural Language Understanding (NLU) are highly active areas of research and development

Katja Prinz · Gerhard Backfried
HENSOLDT Analytics GmbH, Austria,
katja.prinz@hensoldt.net, gerhard.backfried@hensoldt.net

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_12

leading to novel applications on a continuous basis. Over time, new actors enter the stage and change the course of events. In this highly dynamic landscape it is imperative to constantly monitor progress, remain alert and be able to adapt to newly emerging trends. Consequently, any platform and strategy implemented on and by AI/NLP/NLU need to remain flexible and open to change. Fundamental concepts such as value-generation provide orientation across time and should form the base of any strategies developed.

Neither the marketplace nor innovation make sense without an underlying crowd of committed actors, which drive the cycles of supply and demand, form the ingredients of cooperation and consulting and are at the heart of creation and innovation. Establishing and fostering this community who will take LT one step further thus forms one of the most important tasks to be addressed by ELG.

2 Innovation

In today's agile, interconnected and virtualised world, the paradigm of open innovation (Chesbrough 2006), connecting many different disciplines, sectors and actors in a non-linear fashion has gained considerable traction. Under this paradigm, innovation takes place within as well as outside an organisation with knowledge flowing in both directions. It allows different actors to collaborate and experiment across organisational boundaries, across different sectors and disciplines, and enables them to dynamically produce innovation in a heterogeneous manner. Eco-systems like ELG form a natural habitat for such activities and a powerful environment for innovation. In this chapter, the concept of innovation is viewed from the angle of open innovation, forming the most appropriate and promising approach for a platform like ELG, rather than the silo'd and closed kind of innovation which is limited to individual organisations. For innovation to occur, two fundamental ingredients need to be combined: innovation = invention + adoption (Schrage 2004). Both of these factors must be present for innovation to take place and to put it into effect in order to generate new knowledge, to develop new products, services or processes. Any environment or innovation-strategy consequently has to reflect both factors, balance efforts and encourage and support both kinds of activities.

2.1 Significance of Innovation

Applications in the fields of AI, NLP or NLU reside in a highly competitive and dynamic landscape. As technology leaps are produced in rapid succession and markets and opportunities expand, organisations can and should make use of internal as well as external ideas and paths to market as they seek to advance their technology (Chesbrough 2006). Justin Rattner, Intel's former CTO evangelised the concept of 21st century industrial research where innovation is driven by teams of boundary spanners that possess multidisciplinary skills. Online platforms such as ELG provide ideal multi-sided ecosystems for such teams, offering the means to link up and collaborate and to unite a multitude of participants with the joint aim to create novel products and services ready for swift adoption. However, beyond providing the technical framework, resources and tools, such platforms also foster the sharing and exchange of knowledge and ideas between participants. As a result of the increased diversity and connectedness of actors, the generation of genuinely new knowledge and more radical innovation is possible. Whether and to what extent these goals also materialise in practice depends on a variety of factors, such as acceptance and openness to a culture of open innovation that also supports the useful and selective sharing of research results and data. If exercised successfully, open innovation has the potential to eliminate barriers in research and development and generates a dynamic environment that cannot be achieved with traditional methods.

2.2 Types of Innovation and Innovation Strategies

Innovation may span a wide spectrum concerning products, services, methods, business models and even entire organisations. Figure 1 depicts different dimensions and types of innovation and provides several examples for each kind.

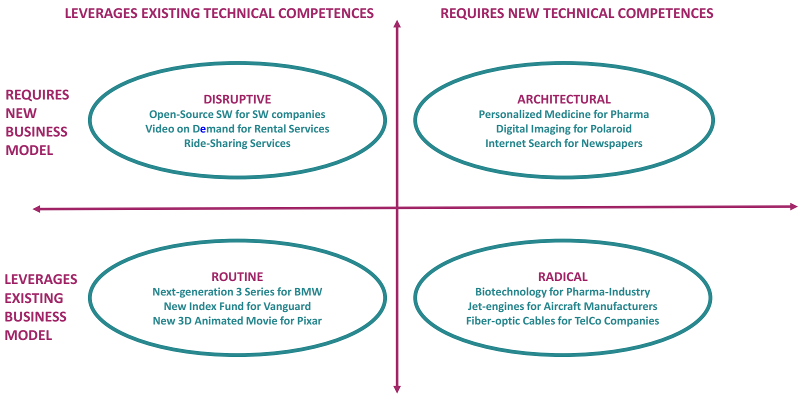

LEVERAGES EXISTING TECHNICAL COMPETENCES
REQUIRES NEW TECHNICAL COMPETENCES
REQUIRES NEW BUSINESS MODEL
DISRUPTIVE
Open-Source SW for SW companies
Video on Demand for Rental Services
Ride-Sharing Services
ARCHITECTURAL
Personalized Medicine for Pharma
Digital Imaging for Polaroid
Internet Search for Newspapers
LEVERAGES EXISTING BUSINESS MODEL
ROUTINE
Next-generation 3 Series for BMW
New Index Fund for Vanguard
New 3D Animated Movie for Pixar
RADICAL
Biotechnology for Pharma-Industry
Jet-engines for Aircraft Manufacturers
Fiber-optic Cables for TelCo Companies

Fig. 1 Innovation landscape (Pisano 2015)

Routine innovation (or incremental innovation) builds on an organisation's existing technological competences and fits with its existing business model and customer base. Routine innovation aims at improving existing products (or services) continuously until the end of their life-cycles. It typically involves activities to improve features, reduce costs or expand production lines and mechanisms. Architectural innovation combines technological and business model disruptions. Disruptive innovation typically requires a new business model but not necessarily a technological breakthrough. For that reason, it also challenges, or *disrupts*, the business models of other actors. Breakthrough innovation can be regarded as the more radical version of disruptive innovation causing fundamental changes in the market through the introduction of new products, methods or business models. These categories are not clear-cut and overlap to some extent. However, the dimensions can serve to locate different types of innovation when designing an innovation strategy. Aside from these categories, innovation can also be characterised by the kinds and magnitude of impact caused by it.

Any innovation strategy needs to specify how the different types of innovation (as outlined above) fit into the overall business strategy. It must map an organisation's value proposition for the defined markets and at the same time set realistic boundaries. Furthermore, the strategy must be clearly communicated in order to assure a common goal for all participants involved, secure their commitment and to streamline activities between all partners. Innovation for innovation's sake or for generic goals such as *"we need to be innovative"* are neither sufficient nor effective. Pisano (2015) emphasises the importance of these inter-connections by defining the term "innovation strategy" as the "commitment to a set of coherent, mutually reinforcing policies or behaviours aimed at achieving a specific competitive goal, promoting alignment among diverse groups within an organisation, clarifying objectives and priorities, helping focus efforts around them and specifying how various functions will support it". Innovation – and an innovation strategy – can neither be developed nor executed in isolation, but need to be carried out in sync with the defined business strategies of an organisation to be successful.

2.3 Open Innovation in the ELG Platform and Marketplace

Innovation does not take place in a vacuum, but is tightly connected to the vision, business, marketplace and sustainability strategies aiming to establish and sustain ELG as the primary marketplace for LT in Europe. The platform and community are positioned at the centre around which these different strategies are aligned, supporting each other in the overall goal as depicted in Figure 2.

ELG is a multi-sided and integrated platform and envisoned to function as an innovation driver during the lifetime of the project as well as beyond. The platform itself is complemented by a vibrant and active community of users and stakeholders. These are a key ingredient in creating the critical mass required to make ELG an established marketplace. Building and strengthening this community consequently forms an essential element of the ELG innovation and communication strategies.

Placing the platform and community at the core allows us to adopt an open and collaborative approach to innovation, which needs to become an inherent element (a process) of ELG. The principles of Open Innovation as coined by Chesbrough (2006) form the over-arching theme of this continuous process. Figure 3 provides a schematic overview of the actors and interactions which need to be aligned for innovation and value creation. It is imperative that all groups are present and participate

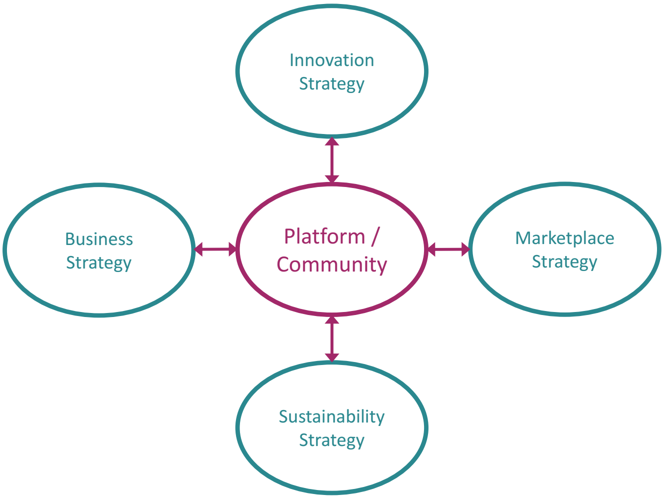


Fig. 2 Strategies centred around the ELG platform and community

actively in the process. To attract and motivate these groups, targeted communication is required.

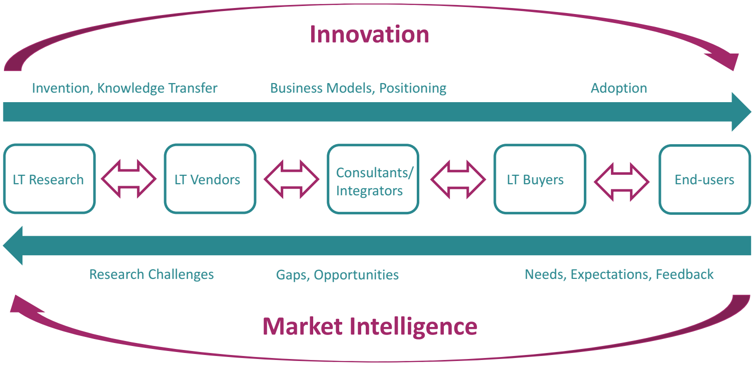


Fig. 3 ELG innovation cycle

In line with the overall approach of ELG, in Figure 3 the process of innovation spans the complete set of activities and actors from invention to adoption. The goal to generate value within the scope of the business model forms the central element. Continuous feedback regarding the needs, gaps, expectations and opportunities is collected via the community, leading to further cycles, which need to be carried out repeatedly and continuously. As a result of the continuous feedback mechanism, strategies can be updated and the speed of adoption increased over time, hence allowing for more rapid cycles of innovation. Figure 4 shows four main dimensions and associated issues to be addressed and considered regarding innovation in ELG.

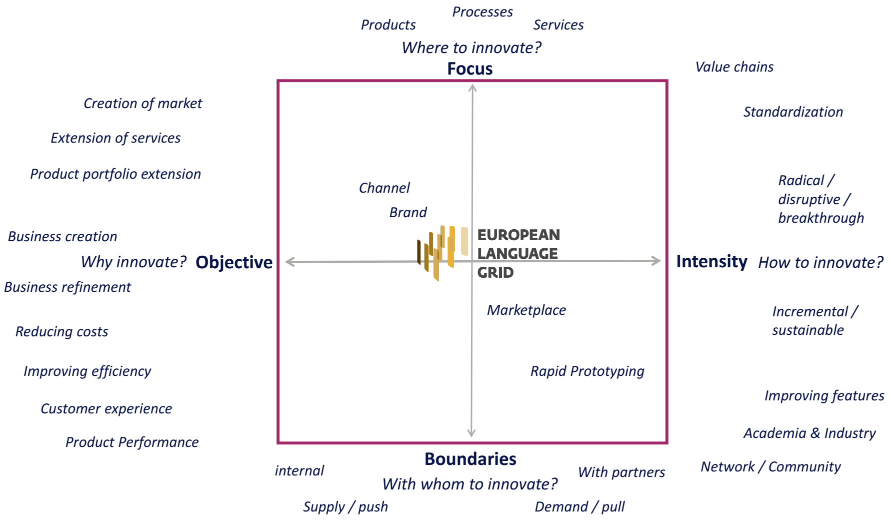


Fig. 4 Dimensions of innovation

For each dimension, several possible approaches are outlined. Together, they form a portfolio of possibilities and opportunities which need to be monitored continuously. Depending on the evolution of ELG, they may need to be adapted to changing conditions and (re-)prioritised. The innovation cycle shown in Figure 3 forms the blueprint for these continuous activities.

For the duration of the ELG project the most important element of innovation is the creation of ELG itself. The use of a platform in the scope of LT as a multi-sided marketplace, allowing participants to create value together by interacting with each other represents an innovative business model (Still et al. 2017). Beyond the platform itself, the creation of products (Section 2.3.1) and services (Section 2.3.2) form two further promising alleys for innovation activities.

2.3.1 Products

ELG provides a large set of technological components and resources which provide a broad basis for product offerings as individual products or product bundles. In terms of innovation potential, both bundles as well as individually improved and adapted LTs provide a wide range of opportunities. Different setups of where services are hosted and run are provided by ELG to optimise resource usage and adapt to the particular needs of customers. An extensive catalogue of tools and resources provides a single point of entry and access to these tools and LTs.

2.3.2 Services

Two kinds of services are provided by ELG: services in the sense of running components (technological services) and services in the sense of experts providing their expertise (human services). In terms of the former, the services can be used individually or in combination (as chains of services) to create value-chains. Combination and composition allow us to establish more complex workflows, enabling end-users to benefit from the capabilities of individual providers without having to worry about any inner workings or being locked in the products of an individual supplier. Corresponding mechanisms regarding billing, licensing and support will provide a single point of contact for customers. Regarding the latter, ELG provides a virtual agora, a business-space for connecting stakeholders developing or deploying (complex) solutions which require skills beyond that of individual actors. This includes services of consultants and integrators who are crucial elements in broadening the adoption (and hence boosting innovation) of LT. They are expected to act as enablers and multipliers for putting LT into practice, supporting their introduction into organisational as well as business processes.

2.3.3 Further Aspects of Innovation

Regarding the *intensity of innovation*, ELG is expected to mainly operate on a level of incremental, continuous innovation, improving existing features and extending the portfolio of features. Through this continuous extension, new combinations of services and products are expected to become available over time which allow the implementation of new features. Linking different services and thus producing value chains in a simple and transparent manner will allow for increased experimentation and thus for an agile environment for the creation of new features. Regarding the *boundaries of innovation*, ELG will focus on the community and stakeholders present on the platform. A catalogue of resources (services, corpora, datasets etc.) as well as of LT experts, consultants and integrators provides a prime resource for locating crucial resources for business. The strength, weight and activity of the community is one of the determining factors for the overall success and adoption of the ELG and hence one of the gate factors for innovation. Regarding the *objective of innovation*, the refinement as well as creation of business form viable alleys. The above-mentioned manner of gradual and incremental innovation lends itself to various kinds of business refinement such as reducing costs, improving the efficiency or product performance and improving customer experience.

Business creation may take place via the platform and community and through the creation of novel services or products via the combination of building blocks offered by ELG. The creation of standards for resources, processing services and interfaces can play an important role as it effectively decouples individual components and vendors. In combination with the technical environment of ELG, this enables increased resilience, scalability, composability and replaceability of components, avoiding vendor lock-in situations. Furthermore, standardisation of these elements will allow for a higher level of experimentation and show-casing and lower the risk of failure in the development of innovative solutions.

3 Multi-sided Marketplace Approach

To date, there is no general digital umbrella platform for LT in Europe. The ELG platform is designed to fill this gap: it is envisioned to serve as the comprehensive virtual marketplace, where value is created for all its members in Europe and beyond. Based on a multi-sided marketplace approach (see Figure 5), ELG will facilitate value and business creation and efficient transactions coupled with large developer ecosystems that build innovative technologies and services on top of a digital platform in an open and agile manner. The advantage of this approach lies in the nature of multi-sided marketplaces as enablers of transactions driving positive network externalities. They make it easier and more efficient for the participants from diverse markets to interact with each other, as the friction between different contact points is reduced. In addition, these interactions increase the value created together which almost comes naturally due to the network effects. A platform becomes more attractive to potential new users the more users meet and interact on it. In other words, value increases for all participants when more users actively use the platform (Sánchez-Cartas and León 2021). As a marketplace, ELG is designed to make it easy and efficient for participants to connect and exchange ideas and products. These can be as diverse as language resources, technologies, services, components, expertise, innovation or even information. The distinctive feature of the multi-sided approach is that the marketplace enables direct interactions between two or more sides, who can be both – product suppliers and demanders at the same time. In other words, value creation is two-way and continuous.

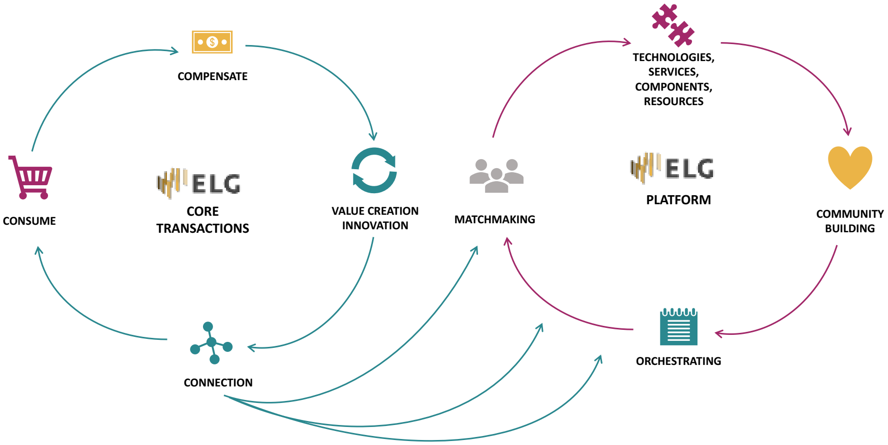


Fig. 5 ELG multi-sided marketplace approach

The core transactions of the multi-sided marketplace are represented by the left part in Figure 5 and are mainly concerned with creating value, establishing connections between supply and demand, and facilitating consumption and compensation of the products (technologies, services, components and resources) offered by ELG. *Connection* is a complex mechanism consisting of the elements portrayed in the right half of Figure 5. Various kinds of connections are supported and promoted by the platform, from matchmaking, to matching of technologies, resources and services vertically and horizontally in order to provide a more comprehensive offering, to orchestrating all interactions between, users, providers and innovators, as well as nurturing a vibrant and active community. The multi-sided marketplace approach encompasses the following principles.

Value Creation ELG aims to be a platform for value creation which will be achieved by facilitating reciprocal exchanges between multiple marketplace participants. In addition, participants can create value by tapping into resources and capacities that they do not have to own. Any resource exchange handled via ELG will reduce transaction costs for each participant and enables access to externalised innovation. The cornerstone of the ELG marketplace positioning is the value it provides to its participants. As *the* European marketplace for LT, it connects previously unmatched supply-side and demand-side participants through innovative forms of value creation, capture and delivery. The value proposition depends on the components and services, their uniqueness, and the means of delivering value to target groups as well as on the right balance between the perceived value and the set price. Furthermore, ELG is the orchestrator to ensure value creation and high quality of participation on the platform. As such, the unique positioning as a marketplace will be based on the value generated and offered across verticals (see Figure 6). For example, a particular buyer receives a vertically packaged LT solution for their desired domain (e. g., the health industry) in the form of a unique combination of components and services from ELG. In addition, they can select the languages for the desired technologies, services and resources for the particular domain.

Connection, Gravity and Flow Whereas traditional offline marketplaces tend to push products and technologies to the market, ELG will rather create a pull-effect. As a multi-sided marketplace it will be equipped to create network effects, i. e., effects that attract new users to enter the marketplace to be part of an ever-growing number of partners who are also part of the network. Together they engage in a mutual value exchange process which is orchestrated by the marketplace. ELG will enable easy access, meaning that participants can easily plug into the platform to share, transact and *connect*. ELG will function like a magnet in creating a pull that attracts participants to the platform with its *gravity*. Because it is both, a transaction and innovation platform, both LT providers and LT users (supply and demand) will be present to achieve critical mass. The *flow* of value will be fostered by matchmaking, i. e., making connections between LT providers and LT users. Rich data will be used for successful matchmaking and the co-creation of value.

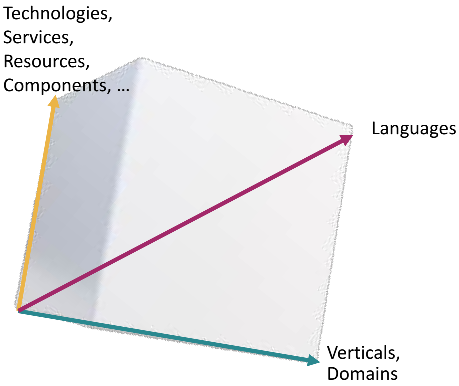


Fig. 6 Value dimensions of the marketplace

Compound Growth The marketplace aims at providing its participants a broad base that enables compound growth and scaling. Growth will be mainly driven by the network effects described above.

Visibility ELG is designed to enhance the visibility of each of its participants, extending their reach and networking power. From the LT vendor perspective the main interest is to acquire customers. As an umbrella platform for European LT, the ELG aims at removing geographic boundaries and language barriers, thus fostering the European Digital Single Market.

Community Building A very important aspect of this approach is to attract, grow and nurture a vibrant and active community around ELG thus promoting an interactive marketplace. The stakeholders include LT providers, academic research organisations, LT customers, EU institutions, public administrations, NGOs, policy makers, project consortia, research projects, as well as the ELG National Competence Centres (NCCs) in 32 European countries. This critical mass of active participants also generates the necessary market pull: an excellent case in point for this are the several pilot projects funded by ELG (see the chapters in Part IV), e. g., Lingsoft, Inc., Coreon GmbH and Elhuyar, among many others, have successfully enhanced the attractivity of the marketplace by contributing highly demanded services, technologies and languages to the platform.

3.1 Foundations for a Successful Marketplace

What are the key ingredients for a successful marketplace? The answer is not straightforward because the formation and growth of marketplaces depends on many factors such as the availability of capital, sufficient demand, talent, legal situation, tax systems, the innovation and startup culture of a country and many more. Nonetheless, there are certain elements successful marketplaces have in common which are equally important for ELG.

Attraction Indisputably, success can only be achieved if enough participants are attracted to join the ecosystem. This *gravity*, which is one of the most important ingredients, will be supported by a well-balanced interplay of supply and demand all of which will be governed by ELG. It is vital for the marketplace to generate a market pull in order to fulfil the goals of self-sustainability. The more participants the marketplace attracts, the greater will be the network effect and compound value growth (a critical mass has to be reached, cf. Bonchek and Choudary 2013). The technical foundation to ensure that people are attracted to ELG is an innovative and state-of-the-art solution for containerised LT components, services and resources coupled with cloud solutions to enable fast and efficient interaction and speedy and scalable innovation.

Demand Economies of Scale ELG will also rely on demand economies of scale, which take advantage of technological improvements on the demand side and are driven by demand aggregation, efficiencies in networks, and other phenomena (like crowd sourcing of software development) that make bigger networks more valuable to their users (Osterwalder and Pigneur 2010). Once the gravity of the marketplace is functioning, network effects will be the natural result. Growth via network effects leads to market expansion. New buyers enter the marketplace, attracted to ELG by the growing number of partners who are part of the network.

Time-to-Market Strategically speaking, ELG will also focus on reduced time-to-market objectives: the corporate strategy of the future marketplace will be designed to truly fulfill the role as accelerator for business creation and will consider concepts like "lean management" and "just-in-time" supply chain delivery. Furthermore, the agile environment will provide a flexible test-bed for trying out new technologies and approaches.

Quality Standards In order to be successful, the marketplace needs to facilitate the exchange of value which means that the components, services, resources provided through ELG require certain quality standards. In order to safeguard the quality of products (technologies, services, resources and components) provided, ELG standards and quality seals will eventually be implemented. In any case, the provision of high-quality state of the art LT, open architecture, reusable software, industry-grade robust components provide key ingredients for establishing confidence and trust in ELG as a whole. In addition, trust in the marketplace will be created through transparent product offering and by providing feedback and reviews of participants concerning their prior transactions.

Orchestration Furthermore, a proper organisation and infrastructure have to be provided to guarantee that the platform smoothly works as enabler of transactions: ideally, the whole setup fosters the exchange and creation of value and supports doing business in an easy and smooth manner. A prerequisite for this is an attractive, simple and transparent licensing and pricing model, and a simple business processing scheme (Täuscher and Laudien 2018).

Ecosystem of Participants Successful ecosystems have the ability to provide for *coopetition* (competition and cooperation) and value co-creation, which are ideally governed by structure and orchestration to work best. ELG will provide for the ideal environment to foster the structured creation and well-coordinated growth of the ecosystem. This principle is also reflected in the paradigm of open innovation adopted and encouraged by ELG.

3.2 ELG Ecosystem of Participants

One of the most important ingredients for a sustainable and successful marketplace rely on the ability of ELG to create, nurture and grow an ecosystem of participants. ELG is in the process of expanding and sustaining a unique ecosystem by attracting diverse stakeholder groups holding different roles – reaching from LT suppliers and demanders to networks and associations, industry members and academia, as well as policy makers and national competence centers (see Figure 7). By aligning itself with key associations and initiatives, ELG aims at establishing itself as a central element in a platform-of-platforms landscape.

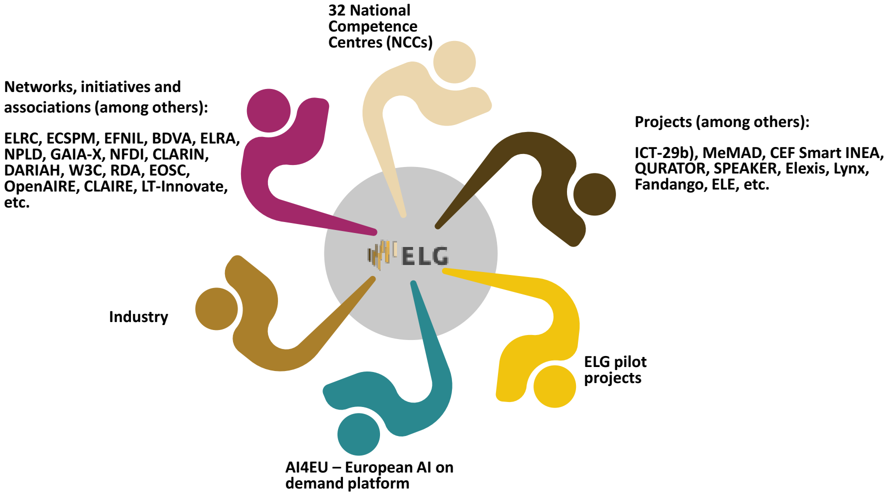

32 National Competence Centres (NCCs)
Networks, initiatives and associations (among others):
ELRC, ECSPM, EFNIL, BDVA, ELRA, NPLD, GAIA-X, NFDI, CLARIN, DARIAH, W3C, RDA, EOSC, OpenAIRE, CLAIRE, LT-Innovate, etc.
Projects (among others):
ICT-29b), MeMAD, CEF Smart INEA, QURATOR, SPEAKER, Elexis, Lynx, Fandango, ELE, etc.
ELG
Industry
ELG pilot projects
AI4EU – European AI on demand platform

Fig. 7 ELG ecosystem of participants

The ecosystem is designed to connect people, foster an environment for open and two-way communication, create mutually beneficial relationships, and promote community building. In short, it is there to provide an umbrella platform for its participants enabling them to build relationships and to provide value to one another. The role of community building is very important because it is the driver of the marketplace. It is needed in order to reach a critical mass of active participants which eventually generate the intended market pull. From a business perspective, ELG will provide the infrastructure for an ecosystem allowing to match products, services, providers (supply) and users (demand), within a multi-sided setup. By orchestrating different stakeholders' needs, the ecosystem will allow for matchmaking of demand

and supply and the continuous multi-directional exchange of values. The technological and organisational infrastructure for this matchmaking as well as the interaction governance principles are key building blocks of ELG.

3.3 Technical and Practical Aspects

From a technical perspective, ELG will be the first large-scale LT platform applying containerisation through Kubernetes. This choice and combination of technnologies provides a scalable environment with an web user interface and corresponding backend components and REST APIs. During the course of the project and beyond, it will provide access to a multitude of state-of-the-art technologies, services and components. Furthermore, it will include an overarching LT directory of stakeholders from research, innovation and technology, i. e., it will be the "yellow pages" or the "who's who" of the European Language Technology community.

On the provider side, ELG adheres to a number of standards in order to facilitate the integration of a large number of disparate tools:

1. Definition of common APIs for each class of tool, designed to be powerful enough to support the necessary use cases but lightweight and flexible enough to allow tools to expose their own specific parameters where this makes sense.
2. Containerisation to isolate tools from one another and to allow each tool to manage its own software dependencies. ELG uses the well-established Kubernetes system to manage the deployment, scaling and execution of containers in combination with Knative to handle auto-scaling of containers on demand.
3. Orchestration of services will become an important topic as the set of offered services grows and the demand for complex workflows becomes visible. This may potentially even concern workflows spanning multiple platforms.

With regard to the user interface, standards in user friendliness are adopted and marketplace-related features, such as upload/download, licensing, billing, payment as well as transparent pricing models will be used. In addition, ELG will promote direct contact to its participants which is important to create additional transparency and trust in the platform.

4 Conclusions

ELG has set its goal to become the *primary platform for Language Technologies in Europe* which incorporates many aspects in one setting: marketplace, business space and a scalable environment for innovation. With regard to innovation, an open innovation approach is adopted, putting the combination of creation and adoption at the centre. Different kinds and granularities of innovation (step-wise and gradual to disruptive) are enabled by ELG and the way the community behind it is set up and managed. Innovation, however, is not viewed in isolation but rather as a crucial element within the larger context of the ELG business model. The marketplace will focus on commercial aspects and communities, linking supply and demand and enabling reciprocal value exchange. In addition, ELG will form a business space and innovation platform in the sense of becoming a virtual agora, bringing researchers, experts, end-users, requirements and capabilities together in one forum. Moreover, it will serve as a promoter for open innovation, providing access to (external and internal) resources and ingredients for innovation. As *the* umbrella platform shared by the whole European LT community, it will support the bundling of efforts and forces and facilitate the reciprocal transaction of values for all participants to grow and benefit from this scaling.

References

Bonchek, Mark and Sangeet Paul Choudary (2013). “Three Elements of a Successful Platform Strategy”. In: *Harvard Business Review* (January). URL: https://hbr.org/2013/01/three-elements-of-a-successful-platform.

Chesbrough, Henry (2006). *Open Innovation: Researching a New Paradigm*. Oxford University Press.

Osterwalder, Alexander and Yves Pigneur (2010). *Business Model Generation – A Handbook For Visionaries, Game Changers, And Challengers*. Wiley.

Pisano, Gary (2015). “You Need an Innovation Strategy”. In: *Harvard Business Review* (June). URL: https://hbr.org/2015/06/you-need-an-innovation-strategy.

Sánchez-Cartas, Juan Manuel and Gonzalo León (2021). “Multisided Platforms and Markets: A Survey of the Theoretical Literature”. In: *Journal of Economic Surveys* 35 (2). URL: https://doi.org/10.1111/joes.12409.

Schrage, Michael (2004). “Interview in Ubiquity”. In: *ACM Ubiquity* (December). URL: http://ubiquity.acm.org.

Still, Kaisa, Heidi Korhonen, Miika Kumpulainen, Marko Seppänen, Arho Suominen, and Katri Valkokari (2017). “Business Model Innovation of Startups Developing Multisided Digital Platforms”. In: *IEEE 19th Conference on Business Informatics*. Vol. 2. Thessaloniki, Greece: IEEE, pp. 70–75.

Täuscher, Karl and Sven Laudien (2018). “Understanding Platform Business Models: A Mixed Methods Study of Digital Marketplaces”. In: *European Management Journal* 36 (3), pp. 319–329. DOI: 10.1016/j.emj.2017.06.005. URL: https://doi.org/10.1016/j.emj.2017.06.005.

Chapter 13
Sustaining the European Language Grid: Towards the ELG Legal Entity

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajič, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Katja Prinz, Jose Manuel Gómez-Pérez, and Ulrich Germann

Abstract When preparing the European Language Grid EU project proposal and designing the overall concept of the platform, the need for drawing up a long-term sustainability plan was abundantly evident. Already in the phase of developing the proposal, the centrepiece of the sustainability plan was what we called the "ELG legal entity", i. e., an independent organisation that would be able to take over operations, maintenace, extension and governance of the European Language Grid platform as well as managing and helping to coordinate its community. This chapter describes our current state of planning with regard to this legal entity. It explains the different options discussed and it presents the different products specified, which can be offered by the legal entity in the medium to long run. We also describe which legal form the organisation will take and how it will ensure the sustainability of ELG.

Georg Rehm · Katrin Marheinecke · Stefanie Hegele
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de, katrin.marheinecke@dfki.de, stefanie.hegele@dfki.de

Stelios Piperidis
Institute for Language and Speech Processing, R. C. "Athena", Greece, spip@athenarc.gr

Kalina Bontcheva
University of Sheffield, UK, k.bontcheva@sheffield.ac.uk

Jan Hajič
Charles University, Czech Republic, hajic@ufal.mff.cuni.cz

Khalid Choukri
ELDA, France, choukri@elda.org

Andrejs Vasiļjevs
Tilde, Latvia, andrejs@tilde.lv

Gerhard Backfried · Katja Prinz
HENSOLDT Analytics GmbH, Austria, katja.prinz@hensoldt.net, gerhard.backfried@hensoldt.net

Jose Manuel Gómez-Pérez
Expert AI, Spain, jmgomez@expert.ai

Ulrich Germann
University of Edinburgh, UK, ulrich.germann@ed.ac.uk

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_13

1 Introduction

One of the challenges the European Language Grid initiative aims to address is the fragmentation of the European Language Technology landscape, with regard to academia, research institutions and commercial entities. ELG aims to bring together all stakeholders, currently scattered all over Europe, under the European Language Grid platform as a common umbrella (Rehm et al. 2021; Vasiljevs et al. 2019). However, the efforts taken within the project can only be translated into a large-scale success if ELG continues to exist beyond the project runtime of 42 months. This is why it had already been foreseen in the ELG project proposal to develop a long-term sustainability plan during the project. Its centrepiece is the idea of establishing, in the second half of 2022, a dedicated ELG legal entity, which is meant to take over operations, maintenance, extension and governance of the European Language Grid platform as well as managing and helping to coordinate its community. Only with such a sustainable, long-term activity can the overarching goal of strengthening, harmonising and bringing together the European LT business and research community be met. In other words, the sustainability plan and the legal entity are mission-critical for the success of the project.

After a brief presentation of the long-term vision of ELG (Section 2), this chapter describes business and operation models that have been examined in order to assess if they are suitable for the ELG legal entity (Section 3). Not only shall the ELG platform and initiative continue to exist, we also want to expand its functionalities further in order to serve and adapt to evolving user needs even better and to fulfil ELG's mission for the European LT community. We explore a number of different dimensions with regard to the shaping of the ELG legal entity and place special emphasis on the description of a set of products we specified that can be offered by the legal entity. At the same time, it is important to point out that the AI landscape – including LT – must still be characterised as highly dynamic (Rehm et al. 2020b). Precise predictions of where the field is headed in Europe in the next years are difficult to be made right now. It remains to be seen what the post-COVID market will look like, which breakthroughs will come next in AI and LT, what the impact of the various ongoing large-scale initiatives will be and how the LT/AI-related situation in the different European countries will develop in the future. This dynamic situation creates additional challenges when it comes to specifying the final shape of the ELG legal entity, which must consequently correspond to this agile and dynamic environment.

2 Long-term Vision and Mission of ELG

Our vision and long-term goal is to establish ELG as the primary platform and marketplace for all commercial and non-commercial Language Technologies developed and offered by the European LT community. In order to achieve this goal, multiple prerequisites need to be in place, e. g., the ELG cloud platform must have very high availability and it must exhibit near real-time performance for individual services, legally safe service level agreements need to be prepared so that services can be applied in production environments, simple mechanisms for billing need to be available and technical support needs to be offered. Trust in the platform and its reliability need to be established in a transparent manner. Operating these and other components of the platform and initiative incurs various system-relevant costs (Teece 2017).

2.1 Mission of the European Language Grid

To achieve the goal of becoming the primary platform for European LTs, ELG follows its mission of creating impact beyond the platform itself:

- Grow a vibrant community and help coordinate all European LT activities: ELG is an initiative *from* the European LT community *for* the European LT community, including industry, innovation and research. ELG can only be successful if the whole community makes active use of the platform and contributes as well as uses datasets and services. ELG collaborates with many related projects, companies, research organisations and further initiatives (see Chapters 10 and 11), most notably its sister project European Language Equality (ELE), which is currently developing a strategic agenda and roadmap that specify how to achieve digital language equality in Europe by 2030. In the agenda developed by ELE, ELG functions as the main technology platform of the ELE Programme so that the support of Europe's languages through technologies can be measured and monitored over time (Gaspari et al. 2022; Grützner-Zahn and Rehm 2022).
- Create and maintain a powerful, scalable and useful Language Technology platform: ELG's novel technological approach enables innovations and synergies between commercial and non-commercial LT demanders, suppliers and users (see Chapter 12). The unique ELG platform is based on the principle of encapsulating services in containers. This approach tackles and solves some of the issues of technical interoperability, which is a crucial obstacle on the way of cross-provider and cross-platform interoperability. ELG enables providers to deposit and deploy their services.
- Support the Multilingual Digital Single Market: ELG strengthens the commercial European LT landscape through the pan-European platform and marketplace. Offering powerful multilingual, cross-lingual and monolingual technologies, ELG aims to contribute to the emergence of a truly connected, language-crossing Multilingual Digital Single Market. European companies can showcase and offer their LTs and consulting services to customers on the ELG marketplace (see Chapter 12).

2.2 Added Value for Stakeholders

The implementation of this mission in the form of the ELG platform provides added value for all stakeholders, e. g., 1. ability to attract participants (i. e., customers, buyers, users, providers etc.), 2. ability to create demand economies of scale, 3. benefit of reduced time-to-market (especially from lab to market), 4. standardised quality, 5. ease of doing business and a 6. coherent ELG technology exploitation ecosystem.

Traditional, linear value chains are focused on a one-way process of value creation, e. g., raw materials are used and manufactured into products, which are then distributed and used by the consumer, until they are disposed of. For ELG, we foresee a two- or multi-way value creation. As a digital platform, ELG will maintain an ecosystem of reciprocity. LT providers, LT consumers, ELG stakeholders and the whole ELG community help to generate two-way and reciprocal value as a result of the combination of resources of its participants, cost benefits (demand economies of scale) and network effects. As such, marketplace participants will create value by tapping into resources and capacities that they do not have to own themselves. In addition, marketplace participants will enjoy cost benefits and positive compound effects, arising from demand aggregation, from efficiencies in networks and from technological improvements on the demand side. Third, there is value within the network itself: growth via network effects will lead to market expansion for each of the members of the ecosystem. New participants (buyers and suppliers) enter the marketplace, because they are attracted to ELG by the growing number of participants who are also part of the network. That way, value is created in a reciprocal, multi-sided (almost infinite) way. For more details, see Chapter 12.

3 Main Pillars of the Business and Operational Model

Given the large number of possible routes to evaluate as well as decisions to be made eventually, we stretched the consortium-internal discussion of the main pillars of the ELG legal entity's business and operational model over the whole project duration, initiating the consortium-wide discussion in late 2019, i. e., we started immediately after the implementation of the proof of concept of the ELG platform. The goal was to specify, in a step by step fashion, the main ingredients of the sustainability plan. Relevant intermediate results were presented at META-FORUM 2020 and 2021 as well as in a number of talks.

At the very start of the overall process we looked at the setup and models of various other organisations that might serve as potential blueprints for ELG or, the other way around, as examples of organisations that would *not* work for ELG. We paid special attention to the domain of Language Technology and related fields, to the aspect of community-driven organisations, to combining industry and research and to the relevance of Europe as an overarching umbrella. All organisations we examined in more detail operate in the sphere of IT, LT or AI. Some of them have been created as spin-offs of research projects. With regard to their size and setup, though, these organisations are very diverse; the similarities with ELG in terms of their respective starting points and target groups also vary considerably. The organisations are: DBpedia Association[1], World Wide Web Consortium (W3C)[2], Industrial Data Spaces (IDS)[3], LT Innovate[4], OpenAIRE[5], CLARIN ERIC[6], Big Data Value Association (BDVA)[7], Translation Automation User Society (TAUS)[8], ELRA/ELDA[9] and GATE Cloud[10]. While discussing and learning more about these organisations – especially with regard to the type of legal entity they use, their membership as well as governance and fee structure, revenue streams etc. – we realised that despite some superficial similarities, none of them could serve as a direct model for the ELG legal entity, i. e., we are not aware of any organisation that could serve as an actual blueprint. However, we have been able to derive some important questions from this comparison that have informed the subsequent steps of the process.

The following sections present the main pillars of the legal entity approx. in the order in which we discussed and designed them.

3.1 Expectations by the ELG Consortium's SME Partners

Next up in the overall process of designing the ELG legal entity, we initiated a discussion with the ELG consortium's SME partners, primarily to collect their expectations and demands towards a legal entity that operates and maintains the "primary platform for Language Technology in Europe". The most important aspects of their considerations can be summarised as follows.

Sales channel: ELG is, first and foremost, understood as a channel to promote and to sell the products and services offered by the SMEs. ELG should stir interest and convince potential customers to invest in European LT. This is also true for public administrations and governmental bodies, the European Institutions and NGOs with the general idea being that interested parties and stakeholders look at ELG first in their procurement processes for LT. It was suggested that, in the medium to long run, ELG should consider fulfilling or even establishing certain quality and security standards as well as some kind of quality seal.

Strategy and collaboration: Europe has strengths in certain areas and language combinations but new business opportunities can only be reached by joining

[1] https://www.dbpedia.org

[2] https://www.w3.org

[3] http://www.industrialdataspace.org

[4] https://lt-innovate.org

[5] https://www.openaire.eu

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

[7] https://www.bdva.eu

[8] https://www.taus.net

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

[10] https://cloud.gate.ac.uk

forces and combining the offers with those of other European players. Missing or needed tools and services from others will help expand one's own set of tools and services. The SMEs expect ELG to help in this regard, i. e., identifying and closing strategic partnerships (also see *Interoperability* below).

Buy-in from the whole community: According to the SME partners, ELG must be positioned in the right way with regard to other platforms and infrastructures, e. g., a controlled transition from META-SHARE to ELG should be achieved by also integrating those organisations who have participated in META-SHARE. Furthermore, ELG should be backed, i. e., supported and *actively* used, by national centres and institutions. In terms of the governance model, all stakeholders should be able to have their say, yet dominance must be avoided. ELG can also provide a channel so that the results of national and international funding programmes can be disseminated efficiently on an international level.

Information channel: The goal is for ELG to become the primary European platform for participants from academia, research institutions and commercial entities. Especially with regard to industry, the relevance, understanding and benefits of LT for companies of all sizes needs to be increased. ELG could function as a means to keep interested stakeholders informed by serving as an information source and matchmaker for buyers and suppliers alike (marketplace approach, see Chapter 12).

Interoperability: 1. Throughout Europe, there is a sizable number of other relevant platform and infrastructure initiatives including, among others, Gaia-X[11], the European AI-on-demand platform[12], EOSC[13] and NFDI[14]. The SMEs mentioned their expectation that ELG becomes part of this larger ecosystem of platforms around Artificial Intelligence, data economy, research data management and Open Science, i. e., that ELG should ideally be fully interoperable with these other infrastructures, eventually opening up additional markets (Rehm et al. 2020a). 2. Furthermore, providers of LT need to understand what the requirements are to participate in ELG and why it is beneficial for them. ELG needs to be compatible with existing businesses and should not duplicate existing systems. Since various companies already operate their own or managed cloud platforms, platform interoperability should be ensured so that ELG complements existing or emerging clouds rather than appearing like competition. ELG should avoid creating the impression of being yet another collection of data and tools but rather emphasise the ability to combine services and resources from different companies. 3. For this, however, full interoperability on the level of the actual tools and services, i. e., on the level of APIs, annotations, semantic descriptions, closed vocabularies etc. needs to be achieved (also see *Strategy and collaboration* above).

[11] https://gaia-x.eu

[12] https://www.ai4europe.eu

[13] https://eosc.eu

[14] https://www.nfdi.de

3.2 Key Aspects of the ELG Legal Entity

Informed by the SME partners' expectations and other desk research we performed (see above), we started defining key aspects of the ELG legal entity, as follows:

Not-for-profit or for-profit organisation? There was a broad consensus in the consortium-internal discussions that the legal entity should be a not-for-profit organisation. This decision is rooted in the overall approach of ELG as an initiative *from* the European LT community *for* the European LT community. Moving into the for-profit direction would constitute a significant change of plan, effectively compromising the initiative's independence and ability to be perceived as neutral and non-competitive; this could also jeopardise the initiative's political standing with national and international administrations and funding agencies. In addition, the not-for-profit direction comes with additional benefits (e. g., in terms of taxation, more favourable funding conditions when participating in EU projects etc.).

Distributed team or central location? Due to the fact that the ELG consortium is already a distributed team and that the development of the platform and its technical infrastructure is spread across different European countries, the decision was made to keep this distributed setup and to build the team virtually rather than in one physical location. Current technical setups for remote work enable efficient virtual meetings and distributed teams are very common in business by now anyway, which is why we made this decision. The suggestion was made to position the legal entity's "headquarter" in the country where the majority of the costs are likely to be incurred, which, for the time being, will be the rented cloud infrastructure plus part of the personnel costs.

Start small or big? Given that developments in the AI/LT field and in Europe as a whole are very dynamic, the preparation of a detailed ten-year plan does not seem to be the right approach. A large organisation with a rigid hierarchical structure was perceived to be an obstacle in our consortium-internal discussions. Instead, we favour a flexible and agile setup that can react quickly and efficiently to changes and new framework conditions. However, the organisation must be large enough to ensure that the existing infrastructure and platform can be maintained and extended in a meaningful way and so that growth is possible. We currently assume a headcount of 10-15 employees for Phase 3 (see Table 1).

Abrupt transition or soft launch? While the ELG EU project will end on 30 June 2022, various partners of the ELG consortium are involved in a number of new projects, in which the European Language Grid plays a certain role. Through these new projects, some of the costs of operating the cloud platform can be covered. This situation is ideal because it gives the consortium a bit more time and flexibility for completing the overall setup of the legal entity. Our goal is to establish the legal entity in the second half of 2022, performing a rather soft launch.

Membership organisation? There are good reasons for having a setup that includes a membership structure, especially for actively including the many members of the European LT community and also because membership fees can be considered a constant, reliable source income if the ELG legal entity is able to continuously provide added value. On the other hand, the membership fee needs to be reasonable to make sure that interested parties are not deterred from the very outset. The specifics are still under discussion.

3.3 Assessment of Operational Costs

Operating the ELG legal entity will create costs, that need to be covered, even if the organisation itself will be a not-for-profit one. While the key tangible outcome of the EU project, the implemented and populated cloud platform, is an important prerequisite for the legal entity, several additional components need to be put in place. Crucially, the legal entity needs a team and director to take care of operations, maintenance and further development of the platform, associated tools and the ELG community. The main cost items are as follows.

Staff Labour costs represent the largest share of the organisation's expenses. Even a minimal team includes employees for operations, development, marketing, support and management. It might not be necessary to hire full-time employees for each of these areas right away but in order to run a successful organisation, a stable team is essential.

Cloud hosting To enable the legal entity to operate the ELG platform, a cloud infrastructure (including CPU, GPU, RAM, SSD and bandwidth) needs to be rented from a cloud service provider.

Overhead This refers to costs like rent of office space, hardware like workstations and printers, furniture, electricity, heating, etc. Even if remote and part-time work might reduce these costs because there is no need to rent larger office spaces, overhead still accounts for part of the fixed costs of the organisation.

Legal Especially in the ramp-up phase of an organisation, comprehensive and sound legal advice is crucial. The ELG legal entity will have to draw up and maintain model contracts and service level agreements for its products. Moreover, advice on GDPR, tax legislation and human resources issues is needed. The legal entity will not have the capacity for an inhouse legal expert, instead, legal services will be outsourced.

To facilitate future planning, a preliminary cost-structure has been developed (Table 1). It illustrates the foreseen soft start of the legal entity, which is separated into three phases. The gradual soft launch is meant to go from a small team that is working part-time (Phase 1) to a team of 10-15 full-time employees (Phase 3).

Cost Item	Phase 1 (start)	Phase 2 (ramp-up)	Phase 3 (stable)
Staff	2,500€	25,000€	100,000€
Cloud hosting	2,500€	10,000€	20,000€
Overhead	500€	2,500€	7,500€
Legal	–	2,500€	5,000€
Total	5,500€	40,000€	132,500€

Table 1 Estimated monthly costs in three phases (numbers are preliminary and indicative)

3.4 Business Model Canvas

The Business Model Canvas (BMC)[15] is a template used in strategic management for the development or documentation of existing or new business models. It is widely known and often serves as the first instrument applied when it comes to the visualisation and structuring of business models. The BMC helps to bring all essential elements of a business model into a scalable system. It consists of a visual chart with all necessary elements of an organisation or company. The idea is that the company or startup recognises its potential and weaknesses and understands where to align their activities by illustrating potential trade-offs (Osterwalder and Pigneur 2010). The nine "building blocks" of the business model design template that came to be called the Business Model Canvas were initially proposed by Osterwalder (2004) based on his work on a business model ontology. It outlines nine segments for the business model in a simple one-page canvas that can be inspected alongside each other. The nine BMC segments are: 1. Key Partners, 2. Key Activities, 3. Key Resources, 4. Value Proposition, 5. Customer Relationships, 6. Channels, 7. Customer Segments, 8. Cost Structure and 9. Revenue Streams. Below we explain how the ELG legal entity relates to each of the nine segments of the BMC. This ELG-specific BMC was prepared by all nine ELG consortium partners. First, we asked all partners to prepare a partner-specific BMC, i. e., to prepare their own vision and approach of the ELG legal entity. Afterwards we processed the nine individual, partner-specific BMCs into one consolidated BMC, which is the basis of the following description.

Segment: Key Partners "Who are the key partners/suppliers? What are the motivations for the partnerships?"
One key partner in the ELG BMC are commercial and non-commercial LT service providers, either with or without their own cloud platform. Equally important are Language Resource and data providers that own existing data sets and repositories. These two key partners contribute to the thriving of the ELG platform. Their motivation is not (or not only) to use available services and resources, but they offer their own services and resources and create value or profit for their own organisations. Another key partner is the wider ELG community, including the ELG consortium, the 32 National Competence Centres, the national language communities, and all running EU projects and initiatives in the field of LT (includ-

[15] https://en.wikipedia.org/wiki/Business_Model_Canvas

ing ELE). This community consists of academic and research partners as well as a number of companies that need multilingual datasets and services for their research. Equally important for raising awareness are the European Commission and the European Parliament as well as national institutions such as ministries and funding agencies and other established networks and associations.

Segment: Key Activities "What key activities does the value proposition require? What activities are the most important in distribution channels, customer relationships, revenue stream, etc.?"

The most crucial key activity is the maintenance, further development and operation of the ELG platform. It needs to provide an interesting and relevant offering in order to grow a critical mass of members and users and gain popularity in the whole European LT community and beyond. Regular posting of content and other outreach activities (such as events, tutorials, talks, publications, meetups etc.) are essential to generate visibility and create a strong reputation (see Chapter 10). All communication and dissemination activities have to be treated with the highest priority to retain existing users and keep attracting new ones. Leveraging existing communication networks and sales channels can support this process and will be further explored. Quick and reliable service and support helpdesks are needed to strengthen customer relationships. Licensing and billing models need to be maintained and promoted. Maintenance and management of cloud storage and computing for running services has to be ensured.

Segment: Key Resources "What key resources does the value proposition require? What resources are the most important ones in distribution channels, customer relationships, revenue stream etc.?"

The most important resource is the ELG platform itself with all its functionalities and included services, corpora and additional information. ELG can be regarded as a set of seed technologies, tools and components that are extended over time. Customer feedback can be seen as a useful resource as well. It can come in many different forms such as evaluation from market data or helpdesk and user support feedback. Equally important is a dedicated ELG team, committed to not only maintaining existing technology, but growing it and promoting the importance of ELG on an international level. To achieve this, a wide international network is a key resource. The consortium combines vast experience and expertise, good knowledge of ongoing trends and access to numerous European networks in academia and industry.

Segment: Value Proposition "Which customer needs are being satisfied? What core value is delivered to the customer?"

ELG is envisioned to become the primary LT platform for Europe and to function as a one-stop-shop, offering a rich portfolio of LT services, tools and datasets. One of its core values is the availability of state of the art services which are fast, effective, robust and high-quality. Another special attribute is the fact that ELG is "made *in* Europe, *for* Europe". This strong branding inspires trust and confidence and ensures that the system is compliant with European regulations, security constraints and ethics. For customer satisfaction, ELG needs to be customisable, cover niches, address verticals and offer direct access to providers. Furthermore, all solutions come with high usability and are easy to integrate. Stakeholders familiar with the European LT landscape are aware of the fragmentation of the community which impairs an effective exchange of resources. ELG is committed to tackle this existing fragmentation. Competitive pricing is another value that makes ELG attractive for customers. Unique about ELG is that it offers a new or additional channel for service providers and consumers. Suppliers can gain more visibility, easy portability between providers is guaranteed through joint standards. Workflow functionalities will eventually be integrated to combine services from different providers and even their own clouds. ELG also offers added value to academia. It allows the use of services and data and offers easy comparison between systems on the same data or different data with the same system. ELG is meant to act as a broker for European LT and as a catalyst to boost innovation that also makes both the European industry LT sector and academic institutions an attractive employer for young high-potentials.

Segment: Customer Relationships "What relationship that the target customer expects are you going to establish? How can you integrate that into your business in terms of cost and format?"

The ELG brand is intended to be a quality seal for customers that guarantees state of the art services, a high level of security and compliance with all relevant EU regulations. Customers can use ELG through the web UI including code samples and libraries or through the APIs or SDKs. High quality guidelines and a user-friendly design make processes intuitive. Support through a service helpdesk is also possible. Technical onboarding and support packages will be offered and a fine-grained customer relationship model is being developed. Essential for targeting customers is strong brand building. Related marketing activities are tailored to different audiences and distributed regularly. While retaining customers is essential, new potential customers can be attracted through outreach and training events, tutorials, webinars and conferences. A brand that has earned people's trust can also create a need for other customer services such as consulting services around ELG and language-centric AI.

Segment: Channels "Through which channels do customers want to be reached? Which channels work best? How much do they cost? How can they be integrated into customers' routines?"

Customers will be reached through a variety of channels. Events, both established and new ones, will play an important role, for example, events targeted at stakeholders in a specific industry domain. Dedicated networking sessions, conferences and presentations are also foreseen. Online advertising campaigns will accompany all events. Since ELG builds on an existing network of stakeholders, email marketing and social media campaigns have proven to be successful means of reaching out. Presence on social media channels such as Twitter or LinkedIn helps to promote events and maintain customer relationships. ELG itself is a channel through which customers can retrieve information, not only about services and datasets, but also about the community and events. Cloud platforms that are either currently being developed in other EU or national projects as well as existing commercial platforms can also act as channels to point potential customers to ELG. SEO can also help promoting ELG since users trust search engines.

Segment: Customer Segments "For which segment is value being created? Who is the most important customer?"
The ELG platform offers value to different customer segments. LT providers, both commercial and academic ones, can use ELG to offer their services and datasets. Research organisations can benefit immensely from the wide offer. Customers from industry that demand LT (including large enterprises, SMEs, startups etc.) represent an essential customer segment that contributes to turning ELG into a flourishing marketplace. The European Union, public administrations and NGOs can also integrate ELG services into their current solutions. The same holds true for funding agencies and policy makers, advertising companies etc. Other EU project consortia as well as project consortia on the national level can benefit from the value created by ELG.

Segment: Cost Structure "What are the highest costs? Which key resources or activities are most expensive?"
As mentioned earlier, the highest costs are created by the human resources and the digital infrastructure. Personnel costs are created by the team maintaining and further developing ELG including daily operations as well as customer support, but also community management work that requires marketing and communication activities. Further resources need to be assigned to management and administration work that includes budgeting, accounting and legal counselling. Moreover, overhead costs are to be covered.

Segment: Revenue Streams "For what value are customers willing to pay?"
Part of the overall revenue will be generated through different products including usage or subscription fees, brokerage fees (marketplace approach), commission fees and products such as LT as a Service (LTaaS; hosting of services, models, datasets), LT Platform as a Service (PaaS; combining ELG services into workflows) and Repository as a Service (RaaS; hosting service for whole repositories). Advertisements can, for instance, showcase companies, services, conferences etc. Sponsored content, services, data sets, companies etc. present another revenue stream as well as commission fees. Paid training events, tutorials, webinars etc. can be offered to commercial stakeholders. Conferences (event registration fees; sponsorship packages for companies) are also an opportunity to generate income as well as general consulting services around ELG and language-centric AI.

This brief summary of the nine segments is an extract of the ELG BMC, produced by consolidating the BMCs prepared by the ELG consortium partners. For many segments, there was broad agreement within the individual BMCs, especially with regard to *key partners*, *key activities* and *key resources*. Also, in *value proposition*, *customer relationships* and *channels* the answers were largely similar. The *customer segments* are quite heterogeneous, though, which may make a targeted approach more difficult. As far as the *cost structure* is concerned, there are few deviations. A crucial open question concerns the appropriate size and ambition of the ELG, in particular with regard to team size. The answers were rather diverse in the case of *revenue streams*; here, positions could be aligned more closely through the subsequent step of specifying and discussing the different ELG products. As a follow-up step, the exact revenue streams will be evaluated with regard to cost-effectiveness and sustainability.

3.5 Product Portfolio and Revenue Streams

Together with all partners of the ELG consortium we defined, in a process that included several iterations, a portfolio of products that the ELG legal entity can potentially offer. These products are targeted at members of the European LT community and also at stakeholders interested in using, implementing, integrating or purchasing European LT. The products are primarily foreseen as revenue streams for the ELG legal entity so that it is able to cover the fixed costs associated with operating the ELG legal entity and platform (Section 3.3).

Such a structured portfolio of products, including associated fees, is necessary for eventually preparing the budget plan of the legal entity. In the following, we briefly describe the main categories of the ELG product portfolio; due to space restrictions we are unable to include all the details (especially aspects such as competitors, pricing, technical preconditions and general prerequisites are left out), i. e., the description in this chapter is not meant to be exhaustive but rather indicative of the overall plan and vision of the legal entity. It is also important to note that not all products will be offered right from the start but that the set of products will be expanded gradually over time.

3.5.1 Product Category: Marketplace

Marketplace Commission ELG features a directory of all European LT developers and can enable a match-making process, i. e., ELG facilitates, for potential buyers or integrators of LT, the discovery of the right LT provider. In this product, ELG receives a commission from every contract generated through the marketplace (approx. 5-10%). This product can be used by commercial LT developers to broaden their reach and to penetrate new markets, especially if the current is limited or if the developer is operating in a niche. On the demand side, we foresee this product to be used by larger organisations that want to buy LT or integrators that need a specific LT for a customer project. In order to participate in this marketplace, LT developer companies have to agree and to sign a marketplace participation framework agreement.

Public Request for Bids Model This product is a potential extension of the *marketplace commission* product: Customers can publicly and maybe anonymously post the need for a certain technology or resource or perhaps for an integration task and ask supplier companies for bids. Multiple LT developers and integrators can post their bids (not publicly) so that the organisation that posted the original request for bids can identify a cost-effective way to move forward. Posting the original request for bids would require a small fee to be paid. If a contract is established, the usual ELG marketplace fee applies on top of this.

3.5.2 Product Category: Consulting

Technical ELG Platform Consulting The ELG legal entity has enough expertise so that it can offer various types of technical consulting services, for example, regarding ELG, providing or using ELG services, combining services, training new models and making them available, i. e., services with a clear focus on the ELG platform, ecosystem and technical basis. This product is likely to be purchased by organisations that have a certain need for LT and that want to test and explore certain functionalities, models or tasks, but these organisations realise that they need some kind of help, e. g., implementation of prototypes, selection of technologies, evaluations etc. Using this product, organisations are able to make full use of the ELG platform and all its services. This product can be offered for a one-time fee or, for larger companies, also as part of a framework contract.

Conceptual ELG Community Consulting This product is similar to the one described above; it primarily makes use of the ELG team's in-depth knowledge of the ELG community, i. e., of the European LT developer or provider landscape. In that regard, the ELG team can support organisations with a certain need for a general or specific type of LT in finding the right technology provider. Customers interested in this type of product know that they have a certain need for LT but they are unsure about the concrete next steps, i. e., where and how to find the provider company.

LT Market Intelligence Report The ELG legal entity could exploit its in-depth knowledge of the European LT landscape and community and publish an annual or semi-annual market intelligence report about the European and maybe also global LT landscape including topics such as, among others, emerging trends, new players and rising stars, new projects and success stories. Such market analyses are highly relevant for a larger group of stakeholders including larger companies and enterprises (LT developers, LT users), non-governmental organisations, venture capital companies and others. These reports could be offered for a one-time fee or as packages that cover multiple reports with a slightly reduced fee.

3.5.3 Product Category: ELG APIs

ELG Power User Flatrate (for commercial users) Through this product, commercial customers get unlimited and unrestricted access to the ELG APIs of all integrated services and tools. This product targets companies of any type (SMEs, integrators, enterprises) that have to pay a small monthly or annual fee to be able to use it. This subscription product provides direct to all ELG APIs for experimentation and evaluation purposes, enabling fast comparisons and immediate results.

It can also be used to develop smaller LT-driven applications by integrating ELG APIs into existing systems. Like with many other products, any surplus generated through this product will be transferred to those LT developing companies that have provided the ELG-integrated services that were used in the relevant month, based on the proportionate number of API calls.

ELG Power User Flatrate (for academic users) Technically, this product is exactly like the first one but it targets academic users exclusively. The monthly or annual fee will be significantly lower than the fee of the power user flatrate for commercial users.

ELG Professional Flatrate Conceptually, this product is similar to the first one but the professional flatrate includes additional features and support services, e. g., faster tools, more compute resources, faster helpdesk support, workflow or pipeline functionality etc. The price of this product will be significantly higher than the pricer for the first product.

3.5.4 Product Category: LT-as-a-Service

LT-as-a-Service (for commercial users) This product targets commercial LT developers. Paying a certain fee, it enables them to host a limited number of LT tools or services within the ELG platform with guaranteed performance and availability. In order to be able to host more services or API endpoints in ELG, a different type of product needs to be purchased (see Section 3.5.6). This product is especially interesting for those companies that do not operate their own cloud infrastructures or that are eager to participate in the ELG initiative, i. e., ELG's LT-as-a-Service product can be seen as an alternative to renting cloud infrastructure. Another benefit of this product is that companies are able to extend their reach and to open up new markets, i. e., once again ELG can be used as an additional sales, promotion and distribution channel. This product can also be set up in multiple tiers, representing different maximum numbers of services and corresponding prices. While companies have to pay a certain fee for this product, the different ELG APIs products (see Section 3.5.3) will generate revenue, from which the companies will benefit. In that regard, it is important to identify the right balance over time.

LT-as-a-Service (for academic users) Technically, this product is exactly like the previous one but it targets academic users exclusively. The monthly or annual fee will be significantly lower than the fee of the LT-as-a-Service product for commercial users. This product also targets research projects, for which ELG can function as a secondary or maybe even primary dissemination and exploitation channel for their research results. Like the ELG power user flatrate for academic users, we consider making this product available for free for academic users if and when the ELG legal entity has established stable revenue streams.

3.5.5 Product Category: Data-as-a-Service

Data-as-a-Service (for commercial users) This product is very similar to LT-as-a-Service but instead of focusing upon running services or tools, it only allows making datasets or other (static) resources available on ELG, again, with guaranteed availability. Like LT-as-a-Service, this is an entry level product and, thus, only allows hosting a limited number of datasets (or up to a certain amount of data) on ELG. In case of more demand on the side of the customer, a different type of product needs to be purchased (see Section 3.5.6). This product needs to be priced lower than the LT-as-a-Service product.

Data-as-a-Service (for academic users) Technically, this product is like the previous one but it targets academic users. The monthly or annual fee will be significantly lower than the fee of the Data-as-a-Service product for commercial users.

3.5.6 Product Category: Repository-as-a-Service, Platform-as-a-Service

Repository-as-a-Service, Platform-as-a-Service Using this product, customers can host whole LT platforms or repositories on ELG while the ELG team takes care of all technical aspects including branding, availability, backups etc. This product targets a variety of stakeholders including goverments and ministries (e. g., for hosting national LT platforms on ELG), smaller or larger companies, smaller research groups and also whole research centres. The idea behind the product is that setting up and operating a cloud with an LT repository requires a lot of effort and expertise, which can be fully avoided by purchasing the corresponding ELG product. While the branding of the respective hosted platform or repository can be adapted to the brand and logo of the respective customer on the user interface level, at the same time, all hosted services, tools and other resources are automatically also part of the 'wider' ELG cloud platform, which will automatically broaden their reach significantly. We currently foresee three different tiers of this product: one entry level tier for research groups, one for SMEs and research centres and one for national LT repositories.

3.5.7 Product Category: Events

Training Events and Tutorials In addition to the more technical products described above, training events and tutorials can be offered as products, especially for commercial customers. These can be, among others, general ELG-related training events (from half a day to multiple days) where the training relates to the ELG platform, using, providing and combining services etc. This type of event can be offered to organisations that have a need for certain LT and that want to be able to make the most of the ELG platform. This product is a pre-packaged and generic course, while those training events that involve customisation of content, tailoring the course to the respective customer and its specific needs, would be rather considered technical ELG platform consulting (see Section 3.5.2).

Annual Conference The annual ELG conference assembles the whole ELG community, including commercial and academic participants, related projects and initiatives etc., and also the ELG team. While the annual ELG conference organised by the ELG EU project has been free of charge, this model could change (e. g., registration fees, sponsorship packages, paid presentation slots, booths for a fee in the industry exhibition etc.).

3.5.8 Product Category: Marketing and Advertisements

Conference Sponsorship This product relates to typical conference sponsorship packages, which can be purchased by, typically, companies to position themselves as gold or platinum sponsors of the annual ELG conference. This product model is well established and accepted in industry and research but to be successful it requires the respective conference to be of very high relevance for its community.

Online Advertisements The ELG platform could offer a small part of its screen real estate for online advertisements that can be purchased, among others, by members of European LT community to position their products or services in a more targeted way on the ELG website, for example, when certain keywords or search terms are used. In terms of revenue generated, this product only makes sense if the website has a very high number of users. Furthermore, it remains to be discussed and seen if online ads are a welcome addition on the ELG website or if they are perceived as not appropriate.

Sponsored Content Similar to online advertisements, the idea behind this product is that customers can pay a small fee to get one or more of their products, services or resources or perhaps even their own organisation's or project's page in the ELG catalogue featured on the ELG website, clearly marked as "sponsored content" (for example, the first search result).

Merchandise The final product relates to ELG-branded merchandise, which could be sold online, for example, tshirts, hats or pens with the ELG logo.

3.5.9 Miscellaneous

In addition to the actual products offered by the ELG legal entity, there are at least three other potential revenue streams or activities related to marketing the ELG products. These additional revenue streams cannot be considered products per se.

Foundations The ELG legal entity could approach one or more foundations with the request to grant financial support. In return, the foundations could position themselves as supporters of the ELG initiative.

Project Grants EU or national project grants are an obvious mechanism to support part of the ELG team and platform as well as its operation.

ELG Use Cases as Show Cases Together with larger enterprises and some of the commercial LT developer companies represented in ELG, interesting and relevant show cases as well as success stories can be published on the ELG website, which can function as marketing instruments and testimonials that demonstrate that ELG is an important and valuable activity.

3.5.10 Summary and Assessment

The ELG product portfolio is diverse and broad, it offers multiple different options of moving forward under the umbrella of the legal entity. As mentioned, we will not start with all products right away but only with a selection. Before we make the final selection, we will validate the products and their chance of being accepted by the European LT community with a number of experts from the field. As the most promising products we currently perceive the ELG APIs (Section 3.5.3) due to the enormous market for this product, the LT-as-a-Service products (Section 3.5.4) due to high demand, the marketplace (Section 3.5.1) as well as the consulting product (Section 3.5.2).

Additionally, we see a lot of potential in offering countries the technical infrastructure for the purpose of supporting national LT platforms (Section 3.5.6). Especially for smaller countries or regions, it is challenging to develop, operate and maintain an elaborate technical platform all by themselves. For these, having their *National LR/LT Repository* hosted as a service within ELG can be an attractive offer. For ELG, in return, it appears to be an interesting financial pillar to operate such platforms, charging an annual hosting fee.

Making use of the ELG platform as the primary dissemination and exploitation channel for research projects is another product idea that has a lot of potential (Section 3.5.4). It enables research projects to fully concentrate on the actual research work without a need for developing complicated exploitation plans on their own because they can fully rely on ELG for this purpose. This approach can increase the general visibility of European research results significantly.

3.6 Legal Entity Type

For the creation of a dedicated legal entity with European scope, we considered a number of different entity types. The decision to move forward with a not-for-profit organisation was made rather early in the process. The main options that we explored were a professional association or a foundation. In that regard, each EU country has its own set of different types of business entities as part of their legal system, which, generally, all have their own specific sets of rules. These include, among others, cooperatives, partnerships and limited liability companies. Looking at Germany, for a not-for-profit organisation, a *gGmbH* (a not-for-profit private limited company), or an *e. V.* (eingetragener Verein, registered association) would be two obvious options.

An alternative that enjoys some popularity with EU-funded projects is the Belgian Association without lucrative purpose (*AISBL*). As the ELG consortium does not have any partners in Belgium or Luxembourg, the AISBL option was ruled out for reasons of efficiency. In addition to national entities, there are several types of legal entities on the level of the European Union.

The *EEIG* (European Economic Interest Grouping) is part of European Corporate Law, created in 1985. An EEIG makes it easier for companies in different countries to do business together. Its activities must be ancillary to those of its members. Any profit or loss is attributed to its members. It is liable for VAT and social insurance of its employees but it is not liable to corporation tax and it has unlimited liability. Several thousand EEIGs exist and are active in various fields. This legal entity only applies to companies, it does not include research institutions.

The *SE* (Societas Europea) is a European company, established in 2001 by an EU Regulation. The SE has been growing in popularity ever since. It is a type of public limited-liability company and allows an organisation to operate its business in different European countries under the same rules. An SE offers many advantages such as easily setting up Europe-wide subsidiaries as well as an international holding company. The company headquarters can be relocated easily and the SE legal form conveys a strong European image. However, the SE comes with strict foundation criteria, such as the requirement of high initial capital.

The *SCE* (Societas cooperativa Europaea, European Cooperative Company) was established in 2006, it is related to the SE. An SCE can be established in the European Economy Area. This entity type was created to remove the need for cooperatives to establish subsidiaries in each EU Member State in which they operate, and to allow them to move their registered office and headquarters from one EU Member State to another. SCEs are governed by a single EEA-wide set of rules and principles which are supplemented by the laws on cooperatives in each Member State.

The *SPE* (Societas privata Europaea) is a European private limited company, it corresponds to an Ltd. in Anglo-Saxon countries or a GmbH in Germany, Austria and Switzerland. This legal entity type has been a European Commission proposal for more than ten years. As of now, it still does not exist.

For ELG, a crucial requirement is that the selected solution provides flexibility, agility and the ability to ramp up the operation of the legal entity in a careful way. The final decision must also be made on the basis of financial considerations, i. e., it must be specified which products or services can be offered to generate which profit.

At the time of writing, we will establish a registered association headquartered in Germany (e. V., *eingetragener Verein*). This option does not require any initial capital and frees ELG from the pressure of having to generate income immediately. Since some of the staff members who will be active in the ELG e. V. in the first phase are based in Berlin, it appears practical to set up the entity in Germany and under German law. It must be noted, however, that the legal entity will work in virtual teams primarily. The only legal entity type on the European level that could be appropriate for ELG, the SPE, does not exist yet.

4 Summary and Next Steps

This chapter presents the current state of planning of the ELG legal entity, which is foreseen to be established as an *eingetragener Verein, e. V.*, as a registered, not-for-profit association, in the second half of 2022. The legal entity will start small, with a soft launch, and is meant to be flexible and agile. The main pillars of this concept have been under development since late 2019 and cover most of the crucial aspects of the legal entity. In terms of financing, a mixed model is envisaged, driven by the product portfolio (Section 3.5), that includes shared revenue streams through LT provider companies that use ELG as a sales channel and their customers who use ELG to find the right providers and suppliers as well as services.

One aspect that still needs to be specified in more detail is the inclusion and active involvement of the European LT community and the governance structure of the legal entity. As an initiative *from* the European LT community *for* the community, its involvement is crucial to create trust and transparency as well as to provide representation to academic and industrial European LT developers. The proper inclusion of the community in a representative manner will require a number of discussions and deliberations. Fortunately, with regard to an *e. V.*, these matters do not need to be fully resolved before establishing the organisation but can also be taken on board and revised through updates of its statutes.

Originally we had envisioned to establish the legal entity within the project runtime and to start with a 'bigger' approach than is currently foreseen. The aforementioned delay of a few months in establishing the entity does not pose a problem because the overall framework conditions have changed in the last 12 to 18 months. Through recently started and publicly funded projects including ELE, ELE2, OpenGPT-X, NFDI4DataScience and AI as well as the upcoming EU projects DataBri-X and SciLake, which are about to start in October 2022 and early 2023 respectively, we are able to operate the ELG cloud platform and we can also perform some maintenance and other ELG-related work, including the extension of the ELG platform itself so that it is compatible with the emerging Gaia-X ecosystem. In addition, SciLake will establish the first bridges to the EOSC ecosystem.

Since the start of the project, we have been collaborating with the European AI on demand platform, especially with the AI4EU project, to ensure compatibility of our approaches in terms of semantically describing resources. Furthering these collaborative efforts will facilitate cross-platform search and discovery enabling ELG resources and other assets to be visible and usable by the wider AI community. Considering the EU's plan to deploy the European AI on demand platform, ELG is ready to act as the central language-related AI hub and marketplace providing access to and direct use of several thousands of LT services and related data.

While the future is always difficult to predict, it is clear already now that over the past three years the interest in ELG has risen constantly and that the legal entity that will take over the initiative after the EU project has ended has very good starting conditions. The ELG brand has been established in the community and a considerable buy-in can be observed already now. However, to take advantage of this momentum, the marketplace, broker, dissemination, exploitation and participation model needs to be extremely simple and easy to grasp to make sure users understand and accept it and the platform needs to be as user-friendly and all-encompassing as possible in every regard, including the various levels of technical interoperability. Quality and security aspects play a crucial role and can become the unique selling proposition as opposed to providers of LT services from the US or Asia.

References

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022). “Introducing the Digital Language Equality Metric: Technological Factors”. In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). “Introducing the Digital Language Equality Metric: Contextual Factors”. In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. URL: http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Osterwalder, Alexander (2004). “The Business Model Ontology: A Proposition in a Design Science Approach”. PhD thesis. France: University of Lausanne.

Osterwalder, Alexander and Yves Pigneur (2010). *Business Model Generation – A Handbook For Visionaries, Game Changers, And Challengers*. Wiley.

Rehm, Georg, Dimitrios Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julián Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John P. McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš (2020a). “Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability”. In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France, pp. 96–107. URL: https://www.aclweb.org/anthology/2020.iwltp-1.15.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020b). “The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe”. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. URL: https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Teece, David J. (2017). "Dynamic Capabilities and (Digital) Platform Lifecycles". In: *Entrepreneurship, Innovation, and Platforms* 37 (Advances in Strategic Management), pp. 211–225. DOI: 10.1108/S0742-332220170000037008.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. URL: https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Part IV
ELG Open Calls and Pilot Projects

Chapter 14
Open Calls and Pilot Projects

Lukáš Kačena, Jana Hamrlová, and Jan Hajič

Abstract We describe the two ELG open calls for pilot projects, the objective of which was to demonstrate the use and the advantages of ELG in providing basic LT for applications and as a basis for more advanced LT-based modules or components useful to industry. Our main goal was to attract SMEs and research organisations to either contribute additional tools or resources to the ELG platform (type A pilot projects) or develop applications using Language Technologies available in the ELG platform (type B pilot projects). We start with the detailed description of the submission and evaluation processes, followed by a presentation of the open call results. Afterwards we describe the supervision and evaluation of the execution phase of the projects, as well as lessons learned. Overall, we were very satisfied with the setup and with the results of the pilot projects, which demonstrate an enormous interest in ELG and the Language Technology topic in general.

1 Introduction

To demonstrate the advantages of ELG (Rehm et al. 2021) in providing LT for applications and as a basis for more advanced LT-based modules or components useful to industry, the ELG project set up a mechanism for using close to 30% of its budget for small scale demonstrator projects ("pilots") through two open calls. The calls were prepared using the ICT-29a call specification, making use of the Financial Support to Third Parties (FSTP) scheme according to the ICT Work Programme 2018-2020 (European Commission 2017). In total, we provided 1,950,000€ to the selected projects as FSTP with an awarded amount of up to 200,000€ per project. We established a lightweight submission procedure and a transparent evaluation process, in which external evaluators participated as reviewers.

The main objective of the open calls was to attract SMEs and research organisations to either contribute tools and services to ELG (type A projects) or develop

Lukáš Kačena · Jana Hamrlová · Jan Hajič
Charles University, Czech Republic, kacena@ufal.mff.cuni.cz, hamrlova@ufal.mff.cuni.cz, hajic@ufal.mff.cuni.cz

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_14

applications using Language Technologies available in the ELG platform (type B projects). The results of the pilot projects are included in the ELG platform for dissemination, testing and external evaluation by other entities or the public.

2 Organisation of the Open Calls

2.1 Management Structure and Organisation

While agile, simple and lightweight from the proposers' point of view, the organisation of the two open calls was an internally complex procedure requiring close collaboration of three different teams (management team, technical team, Pilot Board) with support from a broad panel of external evaluators.

2.1.1 Pilot Board

The Pilot Board (PB) was set up for the supervision of the pilot projects. While the management team took care of the organisation and handling of the open calls and the execution of the pilots, the PB provided a forum so that the ELG project could discuss the progress of the pilots, their feedback and results. The PB was meant to be the main technical and strategic interface between the pilot projects and the ELG project proper, so that ELG could maximise its benefits from supporting the pilots and to make sure that the pilot projects benefit from ELG.

The PB operational procedures were drafted by the management team and approved by the ELG Steering Committee. Afterwards, seven PB members were nominated and approved. The operational procedures defined the main responsibilities of the PB as follows: approval of the open calls and related documentation; pilot project selection process; supervision of pilot project execution, including progress monitoring, evaluation of results and approval of the phased payments.

2.1.2 External Evaluators

An independent panel of experienced external evaluators ensured an open, transparent and expert-evaluation based selection process. The pool of evaluators was created using a separate open call. The evaluators were responsible for evaluating the project proposals and worked remotely using the web interface of the ELG Open Calls Platform. They were selected from the pool, avoiding any conflicts of interest. All evaluators were asked to sign a non-conflict of interest declaration and a confidentiality agreement before being accepted to perform the task.

2.1.3 Management Team

The management team organised the whole Open Calls process, including managing and directing the technical team. In line with Annex K of the Work Programme (European Commission 2017) and other relevant sections of the Rules for Participation, the management team prepared all prerequisites and procedures: the Open Calls Platform, web content, informational materials, forms, contract templates, presentation and reporting forms and templates, submission procedure, hiring and selection of external evaluators, call management structure, internal auditing and project results evaluation procedures. In the initial setup phase, the management team tapped the legal and financial expertise of the Technology Centre of the Czech Academy of Sciences, which is charged by the Czech government to host the National Contact Point (NCP) and other experts related to the preparation, execution and evaluation of EU framework programmes and projects.

2.1.4 Technical Team

An essential task was to set up the ELG Open Calls Platform for the proposal submission, evaluation and reporting process. We decided to develop the platform in-house to ensure that it fit our needs.[1] The technical team was responsible for developing the platform and for support during each phase of the process.

2.2 Timeline

Figure 1 shows the open calls execution timeline. After the announcement, each call was open for submissions for two months, followed by an evaluation procedure of approx. two months. After signing the contracts with the selected projects, the execution phase started. The expected project duration was 9-12 months. Four projects asked for a short extension of one or two months (which was accepted), mainly due to COVID-19 related delays of dissemination activities.

2.3 Communication with Stakeholders

Prospective applicants were targeted through various channels, e. g., the open calls website, a survey for stakeholders and other communication and dissemination activities carried out by all ELG consortium members.

From early 2019 onwards, the open calls were presented on the ELG website.[2] The content was regularly updated, starting from basic information including the

[1] https://opencalls.european-language-grid.eu

[2] https://www.european-language-grid.eu/open-calls

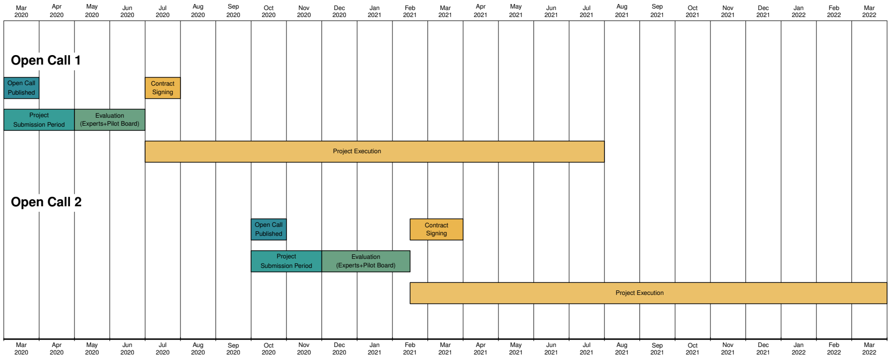


Fig. 1 Open Calls overall timeline

timeline and key parameters at the beginning of the project, followed by the call for evaluators[3] and complete information regarding the open calls[4].

We first monitored the interest in the open calls using a survey, which ran from May 2019 until June 2019. A total of 108 respondents participated. The result showed significant interest in the open calls and also a high demand for more information. Five months before the first call announcement, a second survey was prepared. We disseminated this survey during the first annual ELG conference META-FORUM 2019 in October 2019 in Brussels and collected answers from 47 respondents, 84% of which expressed an interest in taking part in the open calls.

The open calls were promoted through social media (Twitter, LinkedIn), various e-mail distribution lists, internal networks and collaborators, through the META-FORUM conference and through other means whenever an opportunity arose.

2.4 Submission Process

As explained in the previous section, in the preparatory period the overall open call procedure was set up, including all related documents and the development of the online platform for the management and evaluation of submissions. After the official announcement of one of our two open calls, applicants could then prepare and submit their project proposals. There was a continuous need for support, mainly answering questions we received by the participants via email.

With regard to the call announcement, we paid special attention to a well-prepared call documentation, which provided all necessary information for applicants, and a user-friendly submission platform. The documentation was prepared as an easy-to-understand document. It contained several annexes: Guide for Applicants, Third Party Agreement, Project Proposal Template and Evaluation Criteria.

[3] https://www.european-language-grid.eu/open-calls/call-for-evaluators

[4] https://www.european-language-grid.eu/open-calls

In the "Guide for Applicants" the management team showed, using screenshots, how to submit a project proposal through the platform, i. e., how to create an applicant account, how to log in and manage the account, how to create a new project proposal, fill in the forms and finally submit the proposal. We also maintained a list of (expected) frequently asked questions, for example "Who can apply for a pilot project?", "How much money is allocated for the pilot projects?", and "Does Brexit have any implications on eligibility?".

The Open Calls Platform was developed using the open source Content Management System Drupal with the guiding principle to keep the submission and evaluation process easy and straightforward for the participants and manageable for the call organisers. The platform runs under the ELG domain[5], while physically residing with the technical team to ensure quick reactions to any technical problems.

2.5 Evaluation Process

2.5.1 Preparation of the Evaluation Process

The most important part of the preparation of the evaluation process was the selection and specification of evaluation criteria that match the objectives to be achieved by the calls. At the same time, the criteria ought to be clear for the external experts evaluating each proposal.

The criteria were defined and described in detail in the call documentation. First, the submitted proposal should fulfill formal requirements (language, submission date, declaration of honor, legal status, eligible country, number of submitted proposals per applicant and no conflict of interest) which were checked by the management team before any further evaluation. Then, three independent evaluators checked the binary eligibility criteria: uniqueness, relevance for ELG, and whether the proposal contains all the required phases (experiment, integration, dissemination). These were followed by the graded and ranked evaluation criteria: objective fit, technical approach, business, integration and dissemination plan, budget adequacy, and team.

In order to identify evaluators with experience in language technologies and evaluation, a call for evaluators was published in February 2020. All relevant information (description of tasks, eligibility of candidates, selection criteria, contact email for questions, and a link to the registration form on the Open Call platform) was published on the ELG website as well as on the European Commission Funding and Tender portal. In addition, ELG consortium members disseminated the call through various channels. Potential evaluators were asked to fill in a registration form, through which contact information, CV, and professional experience related to evaluation and LT were collected. From about 156 applications, the management team selected 64 evaluators (a total for both project open calls) with relevant expertise in both the subject field(s) and in evaluating projects of at least similar size.

[5] https://opencalls.european-language-grid.eu

Before assigning projects to evaluators, we sent instructions via email and we organised webinars in which the evaluation process and criteria were explained.

All evaluators signed a contract with the ELG project. The contract included a clause to keep in strict confidence any technical or business information about the evaluated projects, as well as a no-conflict-of-interest declaration.

2.5.2 Execution of the Proposal Evaluation Process

Each proposal was evaluated by three independent external experts to ensure an transparent selection process. The evaluators were carefully assigned to the proposals by the management team. We also paid attention to gender (at least one female evaluator per proposal) and country of residence of the evaluator, avoiding at the same time possible personal or nationality-based conflicts of interest. The whole process was monitored by the Pilot Board. Each proposal was assigned to one of the PB members. These project coaches checked and confirmed or rejected the selection of evaluators with special regard to conflict of interest.

After the evaluation, the project coaches prepared summary reports for each proposal assigned to them. In these summaries, the coaches first reviewed the three reports by the external evaluators. They also suggested potential budget adjustments and changes of the total number of points (the maximum was 300 points, i. e., 100 points from each evaluator) in range of at most 30 points (open call 1) or 45 points (open call 2) up or down, where applicable. According to the evaluation criteria, project proposals by SMEs developing applications using LT available in ELG (B type projects) received 30 bonus points. Finally, the project coaches reviewed the eligibility criteria (uniqueness, relevance for ELG and project phases) as checked by the evaluators and suggested their decision on their fulfilment if the evaluators differed in opinion. The coaches also assessed the performance of the evaluators and quality of the reports. After all summary reports had been submitted by the coaches, a Pilot Board meeting was convened, in which the final ranking and selection was decided. All proposals were ranked by the total sum of points assigned. The ranked list was cut at the maximum available financial support (1,365,000€ for open call 1 and 585,000€ for open call 2).

3 Results

3.1 Open Call 1

3.1.1 Overview

The first call was opened on 1 March 2020 and closed on 30 April 2020 in accordance with the timeline (Figure 1). We accepted a total of 110 project proposals for evaluation from 103 applicants.

Submitted by	Type A	Type B	Total
Research organisation	43	5	48
SME	36	26	62
Total	79	31	110

Table 1 Proposals submitted to the first open call and accepted for evaluation

Seven applicants (five SMEs and two research organisations) submitted two proposals (one type A and one type B). Regarding the type of project, 79 submitted proposals were of type A (contribute resources, services, tools, or datasets to ELG) and 31 proposals were of type B (develop applications using language resources and technologies available in ELG), see Table 1. We received proposals from 29 different countries, including eligible countries outside the EU (Iceland, Israel, Norway, Serbia, South Africa, Switzerland, Turkey, United Kingdom). The total amount of financing requested by the submitted projects was 16,900,000€. One project requested 283,000€, which was over the limit of 200,000€ per project, and the lowest requested amount was 50,000€. The average amount requested per project was 153,000€.

At the end of June 2020, the results of the first open call were announced on the ELG website, including the list of projects selected for funding.[6] The two projects from the reserve list were informed that they might be selected for financial support if any of the selected projects rejected the financial support. The remaining projects were informed that they were not selected. In July 2020, contracts with all selected projects were signed, and the first payments were made (half of the awarded financial support), in line with the approved call documentation and procedures. All projects had started their execution phase by August 6. Furthermore, at the end of July 2020, abridged versions of the summary evaluation reports were provided to all applicants through the Open Calls Platform.

3.1.2 Selected Projects

The projects selected in open call 1 are listed in Table 2. All supported organisations are from the EU – three from Finland, two from Austria, Germany and Italy, and one from Spain. The awarded budget varies from 87,445€ to 167,375€.

Although we obtained more proposals from SMEs than from research organisations, there are three SMEs and seven research organisations among the selected projects. Similarly, although B type projects from SMEs were preferred, only two B type projects were accepted for financing which probably reflected the fact that the ELG platform was still being developed at the time of the first open call. Thus, it appeared to make more sense to create missing resources or tools rather than build applications using resources and tools available in ELG.

[6] https://www.european-language-grid.eu/open-calls/open-call-1

Organisation	Pilot Project	Type	Country	Funding
Fondazione Bruno Kessler	European Clinical Case Corpus	A	IT	139,370€
Lingsoft, Inc.	Lingsoft Solutions as Distributable Containers	A	FI	140,625€
Coreon GmbH	MKS as Linguistic Linked Open Data	A	DE	167,375€
Elhuyar Fundazioa	Basque-speaking smart speaker based on Mycroft AI	B	ES	117,117€
Universita' Degli Studi di Torino	Italian EVALITA Benchmark Linguistic Resources	A	IT	126,125€
University of Helsinki	Open Translation Models, Tools and Services	A	FI	154,636€
University of Vienna	Extracting Terminological Concept Systems from Text	A	AT	132,977€
University of Turku	Textual paraphrase dataset for deep language modelling	A	FI	166,085€
Weber Consulting KG	Virtual Personal Assistant Prototype	B	AT	87,445€
FZI Research Centre for Information Technology	Streaming Language Processing in Manufacturing	A	DE	132,160€

Table 2 List of pilot projects selected for financial support in the first open call

Four of the eight A type projects aimed to enrich the ELG platform with language resources and six of them planned to provide various language tools (i. e., two of the projects provide both resources and tools). The two B type projects promised speech applications – a smart speaker and a digital twin based on real-time language translation and analysis. The projects in general often dealt with underrepresented languages such as Basque, the Nordic languages, and European minority languages.

Technologically, the projects targeted a diverse set of goals and areas. There are projects targeting important interdisciplinary areas (medical informatics, manufacturing), modern technologies relating to language and semantic as well as world knowledge (Linked Open Data, paraphrasing) and core scalable technologies (distributable containers). Evaluation platforms as well as advanced and scalable machine translation still are and will be relevant issues for Language Technologies. Finally, the two speech-oriented applied projects broaden the portfolio of the usual Language Technologies in the desired direction, too.

3.1.3 Feedback provided and Survey for Proposers

With the goal of evaluating and improving our open call procedure, we conducted several surveys with everyone involved in the first open call. We started with the project proposers. After the evaluation process we also conducted a survey among all evaluators. The last survey was conducted among the Pilot Board members.

Two short surveys were designed for those who submitted a proposal (proposers) and those who uploaded an initial draft but did not submit a final version (non-proposers). The survey consisted of 15 questions, some open and some multiple choice. The survey topics were clustered into three sections: “motivation”, “project proposals”, and “your organisation”. The information was collected anonymously.

The surveys were conducted in May 2020. Of the proposers, 73 out of 110 (66%) responded, and of the non-proposers, 6 out of 17 (35%) responded. The main conclusions from the proposers’ survey that were relevant for the setup of the second open call: Almost 70% of respondents were interested in ELG because of both (functional) services and datasets. Slightly more than two thirds of the respondents preferred smaller, agile calls over large, consortium-based calls.

There was a demand for more detailed documentation (e. g., in the form of a webinar) that allows proposers to better interpret the strategic goals of ELG and get better information on already existing services in ELG. More details about the ELG API integration and about the infrastructure for working with data, applications and possibly also workflows were requested. Some improvements of the Open Calls Platform and its user-friendliness were made (e. g., limited space).

3.2 Open Call 2

The second open call was launched in October 2020 and experience from the first open call was reflected in its organisation.

3.2.1 Changes made between Open Call 1 and Open Call 2

The basic parameters, specified in the ELG Grant Agreement, remained the same for the second open call. Based on the lessons learned from open call 1, we implemented the following changes in the call documentation and the open call procedure:

- We improved the explanation of the strategic goals of ELG and the goals of the open calls. Links to an overview of ELG, its history and context and to an overview of the ELG platform were provided in the call documentation.
- We also improved the technical documentation of the ELG infrastructure and provided an easy-to-find list of currently available services – this was done with the launch of ELG Release 1 (June 2020).
- We organised a webinar, which took place during the submission period, on 12 November 2020. We explained the goals of the open call and presented the call documentation. The second part of the webinar was dedicated to questions and a discussion. A recording was made available to all applicants.
- The documentation, annexes, templates, and forms along with the Open Calls Platform were further improved.

- In the proposal template, budget breakdowns were requested in a fixed structure as well as a more detailed budget justification.
- New evaluators were recruited and added to the current group, with the aim to attract more experienced evaluators.
- It was decided that the second open call, like the first open call, should have no specific thematic focus.

3.2.2 Overview

The second call was opened on 1 October 2020 and closed on 30 November 2020 in accordance with the open calls timeline (Figure 1). We accepted 103 project proposals in total for evaluation.

Submitted by	Type A	Type B	Total
Research Organisation	38	5	43
SME	28	32	60
Total	66	37	103

Table 3 Proposals submitted to the second open call and accepted for evaluation

Five applicants (four SMEs and one research organisation) submitted two proposals (one type A and one type B). Regarding the project type, 66 proposals were of type A, and 37 project proposals were of type B. A total of 43 applicants who submitted a proposal in the second open call indicated that they had submitted the same or a similar proposal in the first open call. We received applications from 28 different countries, including eligible countries outside the EU (Iran, Israel, Norway, Serbia, Switzerland, Turkey, United Kingdom). The total amount of financing requested by the submitted projects was 13,257,919€. The average amount requested per project was 129,000€, which is less than in the first open call (153,000€).

In February 2021, the results of the second open call were announced on the ELG website.[7] All applicants were informed about the results. In February and March 2021, contracts with all selected projects were signed, and the first payments were made (half of the awarded financial support), in line with the call documentation and procedures. All projects had started their execution phase by 1 April 2021. Furthermore, in March 2021, abridged versions of the summary evaluation reports were made available to all applicants through the Open Calls Platform.

[7] https://www.european-language-grid.eu/open-calls/open-call-2

3.2.3 Selected Projects

The projects selected for financial support in open call 2 are listed in Table 4. The supported organisations are from five EU countries and the awarded budget varies between 85,421€ and 137,227€.

Organisation	Pilot Project Name	Type	Country	Funding
Institute for Bulgarian Language	Multilingual Image Corpus 2021	A	BG	110,960€
EDIA BV	CEFR Labelling and Assessment Services	B	NL	137,560€
University of West Bohemia	Motion-Capture 3D Sign Language Resources	A	CZ	85,421€
Sapienza University of Rome	Universal Semantic Annotator: A Unified API for Multilingual WSD, SRL and AMR	A	IT	113,228€
Sign Time GmbH	Sign language explanations for terms in a text	B	AT	137,227€

Table 4 List of pilot projects selected for financial support in the second open call

Although we obtained more project proposals from SMEs than from research organisations, there are two SMEs and three research organisations among the selected projects. Similarly, only two B type projects were accepted for financing.

Three A type projects aimed at providing tools to enrich the ELG platform. One project contributed multilingual annotated data, tools and services for image processing whilst the second one aimed at improving the ELG offer of linguistic tools by proposing a unified service powered by state-of-the-art neural models for carrying out annotations on three Natural Language Understanding tasks, i. e., Word Sense Disambiguation, Semantic Role Labelling and Semantic Parsing, in around 100 languages. The third A type project expanded the portfolio of language resources available in ELG by adding a dataset and search tool for Czech sign language. Regarding the B type projects, one of the projects also dealt with sign language. Its goal was to simplify text comprehension for deaf people by linking words and phrases to a sign language encyclopedia. The other project aimed to develop a set of tools, datasets, and services to enable automatic classification of the reading difficulty of texts on the Common European Framework of Reference.

3.2.4 Survey for Proposers to the Open Call 2

Just like for the first open call, a survey with 15 questions was designed for those who submitted a proposal. The survey had three sections: “motivation”, “project proposals”, “your organisation”. In total, 39 out of 103 proposers (38%) responded. Regarding the motivation to submit a proposal, contributing services or resources to ELG to make them available to the ELG community and further development of an existing software or data project were the most frequent reasons reported by the respondents. The main expectations toward ELG were that the platform increases the visibility of the applicant's organisation on the European level and to get access to a large repository of tools and datasets. Also, almost all respondents think that more EU-funded activities dedicated to Language Technology and Language-centric AI are needed, preferably in the form of agile calls (with short proposals and quick evaluations, 9-12 months project run-time). Regarding the specialisation of respondents, most frequently they specialised in text analytics, machine translation or speech recognition. Respondents reported more than twenty domains that they specialise in (most frequently health sector), one fourth of all respondents have no particular specialisation.

4 Pilot Project Execution

Once the pilot projects were selected and the contracts signed, the continuous support from the ELG consortium started so that the projects could start their execution. The first opportunity where the newly selected pilot projects could become more familiar with ELG were the online meetings with the Pilot Board and other members of the ELG consortium. During these meetings, basic information about ELG and its technology as well as guidelines for project execution were presented.

Project execution (Figure 2) consisted of three phases: Phase 1 – Experiment; Phase 2 – Integration; Phase 3 – Dissemination. After finishing Phase 1, reporting from the applicants was required, and then the Pilot Board decided whether the project was allowed to continue execution (and consequently, whether the next payment, 35% of the awarded support, is made). After finishing Phase 3, a final report was required, and the Pilot Board evaluated the whole project and decided whether the project receives the final payment (15% of the awarded financial support).

As mentioned, each project was supervised by a project coach who was responsible for training the project team, collecting and answering questions during project execution, collecting reports, and guiding the team through the project phases.

To advertise them to a wider public, the pilot projects were presented at two annual ELG conferences, i. e., META-FORUM 2020 and META-FORUM 2021, in dedicated pilot project sessions in which all projects could present their main approaches and goals. In addition, workshops and training events organised by the ELG National Competence Centres (NCCs) were also used as opportunities to present certain pilot projects in the respective countries and regions.

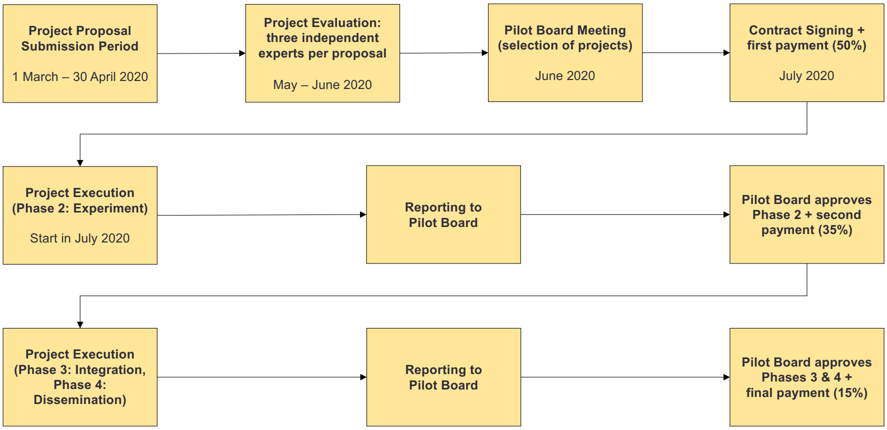

Project Proposal Submission Period
1 March – 30 April 2020
Project Evaluation: three independent experts per proposal
May – June 2020
Pilot Board Meeting (selection of projects)
June 2020
Contract Signing + first payment (50%)
July 2020
Project Execution (Phase 2: Experiment)
Start in July 2020
Reporting to Pilot Board
Pilot Board approves Phase 2 + second payment (35%)
Project Execution (Phase 3: Integration, Phase 4: Dissemination)
Reporting to Pilot Board
Pilot Board approves Phases 3 & 4 + final payment (15%)

Fig. 2 Project execution scheme for pilot projects from the first open call

5 Conclusions

The results of the two open calls demonstrate an enormous interest in the European Language Grid and the Language Technology topic in general. The interest also indicates that the setup, including documentation, proposal template, platform etc., was easy to follow. In total, we received 213 project proposals from 156 different institutions (86 SMEs, 70 research organisations) in 32 different countries (including nine eligible countries outside the European Union); 15 projects were selected for funding, ten in the first open call and five in the second. The total amount requested was approx. 30 mil. €, while the available funding amounted to only 1.95 mil. € (an oversubscription of more than 15 times).

In the following we briefly summarise the main lessons learned, as gathered through the different surveys (see Sections 3.1.3 and 3.2.4):

- We aimed at a simple and light-weight procedure which led to a high number of submitted proposals. At the same time, the simplicity of the proposal template may have led to a higher number of low-quality proposals that were not adequately described or thought through. In both calls this rather high number of proposals required more person days and increased the costs related to the external evaluators.
- The quality of evaluation reports submitted by external evaluators was not entirely stable and, in some cases, could have been more profound. This was usually balanced by the project coach or Pilot Board.
- It was a good decision to develop the Open Calls Platform internally. Among others, it provided us with more flexibility, control over deadlines and quick and reliable support from the technical team.
- In the ELG project budget, the costs for the Open Calls Platform and for the proposal evaluation should have been planned more carefully.

Overall, we were very satisfied with the open calls setup and with the results of the pilot projects. While the results improved the ELG offering in terms of data, tools and services, and the applications developed using the ELG provided mutual benefit to the developers and ELG, we consider the overwhelming interest in the open calls an extremely important, albeit non-technical result: it demonstrates that Language Technologies are of tremendous interest to both researchers and commercial companies. It also shows that the open calls setup, as designed and implemented, was very attractive and can be considered as a model in similar undertakings in the future.

References

European Commission (2017). *Horizon 2020 – Work Programme 2018-2020. Annex K: Actions involving financial support to third parties*. Extract from Part 19 – Commission Decision C(2017)7124. Brussels, Belgium. URL: https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2018-2020/annexes/h2020-wp1820-annex-k-fs3p_en.pdf.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Chapter 15
Basque-speaking Smart Speaker based on Mycroft AI

Igor Leturia, Ander Corral, Xabier Sarasola, Beñat Jimenez, Silvia Portela, Arkaitz Anza, and Jaione Martinez

Abstract Speech-driven virtual assistants, known as smart speakers, such as Amazon Echo and Google Home, are increasingly used. However, commercial smart speakers only support a handful of languages. Even languages for which ASR and TTS technology is available, such as many official EU member state languages, are not supported due to a commercial disinterest derived from their – relatively speaking – rather small number of speakers. This problem is even more crucial for minority languages, for which smart speakers are not expected anytime soon, or ever. In this ELG pilot project we developed a Basque-speaking smart speaker, making use of the open source smart speaker project Mycroft AI and Elhuyar Foundation's speech technologies for Basque. Apart from getting it to speak Basque, one of our goals was to make the smart speaker privacy friendly, non-gendered and use local services, because these are usual issues of concern. The project has also served to improve the state of the art of Basque ASR and TTS technology.

1 Overview and Objectives of the Pilot Project

Commercial smart speakers are increasingly popular despite the fact that their language coverage leaves much to be desired. Many large official national languages and practically all minority languages are unsupported by these devices. In many cases, the lack of support for a language in a smart speaker is not due to the lack of the necessary speech technologies, i. e., Automatic Speech Recognition (ASR) and Text To Speech (TTS). ASR and TTS technologies do exist for the Basque language

Igor Leturia · Ander Corral · Xabier Sarasola
Elhuyar Fundazioa, Spain, i.leturia@elhuyar.eus, a.corral@elhuyar.eus, x.sarasola@elhuyar.eus

Beñat Jimenez
Talaios Koop., Spain, jimakker@talaios.coop

Silvia Portela · Arkaitz Anza · Jaione Martinez
Skura Mobile, Spain, silvia@skuramobile.com, arkaitz@skuramobile.com, jaione@skuramobile.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_15

but it is unlikely that they will be implemented in smart speakers developed by the big technology enterprises because of its relatively small number of speakers.

On the other hand, there is a rather mature, open source smart speaker project called Mycroft AI.[1] Our ELG pilot project develops an open source smart speaker for the Basque language, based on Mycroft AI, that makes use of Elhuyar Foundation's ASR and TTS technologies. Apart from being open source and in Basque, other points of interest were the handling of privacy, gender and service locality issues.

One objective of the project was to improve the state of the art of Basque ASR and TTS technologies, since it would be necessary to adapt them to the context of a smart speaker. Specifically, we wanted to 1. improve the performance of Basque ASR technology for noisy environments; 2. create a grammar-based ASR system instead of a general vocabulary one to only recognise the commands of the speaker and, thus, improve precision; 3. create a neural network-based TTS system for Basque and replace the old HMM one; and 4. try to develop a gender-neutral voice.

2 Mycroft Localisation

A crucial and necessary part of the project was the localisation of Mycroft to Basque in its broadest sense. This involved not only a string translation process, but also making it understand speech commands and respond via speech in Basque. Thus, we had to develop plugins to connect Mycroft to Elhuyar's ASR and TTS services.

The localisation also involved the adaptation to Basque of Mycroft's linguistic module called *lingua-franca*, responsible for parsing numbers, days, times, durations, etc. in speech commands and to pronounce them correctly when responding.

Finally, the routine job of string translation of any software localisation process turned out not to be as straightforward for the commands' part. The parsing of many skills' intents from the commands is done by simply detecting some required or optional keywords and parameters, which is why their translation required more than just a simple sentence translation. We translated the Mycroft core module and 40+ of its skills (volume control, date, time, lists, alarms, audio record, radio, news, Wikipedia, weather, jokes, Wikiquote, e-mail etc.).

3 Privacy, Gender and Proximity

As mentioned in Section 1, we wanted to address the privacy and gender concerns often associated with smart speakers and also promote the use of local services. Regarding privacy, users and potential buyers have concerns with having a device in their homes with a microphone that is always on (Lau et al. 2018). However, respect for privacy is precisely one of Mycroft AI's unique selling propositions. They claim

[1] https://mycroft.ai

that they are "private by default" and that they "promise to never sell your data or give you advertisements" using their technology. This materialises in the fact that the wake word ("Hey, Mycroft") is detected locally, i. e., no audio is sent to remote servers except when saying a command after the detection of the wake word. On the other hand, if some big enterprise's cloud-based ASR or TTS services are used for the recognition of commands and the utterance of responses, there are logically some doubts as to what these companies will do with that data. Using Elhuyar's Basque ASR and TTS remote APIs from Mycroft, no data would be kept or collected.

Regarding gender treatment, smart speakers are known for their improper gender treatment, as stated in the Unesco report "I'd blush if I could: closing gender divides in digital skills through education" (West et al. 2019). According to this report, practically all commercial smart speakers exhibit a female voice and female personalities, and respond obligingly even to hostile requests, verbal abuse and sexual harassment, which may lead to reinforce and spread gender biases. The report ends with some recommendations that range from not making digital assistants female by default to developing neutral voices and personalities, which our project has tried to follow. The Basque voice installed at the moment is a male voice by default. Also, the speaker's name, Mycroft, – although fictional – is male, its "personality" is neutral, and it has no skill to respond in a docile manner to sexual comments or verbal abuse. However, we have also carried out some experiments in order to develop a gender-neutral synthetic voice (see Section 4.4).

We felt that our smart speaker should prioritise the local region and, for instance, allow listening to local radio stations, read the news from local media or buy goods or order food from local stores. We developed half a dozen local skills of our own, including local news, local radio stations, dictionary querying or Basque music.

4 Developments in Basque Speech Technology

4.1 ASR Robustness in Noisy Environments

One of the main challenges regarding the use of ASR technology in a smart speaker is making it robust enough to be reliable under non-optimal conditions: low volume, background noise, music, speech, room reverberation, low quality microphone, etc.

Elhuyar's ASR system for the Basque language is a general purpose system based on the Kaldi[2] toolkit. The speech data used to train the acoustic model comprises high quality clean parliamentary speeches. To make our acoustic model more robust, we used several synthetic data augmentation techniques during the training phase (Alumäe et al. 2018). This means that training data was 1. synthetically augmented by adding background noises from the MUSAN dataset (Snyder et al. 2015), which comprises several recordings of music, speech and a wide variety of noises;

[2] https://kaldi-asr.org

2. artificially reverberated with various real and simulated room impulse responses (Ko et al. 2017); and 3. augmented with threefold speed and volume perturbations.

4.2 ASR Closed Grammar-based Recognition

For general purpose ASR systems, typically a large language model is trained with a vast amount of diverse texts. For a smart speaker, however, where the user is expected to use a closed set of commands, limiting the ASR's vocabulary to just the necessary commands can increase the precision of the speech recognition.

Since Kaldi internally uses weighted finite state transducers (WFST) to model the language, simply by converting all the commands defined in Mycroft skills to the format used by Pynini (a Python library for WFST grammar compilation), we would obtain a language model limited to Mycroft's commands. But although Mycroft's skills were originally defined using its old-style intent parser Padatious (where the whole command is defined), nowadays most skills use the new intent parser Adapt, which defines commands using a few keywords and parameters. This makes it unfeasible to automatically generate all possible commands containing the keywords and parameters. Rewriting all skills to the Padatious format would have made the code much more difficult to maintain as well as losing Adapt's recall gain. This is why the creation of a custom grammar was eventually discarded.

4.3 Neural Network-based Basque TTS

Elhuyar's previous Basque TTS service was based on Hidden Markov Models (HMMs). In the ELG pilot project we developed a new neural network-based TTS service. Since the first neural system was published in 2013 (Zen et al. 2013), these have taken a clear advantage over HMM-based approaches and systems like Tacotron 2 (Shen et al. 2018) have achieved naturalness comparable to natural voice.

The key challenge with neural TTS systems is the size of the training dataset. The original Tacotron 2 monospeaker system was trained with 24.6 hours of speech, and subsequent research concluded that 10 hours is the minimum time required to obtain maximum quality (Chung et al. 2019). The only publicly available database of Basque speech of that size is a multispeaker database created by Google (Kjartansson et al. 2020), which contains recordings from 53 speakers with a maximum of 15 minutes per speaker. Modified configurations of Tacotron 2 using speaker embeddings have proved successful providing good quality multispeaker TTS systems (Jia et al. 2018), i. e., systems trained using combined recordings of multiple speakers, capable to synthesise the voice of each of them. We recorded a small multispeaker database, combined it with the Google database, and trained a multispeaker TTS using speaker embeddings, obtaining our own neural quality TTS voices.

4.4 Gender-neutral Voice

Apart from the interventions to address gender issues (Section 3), we conducted experiments towards obtaining a gender-neutral voice. Tolmeijer et al. (2021) observed that we do not regard voices of intermediate pitch (which is what could be understood as gender-neutral) as genderless, that we assign them one gender or the other, and that those that could be best considered as ambiguous in terms of gender or genderless were those with the greatest division of opinion.

Most of the literature on the field of generating gender-ambiguous voices seek gender neutrality through pitch modification, such as Tolmeijer et al. (2021), or the first genderless voice Q (Carpenter 2019). We employed a different and innovative approach. We first calculate the average speaker embedding for each gender with the embeddings obtained in the training and then we compute the embedding that is midway between the average male and female embeddings. Using this embedding in the trained Tacotron 2, we can synthesise sentences with a voice which has produced divided opinions as to its gender and which can thus be considered genderless.

5 Conclusions and Results of the Pilot Project

This ELG pilot project developed an open source Basque-speaking smart speaker based on Mycroft AI, which respects privacy and which uses a more appropriate approach regarding the voice's gender than commercial smart speakers. We connected Mycroft to Elhuyar's Basque ASR and TTS services, and we improved the state of the art of Basque speech technologies. Our ASR for Basque performs better in noisy environments and we developed a new deep neural network-based TTS for Basque and made experiments towards a gender-ambiguous synthetic voice. We translated more than 40 Mycroft skills and developed half a dozen new ones addressing local services. We tested the Basque Mycroft in PCs and Google AIY Kits.

Anyone can now download, install on a device and try Mycroft in Basque. While the ELG pilot project is finished, we continue to work on the project with the aim of, if possible, bringing a Basque smart speaker device to the market. We believe that the work carried out, the experience gained and the code developed in the ELG pilot project can be very useful for other minority language communities that would like to have access to a smart speaker that speaks their own language.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. The project has also been helped by the Basque Government through its Hazitek programme (project DomEus) and the Gipuzkoa Provincial Council through its Etorkizuna eraikiz programme (project Mycroft.eus).

References

Alumäe, Tanel, Ottokar Tilk, and Asad Ullah (2018). “Advanced rich transcription system for Estonian speech”. In: *Human Language Technologies – the Baltic Perspective: Proc. of the Eighth Int. Conference (Baltic HLT 2018)*. Ed. by Kadri Muischnek and Kaili Müürisep. Amsterdam, the Netherlands: IOS Press, pp. 1–8. DOI: 10.3233/978-1-61499-912-6-1.

Carpenter, Julie (2019). “Why Project Q is More than the World’s First Nonbinary Voice for Technology”. In: *Interactions* 26.6, pp. 56–59. DOI: 10.1145/3358912.

Chung, Yu-An, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan (2019). “Semi-supervised training for improving data efficiency in end-to-end speech synthesis”. In: *ICASSP 2019*. IEEE, pp. 6940–6944.

Jia, Ye, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. (2018). “Transfer learning from speaker verification to multispeaker text-to-speech synthesis”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, pp. 4485–4495.

Kjartansson, Oddur, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara E. Rivera (2020). “Open-Source High Quality Speech Datasets for Basque, Catalan and Galician”. In: *SLTU-CCURL 2020*. 11–12 May, Marseille, France, pp. 21–27.

Ko, Tom, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur (2017). “A study on data augmentation of reverberant speech for robust speech recognition”. In: *ICASSP 2017*, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.

Lau, Josephine, Benjamin Zimmerman, and Florian Schaub (2018). “Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers”. In: *Proc. of Human-Computer Interaction* 2.CSCW, pp. 1–31. DOI: 10.1145/3274371.

Shen, Jonathan, Ruoming Pang, Ron Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. (2018). “Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions”. In: *ICASSP 2018*. IEEE, pp. 4779–4783.

Snyder, David, Guoguo Chen, and Daniel Povey (2015). “Musan: A music, speech, and noise corpus”. In: *arXiv preprint arXiv:1510.08484*.

Tolmeijer, Suzanne, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein (2021). “Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution”. In: *Conference on Human Factors in Computing Systems (CHI)*. New York, NY, USA: ACM, pp. 1–7.

West, Mark, Rebecca Kraut, and Han Ei Chew (2019). *I’d blush if I could: closing gender divides in digital skills through education*. Unesco EQUALS.

Zen, Heiga, Andrew Senior, and Mike Schuster (2013). “Statistical parametric speech synthesis using deep neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7962–7966.

Chapter 16
CEFR Labelling and Assessment Services

Mark Breuker

Abstract Our pilot project aims to develop a set of text collections and annotation tools to facilitate the creation of datasets (corpora) for the development of AI classification models. These classification models can automatically assess a text's reading difficulty on the levels described by the Common European Framework of Reference (CEFR). The ability to accurately and consistently assess the readability level of texts is crucial to authors and (language) teachers. It allows them to more easily create and discover content that meets the needs of students with different backgrounds and skill levels. Also, in the public sector using plain language in written communication is becoming increasingly important to ensure citizens can easily access and comprehend government information. EDIA already provides automated readability assessment services (available as APIs and an online authoring tool) for the CEFR in English. Support for Dutch, German and Spanish are added as part of this project. Using the infrastructure developed in this project the effort for creating high quality datasets for additional languages is lowered significantly. The tools and datasets are deployed through the European Language Grid. The project is scheduled to be completed in the second quarter of 2022.

1 Overview and Objectives of the Pilot Project

The CEFR (Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Council of Europe 2020) aims to provide a comprehensive learning, teaching and assessment method that can be used for all European languages. Indicating the level of learners of foreign languages, the CEFR facilitates the assessment of a person's language proficiency. By now, most are familiar with the six reference levels (A1 – C2) used for this purpose (Figure 1).

CEFR levels are the foundation for a communicative approach to (foreign) language acquisition, teaching and certification. Although the CEFR levels represent a widely supported approach, the availability and quality of (educational) content la-

Mark Breuker
EDIA b. v., The Netherlands, mark@edia.nl

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_16

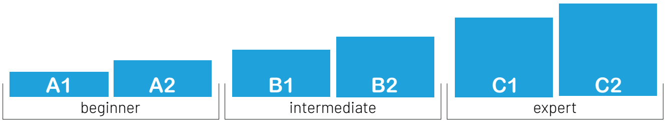


Fig. 1 CEFR proficiency levels

belled with CEFR levels are limited. This is because the highly laborious, error-prone labelling process is performed manually (save for some exceptions). This results in several practical obstacles regarding publishing, teaching, and learning:

- Content creators (publishers, authors, teachers, government officials) struggle to use consistent criteria for checking a text's difficulty level.
- Teachers have trouble finding or creating appropriate texts for their students.
- Content managers struggle to monitor the readability level of their content collections over time.

To tackle this problem, we have developed an automated text classification technology using Natural Language Processing. Our technology can perform CEFR text levelling in a scalable and consistent manner for multiple languages at a very granular level. By removing blockers through automation, we expect to impact the practical application of CEFR, enabling the labelling of more content in less time in a highly consistent manner. This way, we will lay the foundation for making written content with properly labelled text levels more widely available, adhering to the CEFR standard. After all, practical obstacles will have been eliminated.

The European Language Grid (Rehm et al. 2021) provides EDIA with a marketplace to promote, sell and distribute its CEFR services to a broad audience. Through the standardised ELG catalogue and API specification, developers can more easily adopt the services provided by EDIA in their applications.

2 Methodology

The infrastructure for the CEFR readability services developed during the pilot project consists of various components (Figure 2). The infrastructure facilitates the creation of the CEFR readability assessment services, using the following process:

1. Data collection – collect (unlabelled) texts for each language
2. Data labelling – label the texts on CEFR reading level using human experts
3. Model training – train classification models on the datasets
4. Integration – expose the models as REST services on ELG using API proxies
5. Authoring – integrate the services in a CEFR levelling and authoring application

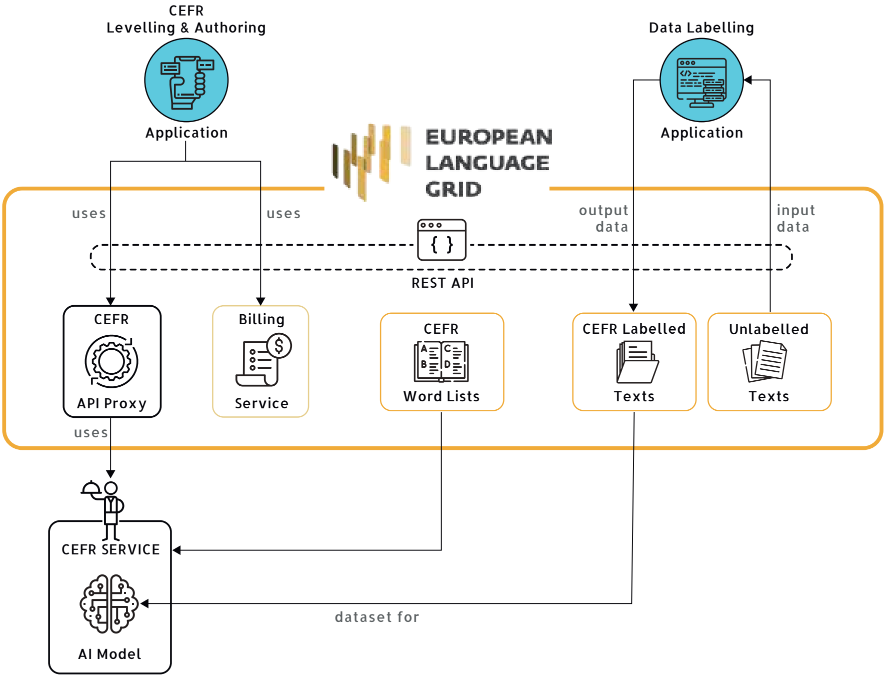


Fig. 2 CEFR infrastructure diagram

3 Implementation

To create the corpus, we collected approx. 1,200 texts per language from various public sources such as newspapers, magazines, educational sources and government websites. To speed up the text collection process we developed several text-scraping algorithms. Each text was stored as plain-text in a database together with information about its source and copyright licence. To ensure that the unlabelled dataset was well balanced and covered both easy and more difficult reading levels, we used texts from sources known to be targeted at basic, intermediate and advanced language users. In addition we used heuristics-based methods of readability assessment. This provided us with an initial indication of the reading difficulty of each text.

Our first attempt at a data labelling application was based on a pairwise comparison algorithm (Crompvoets et al. 2020). We applied this approach on a collection of 1,200 Dutch texts. The rationale for this approach was that comparing two texts on reading difficulty is a relatively easy task for teachers and would suffer less from inconsistent and subjective criteria used when evaluating a text directly on its CEFR-level. This approach resulted in a rank-ordered list of texts on reading difficulty. Next we set the boundaries for the CEFR-reading levels within this rank-ordered list. Unfortunately we found that we were not able to train a classification model on the dataset. Upon closer inspection (based on a random sample of 100 texts) we found that many texts were labelled incorrectly (i. e., 25 percent more than two levels off). Although we compared each text with six other texts (resulting in a total of 7,200 annotations), possibly the number of comparisons per text was still insufficient to create a reliable measurement. This means the pairwise comparison approach also offers no benefits compared to labelling each text on CEFR level by three experts (resulting in a total of 3,600 annotations) with regard to the number of annotations needed to a reliable dataset.

In our second attempt we labelled texts directly regarding their CEFR reading level. This new labelling application provides functionality for organising the unlabelled texts into various projects which supports working with multiple languages and creating subsets from the total corpus to label the texts in smaller batches. This allows us to annotate the texts iteratively which means we can better monitor the quality of the annotations during the labelling process. Within each project, annotation tasks are created and are assigned to language experts. Each text is evaluated by three different experts to ensure high quality CEFR assessments. For each text annotators complete an assessment form with criteria described in the CEFR reading level descriptors (such as vocabulary and grammatical complexity, Alderson et al. 2006). We have based this approach in part on the CEFR Estim Grid project (Tardieu et al. 2010). Prior to completing the content labelling tasks, annotators participate in an (online) workshop to collaboratively assess the CEFR level of a small subset of texts to align on the CEFR level descriptors.

Once we labelled all texts and completed the datasets we were able to develop the CEFR readability classification models. The models we created return the predicted difficulty on a linear scale, which means that we can predict the reading difficulty more granularly than the 6-level CEFR scale. In other words, we can say, for example, that a text is on the more difficult end of the B2 level. Based on the models, we created web services for assessing the overall readability of a text, difficult words in the text and alternative words (suggestions) for these difficult words.

We then integrated our CEFR services into the ELG platform using proxy services. A proxy service maps incoming ELG requests onto our classification API running on our web servers. The proxy service was packaged as a Docker container, stored in our company's Docker registry and then deployed on ELG. To improve performance and avoid blocking requests, we used the Asyncio library to support asynchronous processing of service requests. To speed up the development of the proxy services, we switched to using ELG's Python SDK for later versions of our service implementations.

For the authoring application we chose to integrate our CEFR services with the Fonto editor[1] as an add-on. This allowed us to focus on developing the text analysis rather than basic text editing features. In addition we used the Fonto Content Quality component to highlight relevant sections in the text and provide feedback to authors which allows them to improve the readability and quality of their texts. The Fonto editor is a popular tool by major (educational) publishers, which enables easy integration and adoption of our technology by new clients.

[1] https://www.fontoxml.com

4 Evaluation

For collecting the texts for our dataset we had planned to use the C4 Corpus (public domain part)[2] which is a huge collection of plain texts, released under a Creative Commons licence, which appeared to be very useful for our project. However, upon closer inspection we found that the licence detection algorithm that was used is not very accurate and that the structure of the texts was not very suitable for our purposes. Also, the sheer size of this corpus added to the complexity of its processing. We therefore decided not to use the C4 Corpus, but create a new corpus instead. We tried various methods for data labelling. Unfortunately the pairwise comparison did not yield a useful dataset from which we could create a classification model. Possible explanations may be that the number of comparisons per text was too low, that we did not select the right pairs of texts for the language teachers to compare, or that the teachers did not consistently select the most difficult text from each set. This would need to be investigated further.

Integrating our services into the ELG was straightforward and easy. Using the ELG Python SDK we were able to make our services available through ELG. We also appreciated the thorough review process of our submitted services and datasets by the ELG team. We received good feedback and support to improve the required metadata, code performance and overall compatibility with the ELG API specification. The standards-based ELG integration (e. g., using the ELG Python SDK) makes it significantly easier for third-party developers to consume and integrate our services in their language learning applications. We have not yet been able to evaluate the billing services of the ELG in a production setting. We can see that the services we deployed on ELG have been used multiple times, but we have little information about the use over time and the types of users (e. g., commercial vs. academic).

5 Conclusions and Results of the Pilot Project

Our goals with this project were to extend our CEFR service to additional languages beyond English and to use the European Language Grid as a marketplace for commercialising our services. Although the project has not yet been completed we can already see that the project has helped us to improve our data collection and labelling process, which helps to create high quality datasets for training additional language models. We created CEFR readability classification models using these datasets which we have made available on ELG as services.[3] The services are integrated into a text authoring application which helps authors assess and improve the readability of their (educational) texts in multiple languages. Deploying services on the ELG is currently easy and useful for demonstration and trial purposes. We

[2] https://live.european-language-grid.eu/catalogue/#/resource/service/corpus/1186

[3] https://live.european-language-grid.eu/catalogue/project/5258

believe the ELG SDKs enable third party developers to more easily discover and consume our APIs.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. We thank Cito Labs for their valuable expertise and support in the pairwise labelling experiment; all language teachers involved in labelling hundreds of texts on the CEFR; FontoXML for their support with integrating our NLP services with their product; the ELG team for funding our pilot project and the help we received for deploying our CEFR services and datasets on the ELG platform.

References

Alderson, J. Charles, Neus Figueras, Henk Kuijper, Guenter Nold, Sauli Takala, and Claire Tardieu (2006). “Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project”. In: *Language Assessment Quarterly* 3.1, pp. 3–30. URL: https://doi.org/10.1207/s15434311laq0301_2.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Strasbourg: Council of Europe Publishing, pp. 53–59. URL: https://www.coe.int/lang-cefr.

Crompvoets, Elise A. V., Anton A. Béguin, and Klaas Sijtsma (2020). “Adaptive Pairwise Comparison for Educational Measurement”. In: *Journal of Educational and Behavioral Statistics* 45.3, pp. 316–338. DOI: 10.3102/1076998619890589. URL: https://doi.org/10.3102/1076998619890589.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). “European Language Grid: A Joint Platform for the European Language Technology Community”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Tardieu, Claire, Raili Hildén, Magda Lehmann, and Monique Reichert (2010). *The CEF-ESTIM Grid*. URL: http://cefestim.ecml.at.

Chapter 17
European Clinical Case Corpus

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli

Abstract Interpreting information in medical documents has become one of the most relevant application areas for language technologies. However, despite the fact that huge amounts of medical documents (e. g., medical examination reports, hospital discharge letters, digital medical records) are produced, their availability for research purposes is still limited, due to strict data protection regulations. Aiming at fostering advanced information extraction technologies for medical applications, we present E3C, a corpus of clinical case narratives fully based on freely licensed documents. E3C (European Clinical Case Corpus) contains a vast selection of clinical cases (i. e., narratives presenting a patient's history) that cover different medical areas, are based on different styles and produced in different languages. A portion of the corpus has been manually annotated to be used for training and testing purposes, while a larger set of documents has been automatically tagged to serve as a baseline for future research in information extraction.

1 Overview and Objectives of the Pilot Project

The interest in information extraction from clinical narratives has increased in recent decades, including clinical entity extraction and classification (Schulz et al. 2020; Grabar et al. 2019; Dreisbach et al. 2019; Luo et al. 2017), clinical prediction systems, e. g., MIMIC III (Johnson et al. 2016), and the organisation of challenges at CLEF (Kelly et al. 2019), and Semeval. However, only a few shared datasets have been created, limiting the potential of developing applications in this area.

Bernardo Magnini · Alberto Lavelli · Manuela Speranza · Roberto Zanoli
Fondazione Bruno Kessler, Italy, magnini@fbk.eu, lavelli@fbk.eu, manspera@fbk.eu, zanoli@fbk.eu

Begoña Altuna
Fondazione Bruno Kessler, Italy, HiTZ Centre, University of the Basque Country, Spain, begona.altuna@ehu.eus

Anne-Lyse Minard
Université d'Orléans, France, anne-lyse.minard@univ-orleans.fr

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_17

We report upon the E3C (European Clinical Case Corpus) ELG pilot project, which resulted in a large collection of clinical cases in five European languages: English, Spanish, French, Italian and Basque. A clinical case is a statement of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. Clinical cases are typically reported and discussed in research papers, and are often used for education purposes in medicine. In addition, published clinical cases are de-identified, overcoming privacy issues, and are rich in clinical entities as well as temporal information.

> A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk. Continuous superficial ulcers and spontaneous bleeding were observed under colonoscopy. Subsequent gastroscopy revealed mucosa with diffuse edema, ulcers, errhysis, and granular and friable changes in the stomach and duodenal bulb, which were similar to the appearance of the rectum. After ruling out other possibilities according to a series of examinations, a diagnosis of GDUC was considered. The patient hesitated about intravenous corticosteroids, so he received a standardized treatment with pentasa of 3.2 g/d. After 0.5 mo of treatment, the patient's symptoms achieved complete remission. Follow-up endoscopy and imaging findings showed no evidence of recurrence for 26 mo.

The sample clinical case reported in the box above is about a patient presenting gastric symptoms, who is finally diagnosed with gastroduodenitis associated with ulcerative colitis (GDUC). To reach the diagnosis, two medical tests (colonoscopy and gastroscopy) were performed. Treatment, outcome (complete remission) and follow-up (no evidence of recurrence) are also present in the text.

2 Corpus Collection and Annotation

The document collection was determined by the available resources for each language (e. g., PubMed, scientific journals, medicine leaflets). First, we identified possible document sources as well as their licenses and re-distribution policies. We selected sources that were either already available under Creative Commons licenses (i. e., CC-BY or CC-BY-SA), possibly asking for re-distribution permission to the right holders. In the case of the SPACCC[1] and NUBes[2] corpora, the texts were ready to be used by us in terms of licensing and formatting. We automated the text collection as much as possible, for example, in some cases we were able to identify and extract the section with the clinical case. All English and some French documents were automatically extracted from PubMed[3], through its API, while medicine leaflets were automatically crawled and stored in a single file for each language. Journal articles with clinical cases that could not be extracted automatically were filtered through the search query "clinical case" in the different languages. In addition to the

[1] https://github.com/PlanTL-GOB-ES/SPACCC

[2] https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus

[3] https://pubmed.ncbi.nlm.nih.gov

extraction of the relevant documents, corresponding metadata was stored to allow accurate documentation.

The annotation of temporal information was performed following an adaptation of the THYME annotation guidelines (Styler et al. 2014).[4] Temporal information refers to the events in a text as well as to chronological references and relations. To encode temporal information, we defined the following tags and relation types. Events, time expressions, temporal relations and aspectual relations are widely used in temporal information tasks, while actor, body part and RML annotations were added as they convey relevant information of the clinical domain.

- *Events* are the events or states relevant to the patient's clinical timeline.
- *Time expressions* refer to points and intervals in time.
- *Temporal relations* (TLINK) implement relations that chronologically order events and time expressions.
- *Aspectual relations* (ALINK) are created between an aspectual event and its subordinated non-aspectual event.
- *Actors* are the people (or animals) mentioned in the text.
- *Body parts* are the parts of the body that are bigger than cells.
- *Results, measurements and lab and test results (RML)* are lab test and analytics' results, formulaic measurements and measurement values.

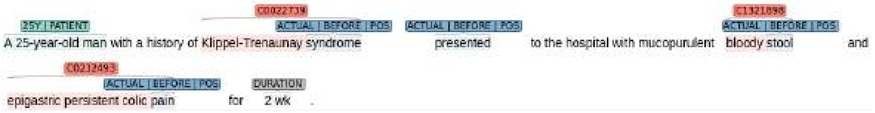

A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk .

Fig. 1 A sentence in a clinical case annotated with both temporal information and clinical entities (i. e., disorders) with their UMLS codes (marked in red)

The annotation of clinical entities is mainly based on the guidelines of SEMEVAL 2015 Task 14 "Analysis of Clinical Text"[5] and on the ASSESS CT guidelines (Miñarro-Giménez et al. 2018). The annotation of Layer 1 was done fully manually, while for Layer 2 the automatic annotation was produced with a distant supervision method that matches clinical entities with disorder concepts in UMLS.

3 Implementation

The E3C corpus is organised in three different layers:

Layer 1: about 25k tokens per language of clinical narratives with full manual or manually checked annotation of clinical entities, temporal information and factuality, for benchmarking and linguistic analysis.

[4] http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

[5] http://alt.qcri.org/semeval2015/task14/data/uploads/share_annotation_guidelines.pdf

Layer 2: 50-100k tokens per language of clinical narratives with automatic annotation of clinical entities. Distant supervision was used to annotate 8,972 clinical entities with their corresponding concepts in UMLS.

Layer 3: about 1m tokens per language of non-annotated medical documents (not necessarily clinical narratives) to be exploited by semi-supervised approaches.

Table 1 shows the sizes of the layers (document and token numbers). Table 2 shows the numbers of Layer 1 tags to indicate information density in clinical cases.

	English	**French**	**Italian**	**Spanish**	**Basque**
Layer 1	84 / 25142	81 /25196	86 / 24319	81 / 24681	90 / 22505
Layer 2	171 / 50371	168 / 50490	174 / 49900	162 / 49351	111 / 12541
Layer 3	9779 / 1075709	25740 / 66281501	10213 / 13601915	1876 / 1030907	1232 / 518244

Table 1 Documents/tokens in each language and layer in the E3C corpus.

Entity	**English**	**French**	**Italian**	**Spanish**	**Basque**
CLINENTITY	1024	1327	869	1345	1910
EVENT	4885	4312	3385	4767	7910
ACTOR	682	427	338	319	505
BODYPART	968	659	328	814	1410
TIMEX3	380	333	298	383	638
RML	480	508	383	391	1101
ALINK	114	71	109	92	113
TLINK	4852	4084	1150	4700	7981

Table 2 Annotations in each language in Layer 1 in the E3C corpus.

4 Evaluation

For temporal information and clinical entity annotation tasks, we performed inter-annotator agreement (IAA) tests. We measured whether the guidelines had been defined and were understood correctly, and we ensured that the quality of annotations in the corpus was similar. The IAA phase had been done on the English part of the corpus. IAA for temporal entities (EVENT, TIMEX3, ACTOR, BODYPART) was measured using three annotators and six documents. To compute the agreement, we used the F1-measure metric, which produced the same results as using the Dice coefficient. The agreement is high for EVENT and ACTOR entities (with an average of 0.81 and 0.87), but a bit lower for TIMEX3 and BODYPART (with an average of 0.50 and 0.57). The IAA for temporal relations (TLINK) was split in two phases: three documents were annotated, the results discussed by the annotators and then three new documents were annotated. To measure the agreement, we used the Tempeval-3 scorer (UzZaman and Allen 2011), implemented for the evaluation of systems based on the comparison of temporal graphs built from annotations. The average F1-measure for the first phase was 0.43 and 0.53 for the second.

The annotation of the clinical entities in Layer 1 was performed by four annotators. Again, the agreement is calculated using F1, whereas for the CUI attribute we computed the accuracy taking into consideration only the entities identified by two annotators. The agreement for clinical entity recognition is 0.70 on average (from 0.64 to 0.78). In the entity linking task, the accuracy on entities identified by both annotators starts at 0.86 (on average 0.89).

The clinical entities in Layer 2 were annotated automatically using distant supervision and UMLS as a controlled vocabulary. A manual assessment of the quality of these annotated entities would be too demanding in terms of human resources. For this reason, the quality of Layer 2 has been estimated through an indirect evaluation that uses the results obtained by distant supervision on Layer 1 (Table 3) as an estimation of the quality of the Layer 2 annotations. This approximation is possible because the documents in Layer 1 and Layer 2 are clinical cases and because they were extracted from the same kind of publications or from the same existing corpora.

	English	**French**	**Italian**	**Spanish**	**Basque**
Accuracy	48.33	54.92	58.09	63.64	55.35

Table 3 Estimated accuracy (F_1-measure) of the clinical entities in Layer 2.

5 Conclusions and Results of the Pilot Project

The E3C pilot project aims at fostering advanced information extraction technologies for medical applications. Results include a large corpus of annotated clinical cases in five languages. The corpus is available on the ELG platform.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects and from the Basque Government post-doctoral grant POS_2020_2_0026.

References

Dreisbach, Caitlin, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken (2019). “A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data”. In: *Int. Jour. of Medical Informatics* 125, pp. 37–46. DOI: 10.1016/j.ijmedinf.2019.02.008.

Grabar, Natalia, Cyril Grouin, Thierry Hamon, and Vincent Claveau (2019). “Recherche et extraction d’information dans des cas cliniques. Présentation de la campagne d’évaluation DEFT 2019”. In: *Actes du Défi Fouille de Textes 2019*. Toulouse, France: Actes DEFT 2019, pp. 7–16. URL: https://www.irit.fr/pfia2019/wp-content/uploads/2019/07/actes_DEFT_CH_PFIA2019.pdf.

Johnson, Alistair E.W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark (2016). “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.35.

Kelly, Liadh, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti (2019). “Overview of the CLEF eHealth Evaluation Lab 2019”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro. Cham: Springer, pp. 322–339.

Luo, Yuan, William K. Thompson, Timothy M. Herr, Zexian Zeng, Mark A. Berendsen, Siddhartha R. Jonnalagadda, Matthew B. Carson, and Justin Starren (2017). “Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review”. In: *Drug Safety* 40 (11), pp. 1075–1089. DOI: 10.1007/s40264-017-0558-6.

Miñarro-Giménez, José Antonio, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg (2018). “Qualitative analysis of manual annotations of clinical text with SNOMED CT”. In: *PLoS ONE* 13.12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6307753/pdf/pone.0209547.pdf.

Schulz, Sarah, Jurica Ševa, Samuel Rodríguez, Malte Ostendorff, and Georg Rehm (2020). “Named Entities in Medical Case Reports: Corpus and Experiments”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: ELRA, pp. 4495–4500. URL: https://www.aclweb.org/anthology/2020.lrec-1.553.

Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. (2014). “Temporal Annotation in the Clinical Domain”. In: *Transactions of the Association for Computational Linguistics* 2. Ed. by Ellen Riloff, pp. 143–154. URL: http://aclweb.org/anthology/Q14-1012.

UzZaman, Naushad and James Allen (2011). “Temporal Evaluation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: ACL, pp. 351–356. URL: https://aclanthology.org/P11-2061.

Chapter 18
Extracting Terminological Concept Systems from Natural Language Text

Dagmar Gromann, Lennart Wachowiak, Christian Lang, and Barbara Heinisch

Abstract Terminology denotes a language resource that structures domain-specific knowledge by means of conceptual grouping of terms and their interrelations. Such structured domain knowledge is vital to various specialised communication settings, from corporate language to crisis communication. However, manually curating a terminology is both labour- and time-intensive. Approaches to automatically extract terminology have focused on detecting domain-specific single- and multi-word terms without taking terminological relations into consideration, while knowledge extraction has specialised on named entities and their relations. We present the Text2TCS method to extract single- and multi-word terms, group them by synonymy, and interrelate these groupings by means of a pre-specified relation typology to generate a Terminological Concept System (TCS) from domain-specific text in multiple languages. To this end, the method relies on pre-trained neural language models.

1 Overview and Objectives

Domain knowledge is paramount to any specialised communication setting. A structured representation of domain-specific terminology fosters the acquisition of new domain knowledge, the expansion of existing knowledge, and optimises specialised discourse by supporting terminological consistency (Budin 1996). Extracting Terminological Concept Systems from Natural Language Text (Text2TCS) is a pilot project supported by the European Language Grid (ELG) to develop a language technology that automatically extracts a Terminological Concept System (TCS) from domain-specific texts in multiple languages. A TCS is a terminological resource that conceptually structures domain-specific terms and provides hierarchical and non-hierarchical relations between them. Within the context of terminology science, a *term* signifies a domain-specific designation that linguistically represents a domain-specific concept (ISO1087 2019). A *concept* groups terms by meaning, which is

Dagmar Gromann · Lennart Wachowiak · Christian Lang · Barbara Heinisch
University of Vienna, Austria, dagmar.gromann@univie.ac.at, lennart.wachowiak@univie.ac.at, christian.lang@univie.ac.at, barbara.heinisch@univie.ac.at

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_18

generally represented as unique characteristics shared by a set of real-world entities. Once terms have been grouped into concepts based on their synonymous meaning within languages and equivalent meaning across languages, terminology science foresees interrelations of concepts by terminological relations. Such relations are categorised into hierarchical, i. e., generic and partitive, and non-hierarchical, e. g., causal and spatial, relations. For instance, the sentence *COVID causes coughing* can be depicted as a causal relation from the concept that represents the cause *COVID* to the effect concept designated by *coughing*. However, in practice, publicly available terminologies rarely contain any relations, since manually creating them is time- and labour-intensive. While Automated Term Extraction (ATE) methods have proliferated (e. g., Astrakhantsev 2018; Lang et al. 2021), additionally structuring extracted terms by concepts and relations has been neglected. To address this issue, Text2TCS provides a method and tool to extract terms and interrelations between domain-specific synonym sets across languages and domains. The Text2TCS implementation has been integrated and is available on the ELG plattform.[1]

2 Methodology

The Text2TCS methodology depicted in Figure 1 builds on a pipeline approach with the following steps: preprocessing, term extraction, relation extraction and postprocessing. The pipeline takes domain-specific natural language sentences or text as input and outputs a TCS in the TermBase eXchange (TBX) format and as a concept map. We experimented with several joint term and relation extraction methods, especially relying on pre-trained Neural Machine Translation and Sequence to Sequence models such as mT5 (Xue et al. 2021). However, a pipeline approach relying on fine-tuning XLM-R (Conneau et al. 2020) was finally preferable due to a smaller model size as well as a substantially higher inference speed and performance reliability. In order to fine-tune pre-trained models, training data needs to be available. To this end, two terminologists annotated 51 texts spanning distinct domains from computer science to ecology in English and German with a total of 6,327 terms and 9,460 relations.

2.1 Preprocessing

In a first step, the input text's language is detected and it is split into individual sentences. The former relies on the Python library PYCLD2[2] that supports 83 languages. Language detection is required in order to issue a warning in case the input language is unsupported and to indicate the language in the final TBX output file. Furthermore,

[1] https://live.european-language-grid.eu/catalogue/tool-service/8122

[2] https://github.com/aboSamoor/pycld2

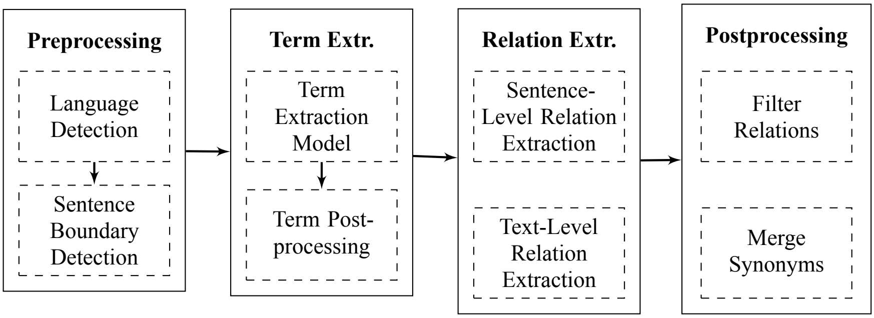


Fig. 1 Text2TCS extraction pipeline

the detected language is passed on to the sentence boundary detection module that relies on language-specific rules.

Sentence boundary detection is achieved using the rule-based Python module pySBD (Sadvilkar and Neumann 2020), which officially supports 22 languages. This step is required due to limited input length of current neural language models and to allow for a sentence-based relation extraction step. Thus, the pipeline can be sure to support 22 languages (two-digit ISO language codes): am, ar bg, da, de, en, es, el, fa, fr, hi, hy, it, ja, kk, mr, my, nl, ru, pl, ur, zh. However, the term and relation extraction models potentially support up to 100 languages.

2.2 Term Extraction

From several distinct experiments with term extraction, which we detail in Lang et al. (2021), the best performing classifies each token of an input sentence separately, utilising the same fully connected layer for all tokens after they have been processed by XLM-R. In term extraction, an established method is (e. g., Hazem et al. 2020) to first generate all possible term candidates from a sequence/sentence and input the candidate together with its context for the model to predict whether it is a term or not. This requires first generating all possible n-grams of a pre-specified length from a text. Instead, the token classification we propose assigns one of three labels to each token in a sequence: `B-T` for beginning of term, `T` for continuation of term, and `n` for not a term (component). For instance, the input sequence “motor vehicle means any power-driven vehicle.” would be labeled as `B-T, T, n, n, B-T, T, n`, extracting the terms “motor vehicle” and “power-driven vehicle”. This approach leads to a substantial reduction in training and inference time compared to previous methods. In XLM-R’s own tokeniser, which we utilise, we noticed an issue with trailing punctuation, e. g. a comma after a term. Thus, we apply an additional cleaning step in which we remove trailing punctuation from a standard punctuation list, unless the punctuation appears multiple times in the term, e. g. “U.S.A.”.

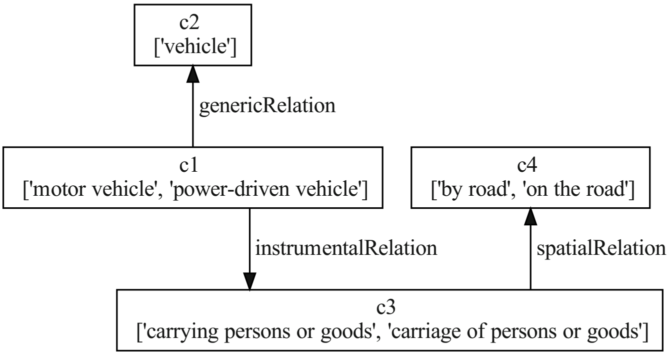

c2
['vehicle']
genericRelation
c1
['motor vehicle', 'power-driven vehicle']
c4
['by road', 'on the road']
instrumentalRelation
spatialRelation
c3
['carrying persons or goods', 'carriage of persons or goods']

Fig. 2 Example TCS from sequence "motor vehicle means any power-driven vehicle, which is normally used for carrying persons or goods by road or for drawing, on the road, vehicles used for the carriage of persons or goods"

2.3 Relation Extraction

Related domain-specific mentions in text can either occur within the same sentence or across sentence boundaries. Thus, two separate models in the pipeline predict relations: a sentence-level and a text-level model. For sentence-level relation extraction, we input a mention pair followed by a contextualising sentence containing both mentions to a fine-tuned pre-trained XLM-R model that predicts a relation taking the relation direction into account (see Wachowiak et al. 2021, for details). We apply our own relation typology of hierarchical relations, i. e., generic and partitive, and non-hierarchical relations, i. e., activity, causal, instrumental, origination, spatial, property, and associative. Generic relations and synonyms frequently occur across sentence boundaries, which is why we additionally train a text-level relation extraction model to detect these two, building on our previous model (Wachowiak et al. 2020) fine-tuning XLM-R. This model takes a mention pair as input and classifies it as a generic relation, synonymy or random, which means no or any other relation. Since predicting relations for individual term pairs drastically impacts inference time, we optimize the pipeline to process multiple term pairs and their context sentence simultaneously.

2.4 Postprocessing

In the last step, synonyms predicted on sentence- and text-level are merged into concepts. Furthermore, the relations predicted by the two models are filtered to only include those with high confidence scores and to remove duplicates to provide the final TCS exemplified in Figure 2.

3 Evaluation

We evaluated individual steps in the pipeline as well as the overall system on manually TCS-annotated texts in English, German, Spanish, Portuguese, French, Italian, Romanian and Russian as well as on standard datasets, where available, for a better comparison. The term extraction model outperforms previous neural approaches (Hazem et al. 2020) from the TermEval challenge by up to 11.6 F1 score and obtained 74% (Precision: 70%, Recall: 78%) on our dataset. The sentence-level relation extraction model obtained a weighted F1 score of up to 53% (Precision: 56%, Recall: 53%) and the text-level relation extraction model of up to 78% (Precision: 78%, Recall: 77%) on our manually annotated datasets. The sentence-level extraction is also compared to a mixed dataset of the SemEval 2007 Task 4 and SemEval 2010 Task 8 relations, on which the model obtains a weighted F1 score of 87% (see Wachowiak et al. 2021, for details).

4 Conclusions and Results of the Pilot Project

Automatically extracting and structuring domain-specific knowledge from text is a challenging task. Text2TCS innovatively fine-tunes pre-trained neural language models in a pipeline approach to first extract terms, second relations on sentence- and text-level, and finally group synonyms. To this end, this pilot project proposed a novel typology of terminological relations. A consistent use of relation types across languages aims to ease the alignment of resulting monolingual TCS across languages. Integrating such an alignment method is future work. At the moment, the method takes terms and relations into consideration, however, text frequently contains (parts of) natural language definitions and their extraction would represent a valuable future addition to the method.

Acknowledgements The work described in this article[3] has received funding from the EU project European Language Grid as one of its pilot projects. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

References

Astrakhantsev, Nikita (2018). "ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala". In: *Language Resources and Evaluation* 52.3, pp. 853–872.

Budin, Gerhard (1996). *Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse*. Vol. 28. Forum für Fachsprachen-Forschung. Gunter Narr Verlag.

[3] https://text2tcs.univie.ac.at

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. ACL, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://www.aclweb.org/anthology/2020.acl-main.747.

Hazem, Amir, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille (2020). "TermEval 2020: TALN-LS2N System for Automatic Term Extraction". In: *Proceedings of the 6th International Workshop on Computational Terminology*. Ed. by Béatrice Daille, Kyo Kageura, and Ayla Rigouts Terryn. Marseille, France: ELRA, pp. 95–100.

ISO1087 (2019). *ISO 1087:2019: Terminology work and terminology science – Vocabulary*. Standard. Geneva, CH: International Organization for Standardization.

Lang, Christian, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann (2021). "Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. ACL, pp. 3607–3620. DOI: 10.18653/v1/2021.findings-acl.316.

Sadvilkar, Nipun and Mark Neumann (2020). "PySBD: Pragmatic Sentence Boundary Disambiguation". In: *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. ACL, pp. 110–114.

Wachowiak, Lennart, Christian Lang, Barbara Heinisch, and Dagmar Gromann (2020). "CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification". In: *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Ed. by Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. ACL, pp. 59–64.

Wachowiak, Lennart, Christian Lang, Barbara Heinisch, and Dagmar Gromann (2021). "Towards Learning Terminological Concept Systems from Multilingual Natural Language Text". In: *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Ed. by Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch. Vol. 93. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 22:1–22:18. DOI: 10.4230/OASIcs.LDK.2021.22.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.

Chapter 19
Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools

Viviana Patti, Valerio Basile, Andrea Bolioli, Alessio Bosca, Cristina Bosco, Michael Fell, and Rossella Varvara

Abstract Starting from the first edition held in 2007, EVALITA is the initiative for the evaluation of Natural Language Processing tools for Italian. We describe the EVALITA4ELG project, whose main aim is to systematically collect the resources released as benchmarks for this evaluation campaign, and make them easily accessible through the European Language Grid platform. The collection is moreover integrated with systems and baselines as a pool of web services with a common interface, deployed on a dedicated hardware infrastructure.

1 Overview and Objectives of the Pilot Project

In Natural Language Processing (NLP), periodic campaigns are a popular means to set benchmarks for specific tasks, stimulate the development of comparable systems and ultimately promote research advancement (Nissim et al. 2017). The validation of NLP models on different datasets strongly depends on the possibility of generalising their results on data and languages other than those on which they have been trained and tested (Magnini et al. 2008). Recent trends are pushing towards proposing benchmarks for multiple tasks (Wang et al. 2018), or for testing the adaptability of systems to different textual domains, genres, and languages, including under-researched and under-resourced ones. The recent specific emphasis on multilingual assessment is also driven by a growing awareness that language technologies can help promote multilingualism and linguistic diversity (Joshi et al. 2020). In this context, the EVALITA4ELG project integrates linguistic resources and language technologies developed under the umbrella of the EVALITA evaluation campaign into the European Language Grid.

Viviana Patti · Valerio Basile · Cristina Bosco · Michael Fell · Rossella Varvara
University of Turin, Italy, viviana.patti@unito.it, valerio.basile@unito.it, cristina.bosco@unito.it, michael.fell@unito.it, rosella.varvara@unito.it

Andrea Bolioli · Alessio Bosca
CELI, Italy, andrea.bolioli@h-farm.com, alessio.bosca@h-farm.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_19

EVALITA[1] is an initiative of the Italian Association for Computational Linguistics (Associazione Italiana di Linguistica Computazionale, AILC[2]). Since 2007, it has been providing a shared framework where different systems and approaches can be evaluated and compared with each other with respect to a large variety of tasks, organised by the Italian research community. The focus of EVALITA is to support the advancement of methodologies and techniques for natural language and speech processing in an historical perspective, beyond the performance improvement, favouring reproducibility and cross-community engagement.

The main goal of the EVALITA4ELG project is to leverage more than a decade of findings of the Italian NLP community, in order to provide easier access to resources and tools for Italian through ELG. We worked towards the achievement of multiples goals, namely: (i) a survey of the tasks organised in the seven editions of EVALITA, released as a knowledge graph; (ii) an anonymisation procedure for improving compliance with current data standard policies; (iii) the integration of resources and systems developed during EVALITA into the ELG platform; (iv) the creation of a unified benchmark for evaluating Italian Natural Language Understanding (NLU); (v) the dissemination of a shared protocol and a set of best practices to describe new resources and tasks in a format that allows a quick integration of metadata into the European Language Grid.

2 Methodology

We started by surveying the tasks organised in EVALITA, collecting the resources and their metadata for upload, and organising this set of information in an ontology. We anonymised the resources according to the current policies for the protection of people's privacy. Finally, we integrated systems and baselines as a pool of web services with a common interface.

2.1 Surveying the EVALITA Tasks

Starting in 2007, EVALITA has been devoted to the evaluation of NLP tools for Italian, providing a shared framework in which participating systems are evaluated on a growing set of different tasks. Rather than being focused on a single task, EVALITA has always been characterised by a wider variety of tasks: each edition of the EVALITA campaign, held in 2007 (Magnini et al. 2008), 2009, 2011 (Magnini et al. 2013), 2014 (Attardi et al. 2015), 2016 (P. Basile et al. 2017), 2018 (Caselli et al. 2018) and 2020 (V. Basile et al. 2020), has been organised around a set of shared tasks dealing with both written and spoken language, varying with respect to the

[1] http://www.evalita.it

[2] https://www.ai-lc.it

challenges tackled and datasets used. The number of tasks has considerably grown, from five tasks, in the first edition in 2007, to 14 tasks in the latest edition held in 2020. Following the trends of other national and international evaluation campaigns, like, e. g., SemEval[3], the typology of tasks also evolved, progressively including a larger variety of exercises oriented to semantics and pragmatics. In particular, the 2016 edition brought a focus on social media data and on the use of shared data across tasks. Open access to resources and research artifacts is deemed crucial for the advancement of the state of the art (Caselli et al. 2018) and the availability of shared evaluation benchmarks is crucial for fostering reproducibility and comparability of results. Organisers were encouraged to collaborate, stimulated to the creation of a shared test set across tasks, and to eventually share all resources with a wider audience. This has resulted in the creation of GitHub public repositories.[4]

2.2 The EVALITA Knowledge Graph

Starting from the semi-structured repositories mentioned in the previous section and from the information collected by surveying seven editions of EVALITA, we built a knowledge graph (KG) that provides the essential information about the editions of the EVALITA evaluation campaign. The KG describes EVALITA in terms of organised tasks, but also of people and institutions that constitute the EVALITA community throughout the years. The KG is structured around an ontology implemented in OWL and it is available both on the website of the EVALITA4ELG project[5] and as a service on the ELG platform. The current version of the ontology comprises 148 classes, 37 object properties and nine data properties. The ontology and the KG are thoroughly described in Patti et al. (2020). As an example, Figure 1 depicts the structure of the KG around the HaSpeeDe2018 task.

The knowledge graph can be queried through a SPARQL endpoint, which allows to inspect the ontology by selecting some variables that occur among the set of triples (subject, predicate, object) composing the knowledge graph. It is thus possible to answer relevant questions related to the EVALITA campaign, extracting information from the KG such as, e. g., “What is the total number of institutions involved as organisers of tasks in all seven EVALITA campaigns?”:

```
SELECT  (COUNT(distinct ?institution) AS ?totalInstitutions)
where {
  ?task e4e:hasInstitution ?institution.
}
>>>> result: 55 <<<<
```

[3] https://semeval.github.io

[4] https://github.com/evalita2016/data

[5] http://evalita4elg.di.unito.it

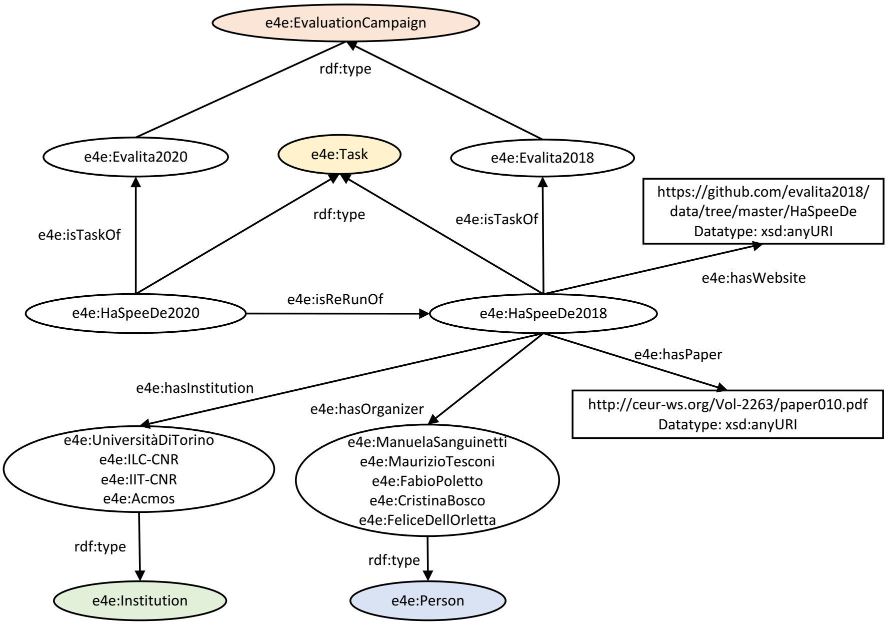


Fig. 1 EVALITA knowledge graph; primary classes are colored and their relations illustrated around the HaSpeeDe2018 task

2.3 Anonymisation of Resources

The EVALITA resources to be made accessible in the ELG platform had to be carefully checked and made compliant with the current policies about data releasing and sharing (e. g., GDPR, Rangel and Rosso 2018), therefore particular attention has been paid to data anonymisation. The datasets collected for EVALITA4ELG were anonymised relying on an automatic anonymisation tool developed in the context of the AnonymAI research project, and then manually reviewed in order to assess their quality. AnonymAI is a nine months research project co-financed by the H2020 project NGI Trust focusing on providing legally compliant anonymisation profiles customised to the needs of end users.

The anonymisation profile applied to the EVALITA4ELG dataset detects and masks person names, phone numbers, email addresses, mentions/replies/retweets, and URLs. The most frequent entities that were masked in the anonymisation process consist of person names and mentions (e. g., in the SardiStance dataset about 50 person names and 150 mentions).

2.4 Release of Data and Models through ELG

At the time of this writing, 51 Language Resources and Technologies are linked to the EVALITA4ELG project in ELG.[6] Eight services were fully integrated into ELG: four of them from the EVALITA 2018 edition, and four of them from the most recent EVALITA 2020 edition. Of the 2018 systems, three are hate speech detection systems (HaSpeeDe 2018 task) and one is Gender Detection (GxG). Of the 2020 systems, two are hate speech detectors (HaSpeeDe 2020 task), one is a POS tagger for spoken language (KIPoS task), and one is a misogyny detection system (AMI task). All datasets and services are accessible interactively from the ELG website or programmatically by means of REST API calls or the ELG-provided Python SDK.

3 Conclusions and Results of the Pilot Project

EVALITA4ELG has been a successful effort towards the inclusion of resources for the Italian language in the European Language Grid. We created a catalogue of resources and models developed during the various editions of the EVALITA campaign, designed in the form of a knowledge graph that can be inspected through SPARQL queries. We collected the original distribution of the resources used for EVALITA tasks and we created 44 entries. For 13 resources, together with CELI, we developed and applied an anonymisation procedure to mask personal and sensitive data. We integrated eight available systems from different tasks into ELG. Finally, we organised an event on September 2021 with hybrid participation[7], including an overview of the project and the results obtained, a tutorial about integrating systems and resources on ELG, and a round table with 14 invited speakers chosen among the most active organisers of tasks of EVALITA.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. The work has also received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 825618 (AnonymAI, NGI Trust).

References

Attardi, Giuseppe, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli (2015). "State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective". In: *Intelligenza Artificiale* 9, pp. 43–61.

[6] https://live.european-language-grid.eu/catalogue/project/1397

[7] http://evalita4elg.di.unito.it/conference

Basile, Pierpaolo, Malvina Nissim, Rachele Sprugnoli, Viviana Patti, and Francesco Cutugno (2017). "EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition". In: *Italian Journal of Computational Linguistics* 3.1, pp. 93–127.

Basile, Valerio, Danilo Croce, Maria Di Maro, and Lucia C. Passaro (2020). "EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian". In: *Proc. of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), 17 Dec. 2020*. Vol. 2765. CEUR Workshop Proceedings.

Caselli, Tommaso, Nicole Novielli, Viviana Patti, and Paolo Rosso (2018). "Evalita 2018: Overview on the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian". In: *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. Ed. by Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. Torino: CEUR Workshop Proceedings, pp. 3–8.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 6282–6293.

Magnini, Bernardo, Amedeo Cappelli, Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Francesca Bertagna, Nicoletta Calzolari, Antonio Toral, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Manuela Speranza (2008). "Evaluation of Natural Language Tools for Italian: EVALITA 2007". In: *Proc. of the 6th Int. Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: ELRA, pp. 2536–2543.

Magnini, Bernardo, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, eds. (2013). *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Rome, Italy, January 24-25, 2012, Revised Selected Papers*. Vol. 7689. Lecture Notes in Computer Science. Springer. URL: https://doi.org/10.1007/978-3-642-35828-9.

Nissim, Malvina, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling (2017). "Last Words: Sharing Is Caring: The Future of Shared Tasks". In: *Computational Linguistics* 43.4, pp. 897–904.

Patti, Viviana, Valerio Basile, Cristina Bosco, Rossella Varvara, Michael Fell, Andrea Bolioli, and Alessio Bosca (2020). "EVALITA4ELG: Italian Benchmark Linguistic Resources, NLP Services and Tools for the ELG Platform". In: *Italian Journal of Computational Linguistics* 6.6-2, pp. 105–129. DOI: https://doi.org/10.4000/ijcol.754.

Rangel, Francisco and Paolo Rosso (2018). "On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks". In: *Language and Law* 5.2, pp. 95–117.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels: ACL, pp. 353–355.

Chapter 20
Lingsoft Solutions as Distributable Containers

Sebastian Andersson and Michael Stormbom

Abstract Lingsoft is one of the leading language technology and language service providers in the Nordic countries. In the Lingsoft Solutions as Distributable Containers (LSDISCO) project, we packaged our language technology tools for distribution as containerised services via the European Language Grid (ELG). As a result, Lingsoft's speech recognition, machine translation, proofing, and morphological analysis was made available to users of the European Language Grid. The services primarily cover Finnish (general and healthcare domain), Swedish (also Finland Swedish), Danish, Norwegian bokmål and nynorsk, and English. The distribution as containerised services is a straightforward way of making our tools available and updated on ELG and we intend to continue to update our service offerings on ELG with new tools and languages as we develop them.

1 Overview and Objectives of the Pilot Project

Lingsoft is one of the leading providers of language technology solutions in the Nordic countries and one of the 100 largest language service providers in the world. The tools and models that Lingsoft contributed to ELG via the Lingsoft Solutions as Distributable Containers (LSDISCO) project already existed and in most cases they were already actively used in production by Lingsoft or our customers. The goal of the LSDISCO project was to make those tools and models available as ELG-compatible services for ELG users (Rehm et al. 2021). This included four types of services:

- *Speech recognition*, with the supported languages being Finnish (general and healthcare domain), Swedish and Norwegian bokmål
- *Machine translation*, for language pairs involving Finnish, Swedish, and English in any combination, as well as both directions of Finnish – German

Sebastian Andersson · Michael Stormbom
Lingsoft, Finland, sebastian.andersson@lingsoft.fi, michael.stormbom@lingsoft.fi

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_20

- *Proofing, entailing spelling and grammar error detection* for Finnish, Swedish, Danish, Norwegian bokmål, and spelling for Norwegian Nynorsk and English
- *Text analysis*, entailing morphological analysis (lemmatization and morphology) and named entity recognition (NER) for Finnish, Swedish, Danish, Norwegian Bokmål, Nynorsk, and English

The end result of the project was a set of high quality NLP tools for the Nordic languages available through ELG, for both commercial and non-commercial use, allowing companies and public organisations throughout Europe to efficiently incorporate Nordic language support in their solutions and services.

2 Methodology

The four types of tools and services in scope for the LSDISCO project – speech recognition, machine translation, proofing and text analysis – have been originally developed at Lingsoft in different periods in the company and software development history and for different primary use cases. The least common denominator was a need for refactoring the tools and service architecture to comply with the ELG requirements. Especially the machine translation tools needed conversion from an internally used tool to enable also external distribution as a service via ELG.

The LSDISCO project was divided into three phases per requirements in the ELG call outline: 1. Experiment; 2. Integration; 3. Dissemination. The Experiment phase consisted of refactoring Lingsoft's tools and architecture to comply with ELG's integration requirements. This phase also included enabling a licensing mechanism for the services and creation or upgrade of the terms of service documentation. For the Integration phase, we selected the option to integrate our services to ELG via a proxy container, as this was the most practical option for us requiring the least amount of additional maintenance. This means that all calls to the ELG service are forwarded to and processed by Lingsoft's back end. Upgrades to the services in Lingsoft's back end per our normal release update cycle, e. g., model improvements, are then immediately available also in ELG. The dissemination phase consisted of advertising Lingsoft's services and the ELG platform on Lingsoft's website and in suitable forums such as conferences and trade fairs.

3 Implementation

Lingsoft's proofing, text analysis and speech recognition services were already to a large extent ready for ELG integration. The improvements made for those largely followed the existing development roadmap. The biggest implementation and refactoring effort in the LSDISCO project was for enabling serving Lingsoft's neural machine translation (NMT) to external users, in this case ELG. The NMT engine and models were migrated from a solution serving “only” Lingsoft’s own translation production to the same Software as a Service infrastructure as our speech recognition. This gave us a scalable back end and the possibility to provide user credentials for NMT usage, thus making important improvements to commercialising Lingsoft’s machine translation and serving also external organisations.

To integrate our services with ELG, we implemented the Lingsoft ELG adapter. The Lingsoft ELG adapter is an API proxy container, illustrated in Figure 1. It exposes the ELG platform’s internal LT Service API specification compatible endpoints and acts as a proxy to the Lingsoft APIs:

ASR API	Lingsoft Speech Recognition API
NMT API	Lingsoft Machine Translation API
LMC API	Lingsoft Language Management Central API (text analysis)

In the proxy container, we implemented the conversion between the ELG and the Lingsoft API specifications. The proxy container also includes the mechanism for forwarding authentication via ELG for Lingsoft’s back end service.

The Lingsoft ELG Adapter was packaged into a Docker image and submitted to DockerHub. Lingsoft then filled in the ELG XML metadata specifications for Lingsoft’s services on the ELG platform, and the ELG technical team could proceed with the actual integration. The DockerHub image of the Lingsoft ELG Adapter was created for ELG, but it can be deployed by other organisations in a Docker environment and integrated with the organisation’s own solutions. All that another

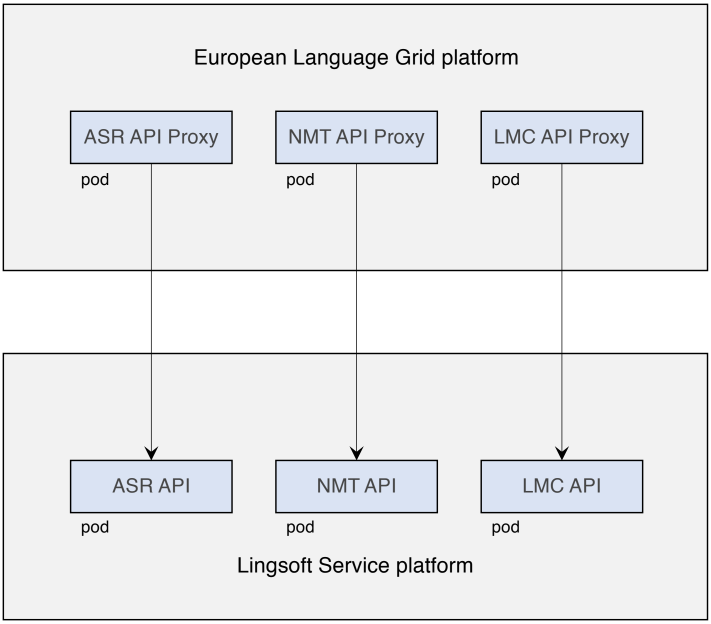


Fig. 1 API proxy containers relay Lingsoft’s services to ELG

organisation would need to deploy the same Docker image into their environment are credentials from Lingsoft that allows calling the Lingsoft back end services.

As the ELG technical team preferred one service per functionality and language. This meant that Lingsoft provided a total of 35 services for ELG integration. The full set of services is presented in Table 1.

Service	Supported Languages/Domains
Speech recognition	Finnish, Finnish Healthcare, Swedish, Norwegian bokmål
Machine translation	Finnish ↔ English, English ↔ Swedish, Finnish ↔ Swedish, German ↔ Finnish
Proofing	Finnish, Finnish Healthcare, Swedish, Finland Swedish, Danish, Norwegian bokmål and nynorsk, English
Morphological analysis (incl. Lemmatization)	Finnish, Swedish, Danish, Norwegian bokmål and nynorsk, English
Named Entity Recognition (NER)	Finnish, Finnish Wikidata, Finnish YSO, Swedish, Danish, Norwegian bokmål and nynorsk, English

Table 1 Lingsoft services and languages

4 Evaluation

Generally, online guidelines and human integration support from ELG were clear and sufficiently detailed throughout the course of the project. The integrated services work per expectation in the “try out” user interface on the ELG platform.

Lingsoft also provided the ELG project with feedback from a commercial perspective regarding the integration process and platform functionality. For example, the demonstration services available in the “try out” box are quite slow. Lingsoft’s speech recognition supports near real-time “live” subtitling/dictation, but this is not yet possible to demonstrate via the ELG platform. The commercial aspects of the platform are also work-in-progress at the time of writing, with no working solution for billing an ELG end user for the use of, e. g., Lingsoft’s services. At present, we provide our solutions through ELG mainly for demonstration purposes, as a marketing channel, and for non-commercial use.

5 Conclusions and Results of the Pilot Project

The ELG project allowed us to upgrade our service infrastructure for easier distribution via ELG as well as through other channels. We believe that we will continue to utilise other providers' ELG resources and services for our benefit, especially open source tools and resources. From our experience with trying to utilise open source tools from the academic community, the ELG approach of researchers (and other developers) providing their open source tools as shareable docker containers with an exposed API is a great improvement over the current situation.

For Lingsoft, ELG can be seen as an additional distribution channel for tools and services we already provide. As an SME from Finland, it is expected that an official EU platform will increase the findability of our services and raise the credibility of our solutions outside of Finland, where we are well known. ELG is therefore expected to facilitate reaching customers outside of Finland and the Nordics.

We provide our tools both for commercial usage (on a Software as a Service subscription model) by companies and organisations, and for research purposes (free of charge for non-commercial use). In our internal work processes, e. g., subtitling and translation, the dockerised tools and API access is ideal, as this facilitates keeping our technology pipeline modular, and the core language technology tools easily replaceable and/or upgradable.

A centralised catalogue of European language technology, if widely adopted, will be beneficial to private providers of language technology, such as Lingsoft, for reaching new customers with our tools and services offerings. Conversely, we hope our contribution to the platform with our services benefit ELG in becoming widely adopted by providing more quality items for the ELG catalogue. Our solutions are robust and widely used with a proven track record. Our spelling and grammar tools have been distributed with the Microsoft Office suite and are used by the Finnish Digital and Population Data Agency, as well as several of the largest newspapers in Sweden; we have collaborated with the Swedish Post and Telecom Authority and the public service broadcaster SVT in creating speech-to-text for Swedish and our Finnish speech-to-text is in use for transcription in a number of Finnish organisations, including the Finnish parliament.

As ELG grows, we believe we will get good exposure for our services by having them on display at ELG. The service adapter ELG integration allows us to continuously improve the content of our ELG services with a minimum of additional maintenance effort. We also intend to continue to release new tools and covered languages in line with our general development roadmap.

Lingsoft is proud to have been one of the selected organisations for the ELG integration projects. We look forward to being part of the continued development of the ELG platform and hope that a substantial part of the ELG visions are fulfilled in the near future.

Lingsoft's services can be found in the European Language Grid.[1]

[1] https://live.european-language-grid.eu/catalogue/search/Lingsoft

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Chapter 21
Motion Capture 3D Sign Language Resources

Zdeněk Krňoul, Pavel Jedlička, Miloš Železný, and Luděk Müller

Abstract The new 3D motion capture data corpus expands the portfolio of existing language resources by a corpus of 18 hours of Czech sign language. This helps alleviate the current problem, which is a critical lack of quality data necessary for research and subsequent deployment of machine learning techniques in this area. We currently provide the largest collection of annotated sign language recordings acquired by state-of-the-art 3D human body recording technology for the successful future deployment of communication technologies, especially machine translation and sign language synthesis.

1 Overview and Objectives of the Pilot Project

Sign language (SL) is a natural means of communication for deaf people. About 70 million people use SL as their first language and there are more than 100 different dialects used around the world. Although significant progress has been made in recent years in the field of language machine learning techniques, the field of SL processing struggles with a critical lack of quality data needed for the successful application of these techniques. SL resources are scarce – they consist of small SL corpora usually designed for a specific domain such as linguistics or computer science. There are some motion capture datasets for American Sign Language (ASL) and French Sign Language (Lu and Huenerfauth 2010; Naert et al. 2017) with a total recorded time of motion of up to 60 minutes. The situation is even worse for "small" languages.

The 3D reconstruction of human body motion using images and depth cameras is a common approach for capturing the movement of the human body (MMPose Contributors 2020). Current large SL datasets are mostly based on 2D RGB videos (Vaezi Joze and Koller 2019; Zelinka and Kanis 2020). The main goal of our project is to deliver a large 3D motion dataset collected using high precision optical marker-based motion capture and to extend the existing ELG portfolio of language resources

Zdeněk Krňoul · Pavel Jedlička · Miloš Železný · Luděk Müller
University of West Bohemia, Czech Republic, zdkrnoul@ntis.zcu.cz, jedlicka@ntis.zcu.cz, zelezny@ntis.zcu.cz, muller@kky.zcu.cz

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_21

by Czech sign language (CSE) data. For comparison SIGNUM, one of the largest video-based SL datasets, contains approximately 55 hours of SL recordings (Koller et al. 2015) and one of the largest 3D motion capture datasets contains only 60 minutes of SL recordings (Naert et al. 2017).

Motion capture technology guarantees precise recording of the signer's movements in 3D space at the cost of a more complex preparation phase compared to standard video recording. Optical marker-based motion capture has become the industry standard for capturing movement of the human body. In Jedlička et al. (2020), we collected the first 3D motion capture dataset for CSE, covering the weather forecast domain. It has a rather limited size and contains recordings of one signer only.

Our contribution can be summarised as follows:

- Proof of concept of large-scale motion capture recording of multiple SL speakers;
- Provide 3D motion capture data to cover wider domains, grammatical context and more signers. We perform proper data post-processing, annotate glosses, and develop tools for data extraction from the collected dataset;
- The largest SL motion capture dataset consisting of recordings of continuous SL phrases and a vocabulary of six native SL speakers from carefully selected domains, in total more than 18 hours;
- Tools that allow searching for individual glosses, phrases, or small movement sub-units (e. g., given hand shape/action) in the dataset.

2 Methodology and Experiment

A new recording procedure for a large amount of 3D motion capturing of SL was investigated to ensure sufficient diversity of SL speakers, grammar, and sign contexts. This makes the new language resource more versatile and useful in many different research fields such as further linguistic and SL motion analyses. The integral part of the experiment is data processing.

In Jedlička et al. (2020), the experimental recording setup with VICON 18 cameras was used as proof of the intended concept. The negative aspect of this setup was its high complexity; the setup was very time demanding and not suitable for large-scale data and multiple speakers.

The new procedure simplifies the process by dividing the setup into two separate parts: large-scale body movement and small-scale, highly detailed finger movement are recorded with two separate motion capture camera setups, each of which uses a reduced number of capture cameras and is adjusted slightly for different speakers.

2.1 Recording Setup

We used our laboratory equipment, i. e., the VICON motion capture system with eight cameras. We extended it with a standard color video camera for a reference video. The frame rate was 100 frames per second (fps) for the motion capture and 25 fps for the reference video. The VICON system records movement using passive retro-reflexive markers attached to the human body. Movement is modeled as a set of movements of the rigid parts connected by the skeleton; the marks are placed on the poles of the rotation axis of the main skeleton joints. Each body part is defined by at least four markers, except fingertips, see Figure 1.

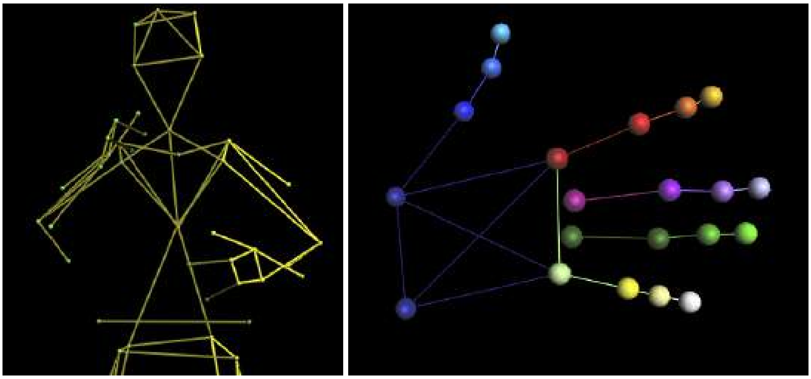


Fig. 1 Visualisation of SL body marker setup (left) and SL hand-shape marker setup (right)

The SL body marker setup is based on marker positions defined by the VICON three-finger standard. It uses a total of 43 markers for tracking upper body, head, arms, and palms movement. A simple hand pose is provided at the same time and incorporates tracking of thumb, index, and little fingertips. Moreover, this setup includes face tracking providing a non-manual component of SL, that is reduced to seven facial markers. The SL hand-shape marker setup is designed for detailed hand-shapes recording. Each hand-shape is recorded separately. Data is recorded for the right hand only. The movement starts from the relaxed hand-shape, then changes to the given hand-shape and back to the relaxed hand-shape. For both setups, data capturing was supervised by CSE linguists.

2.2 Data Annotation

An essential step is the annotation of captured SL utterances. We use time-synchronised reference video, the ELAN tool (Figure 2) and SL experts. The annotation of a sign is done by giving the information of the sign's meaning (gloss), and the right and the left hand-shape. If the sign consists of more than one defined hand-shape, the hand-shapes are annotated as a set of hand-shapes. Both the activities are very laborious and time-consuming. To successfully complete this task, we involved several trained annotators who worked in parallel.

2.3 Data Post-processing

Post-processing consists of data-cleaning, whole-body motion reconstruction, and data-solving. Data-cleaning removes noise and fills gaps in the raw 3D data caused by frequent mutual occlusions of markers during signing, and other noise caused by the environment. Motion reconstruction and data-solving recalculate marker positions into the movement of the skeletal model.

The data of both setups was post-processed. We reconstructed small gaps by the interpolation standard technique as long as the trajectory was simple enough. Note, that the recording speed is 100 fps, which is fast enough to contain minimal changes in trajectory between frames. We used semi-automatic 3D reconstruction of marker trajectories and labeling, and manual cleaning of swaps and gaps. For the body parts defined by at least four markers, filling in the trajectories of the marker is well automatised because at least three points are enough to define the missing position.

The body marker setup uses only one marker per fingertip and some larger gaps caused by more complex self-occlusions of body parts can obscure three or more markers in one rigid segment. Post-processing in those cases is more complicated and gaps must be filled in manually.

The full SL body movement is achieved as a composition of the body movement and corresponding data of the hand-shapes setup. For this purpose, the annotation of hand-shapes provides us temporal segmentation of the recordings. Thus the fingertip motion segments can provide information about dynamic changes during the performance of a particular SL hand-shape in a particular data frame.

The middle part of a given segment is always completed according to the hand-shape(s) assigned by the annotation. We captured full fingers motion only for the transition of the given hand-shape from and to the neutral hand-shape. Thus, for the other frames of the segment, the nearest hand pose with the smallest reconstruction

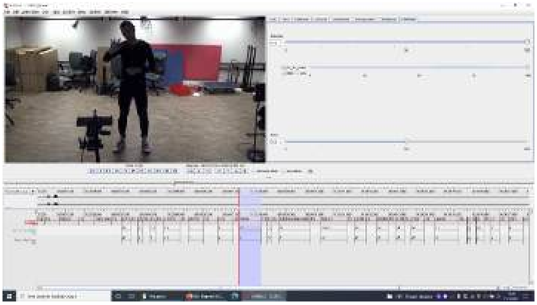


Fig. 2 Example of annotation work in ELAN, specifically designed software for the analysis of sign languages, and gestures

error can be used. We consider only those frames that have an alignment error below a given threshold. The remaining frames will have gaps in the final trajectories.

We solved the above problem as point-set alignment via Procrustes analysis that arises especially in tasks like 3D point cloud data registration. The rigid transformation of two sets of points on top of each other minimises the total distance in 3D between the corresponding markers (Arun et al. 1987). Since the data is noisy, it minimises the least-squares error:

$$err = \sum_{i=1}^{N} ||RM_f^i + t - M_{rf}^i||, \tag{1}$$

where M_f and M_{rf} are current and reference frame(s) respectively as a set of 3D points with known correspondences, R is the rotation matrix and t the translation vector. We define $N = 7$ as three fingertips (thumb, index, little finger), two wrist markers, and two knuckles of the index and little fingers. We aligned just the rotation and translation because the 3D transformation preserves the shape and size (same hand-shape and SL speaker). For the left hand, we mirrored the reference frame(s).

The last step is data-solving. It is a process of reconstruction of the 3D motion of the skeleton from the marker trajectories. For this purpose, we use the VICON software. The skeleton is well defined to directly control the SL avatar animation or handle animation retargeting.

2.4 Dataset Parameters

We limited the linguistic domain to two specific fields to reduce the number of unique signs. Weather forecasts and animal descriptions from the zoological garden domain were selected by CSE linguists. We were also given a list of all hand-shapes which occur in these domains. The dataset is collected from six SL speakers, who differ in their body size, age, and gender.

3 Conclusions and Results of the Pilot Project

SLs are not sufficiently supported through technologies and have only fragmented, weak, or no support at all. Our ELG pilot project offers a new SL resource designed for the development of language technologies (LTs) and multilingual services for Czech. The results contribute to the establishment of the Digital Single Market as one of ELG's objectives. In contrast to the all-in-one recording setup, the body movement is recorded separately from the highly detailed recording of hand poses. This separation reduces the camera setup complexity and the complexity of data during post-processing, which makes SL recording more flexible and adjustments for new SL speakers or data easier.

The project delivered a professionally created SL dataset via state-of-the-art 3D motion capture technology. The project provides data for the wider research community through ELG. We have recorded 18 hours of sign language and recorded six different speakers for two different domains.

We assume our results will be beneficial for other applications such as next generation SL synthesis that uses a 3D animated avatar for natural human movement reproduction or SL analysis or gesture recognition and classification in general.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Arun, K. S., T. S. Huang, and S. D. Blostein (1987). "Least-Squares Fitting of Two 3-D Point Sets". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.5, pp. 698–700. DOI: 10.1109/TPAMI.1987.4767965.

Jedlička, Pavel, Zdeněk Krňoul, Jakub Kanis, and Miloš Železný (2020). "Sign Language Motion Capture Dataset for Data-driven Synthesis". In: *Proceedings of the LREC2020*. Marseille, France: ELRA, pp. 101–106.

Koller, Oscar, Jens Forster, and Hermann Ney (2015). "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". In: *Computer Vision and Image Understanding* 141. Pose & Gesture, pp. 108–125.

Lu, Pengfei and Matt Huenerfauth (2010). "Collecting a motion-capture corpus of American Sign Language for data-driven generation research". In: *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*. ACL, pp. 89–97.

MMPose Contributors (2020). *OpenMMLab Pose Estimation Toolbox and Benchmark*. URL: https://github.com/open-mmlab/mmpose.

Naert, Lucie, Caroline Larboulette, and Sylvie Gibet (2017). "Coarticulation Analysis for Sign Language Synthesis". In: *Universal Access in Human – Computer Interaction. Designing Novel Interactions*. Cham: Springer, pp. 55–75.

Vaezi Joze, Hamid and Oscar Koller (2019). "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language". In: *BMVC*.

Zelinka, Jan and Jakub Kanis (2020). "Neural Sign Language Synthesis: Words Are Our Glosses". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3384–3392. DOI: 10.1109/WACV45572.2020.9093516.

Chapter 22
Multilingual Image Corpus

Svetla Koeva

Abstract The ELG pilot project Multilingual Image Corpus (MIC 21) provides a large image dataset with annotated objects and multilingual descriptions in 25 languages. Our main contributions are: the provision of a large collection of high-quality, copyright-free images; the formulation of an ontology of visual objects based on WordNet noun hierarchies; precise manual correction of automatic image segmentation and annotation of object classes; and association of objects and images with extended multilingual descriptions. The dataset is designed for image classification, object detection and semantic segmentation. It can be also used for multilingual image caption generation, image-to-text alignment and automatic question answering for images and videos.

1 Overview and Objectives of the Pilot Project

Significant progress has been achieved in many multimodal tasks, such as image caption generation, aligning sentences with images in various types of multimodal documents and visual question answering. The shift of traditional vision methods challenged by multimodal big data motivates the creation of a new image dataset, the Multilingual Image Corpus (MIC21).

The MIC21 dataset is characterised by carefully selected images from thematically related domains and precise manual annotation for segmentation and classification of objects in over 20,000 images. The annotation is performed by drawing of or correcting automatically generated polygons, from which bounding boxes are automatically constructed. This allows for wide application of the dataset in various computer vision tasks: image classification, recognition and classification of single objects in an image or of all object instances in an image (semantic segmentation).

The annotation classes which are used belong to a specially designed ontology of visual objects which provides options for extracting relationships between objects in images; the construction of diverse datasets with different levels of granularity of

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria, svetla@dcl.bas.bg

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_22

object classes; and the compilation of appropriate sets of images illustrating different thematic domains. The ontology classes and their definitions, accompanied by illustrative examples, have been translated into 25 languages, which can be used for automatic interpretation of an image, caption generation and alignment of images with short texts such as questions and answers about the image content.

2 Methodology

We have divided the annotation process into four main stages: 1. definition of an ontology of visual objects; 2. collection of appropriate images; 3. automatic object segmentation and classification; and manual correction of object segmentation and manual classification of objects. The dataset contains four thematic domains (sport, transport, arts, security), which group highly related dominant classes such as *Tennis player*, *Soccer player*, *Limousine*, *Taxi*, *Singer*, *Violinist*, *Fire engine*, and *Police boat* in 130 subsets of images. We have used the COCO Annotator (Brooks 2019), which allows for collaborative work within a project, and offers tracking object instances and labelling objects with disconnected visible parts.

2.1 Ontology of Visual Objects

In current practice, WordNet is typically used in generating text queries for the creation of search-based image collections. For example, ImageNet uses 21841 synsets for image collection and their labeling (Russakovsky et al. 2015). A Visual Concept Ontology is proposed which organises concepts (Botorek et al. 2014), containing 14 top-level ontology classes divided into 90 more specific classes. Other datasets use a hierarchical organisation of object classes and mutually exclusive classes (Caesar et al. 2018), however, the number of concepts is usually relatively small.

The ontology of visual objects created for MIC21 embraces concepts that are thematically related and can be depicted in images. The four thematic domains (sport, transport, arts, security) are represented by 137 dominant classes, which show the main "players" within these domains. The ontology also embraces the hypernyms of the dominant classes up to the highest hypernym, which denotes a concrete object, and non-hierarchically related classes (called attributes) (Koeva 2021). The type of dominant class and the type of attribute class determine the type of the relation between them: *has instrument*, *wears*, *uses*, *has part*, etc. For example, the attribute classes for *Billiard player* are *Pool table*, *Billiard ball*, and *Cue*, while for *Bowler* – *Bowling alley*, *Bowl*, *Bowling pin*, *Bowling shoe* etc.; the hypernym classes for *Billiard player* and *Bowler* are *Player*, *Contestant* and *Person*.

Some of the classes and relations are inherited from WordNet (Miller et al. 1990). Additional classes and relations are included in the ontology in case they are not present in WordNet, for example *Bowler wears Bowling shoes*. Using the ontology of visual objects ensures the selection of mutually exclusive classes; the interconnectivity of classes by means of formal relations and an easy extension of the ontology with more concepts corresponding to visual objects.

2.2 Collection of Images and Metadata

The images in the dataset are collected from a range of repositories offering APIs: Wikimedia (images with Public Domain License or Non-copyright restrictions license)[1]; Pexels (images with a free Pexels license allowing free use and modifications)[2]; Flickr (images with Creative Commons Attribution License, Creative Commons Attribution ShareAlike License, no known copyright restrictions, Public Domain Dedication, Public Domain Mark)[3]; Pixabay (images with a free Pixabay license allowing free use, modifications and redistribution)[4]. The Creative Commons Search API is also used for searches on content available under Creative Commons licenses[5]. Over 750,000 images were collected in total and automatically filtered further by image dimensions, license types and for duplication. Each image is equipped with metadata description in JSON format: *filepath*; *source* (name of the repository or service used to obtain the image); *sourceURL* (URL of this repository or service); *license*; *author* (if available); *authorURL* (if available); *domain* (the domain the image belongs to); *width and height* (in pixels) etc.

3 Criteria for the Selection of Images

After the collection of images, we performed additional manual selection to ensure the quality of the dataset, applying the following criteria: i) The image has to contain a clearly presented object described by a given dominant class; ii) The object should (preferably) have no occluded parts; iii) The target object should be in its usual environment and in a position or use that is normal for its activity or purpose; iv) The instances of the target object in different images should not represent one and the same person, animal or artefact; v) Images with small objects, unfocused objects in the background or images with low quality are not selected; vi) Images which represent collages of photos or are post-processed are not selected.

The final selection of images is triple-checked independently by different experts: after the automatic collection, after the automatic generation of segmentation masks and during manual annotation.

[1] https://commons.wikimedia.org/wiki/Commons:Licensing

[2] https://www.pexels.com/license/

[3] https://www.flickr.com/services/developer/api/

[4] https://pixabay.com/service/license/

[5] https://api.creativecommons.org/docs/

3.1 Generation and Evaluation of Suggestions

To accelerate the manual annotation, an image processing pipeline for object detection and segmentation was developed. Two software packages – YOLACT (Bolya et al. 2019) and DETECTRON2 (Wu et al. 2019), and Fast R-CNN (Girshick 2015) models trained on the COCO dataset (Lin et al. 2014) were used for the generation of annotation suggestions. We also performed automatic relabelling for some of the predicted classes (usually for the dominant class and for some of its attribute classes), e. g., the COCO category *Person* within the subset *Golf* from the thematic domain *Sport* is replaced with the class *Golf player*. The performance of the models was evaluated over all domain-specific datasets within the domain *Sport* (see Figure 1).

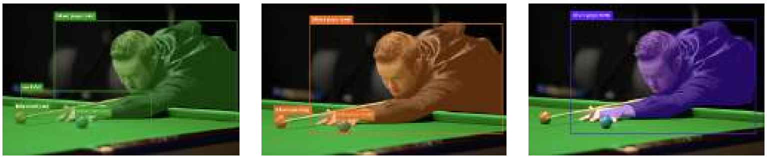


Fig. 1 Annotation results: human (left), YOLACT (middle) DETECTRON2 (right)

The results demonstrate similar behaviour with a slight predominance of one of the models, which was further used to predict the object classes in the datasets from the other three thematic domains. Altogether 253,980 segmentation masks were automatically generated, 194,212 of which were manually adjusted.

3.2 Annotation Protocol

The task for annotators was to outline polygons for individual objects in the image (either by approving or correcting the automatic segmentation or by creating new polygons) and to classify the objects against the classes from the predefined ontology. The annotation follows several conventions:

- An object within an image is annotated if it represents an instance of a concept included in the ontology.
- All objects from the selected dominant class and its attribute classes are annotated (for example, *Gondola* and the related objects *Gondolier* and *Oar*).
- If the object can be associated with different classes, this is recorded within the metadata (for example, for a female soldier – *Soldier* and *Woman*).

Quality control is provided by a second annotator who validates the implementation of the conventions and discusses the quality with the annotation group on a regular basis. If necessary, some of the images are re-annotated.

4 Multilingual Classes

For the purpose of the multilingual description of the images, all ontology classes have been translated into 25 languages: English (Princeton WordNet), Albanian, Bulgarian, Basque, Catalan, Croatian, Danish, Dutch, Galician, German, Greek, Finnish, French, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, and Swedish.

Openly available wordnets have been used from the Extended Open Multilingual WordNet.[6] For the ontology classes which are not inherited from WordNet the appropriate WordNet hypernyms are used. Where WordNet translations are not available, additional sources of translations as BabelNet[7] are employed. The multilingual translations of classes are presented in a separate JSON file which contains information about the language and the translation source. The translations of the ontology classes are accompanied by their synonyms, the concept definition and usage examples (if available in the sources).

5 Conclusions and Results of the Pilot Project

The Multilingual Image Corpus provides fully annotated objects within images with segmentation masks, classified according to an ontology of visual objects, thus offering data to train models specialised in object detection, segmentation and classification (Table 1). The ontology of visual objects allows easy integration of annotated images in different datasets as well as learning the associations between objects in images. The ontology classes are translated into 25 languages and supplied with definitions and usage examples. The explicit association of objects and images with appropriate text fragments is relevant for multilingual image caption generation, image-to-text alignment and automatic question answering for images and video.

Domain	Subsets	Number of Images	Number of Annotations
Sport	40	6,915	65,482
Transport	50	7,710	78,172
Arts	25	3,854	24,217
Security	15	2,837	35,916
MIC21	**130**	**21,316**	**203,797**

Table 1 Multilingual Image Corpus: basic statistics

[6] http://compling.hss.ntu.edu.sg/omw/summx.html

[7] https://babelnet.org/guide

All annotations and image metadata are available for commercial and non-commercial purposes in accordance with the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Bolya, Daniel, Chong Zhou, Fanyi Xiao, and Yong Jae Lee (2019). “YOLACT: Real-time Instance Segmentation”. In: pp. 9156–9165. DOI: 10.1109/ICCV.2019.00925.

Botorek, Jan, Petra Budíková, and Pavel Zezula (2014). “Visual Concept Ontology for Image Annotations”. In: *CoRR*. URL: http://arxiv.org/abs/1412.6082.

Brooks, Justin (2019). *COCO Annotator*. URL: https://github.com/jsbroks/coco-annotator/.

Caesar, Holger, Jasper Uijlings, and Vittorio Ferrari (2018). “COCO-Stuff: Thing and Stuff Classes in Context”. In: *Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218.

Girshick, Ross (2015). “Fast R-CNN”. In: pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.

Koeva, Svetla (2021). “Multilingual Image Corpus: Annotation Protocol”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. INCOMA, pp. 701–707.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár (2014). “Microsoft COCO: Common Objects in Context”. In: *European Conference on Computer Vision (ECCV)*. Zürich, pp. 740–755.

Miller, George, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990). “Introduction to WordNet: An on-line lexical database”. In: *International Journal of Lexicography* 3, pp. 235–244.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision* 116, pp. 157–173.

Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick (2019). *Detectron2*. URL: https://github.com/facebookresearch/detectron2.

Chapter 23
Multilingual Knowledge Systems as Linguistic Linked Open Data

Alena Vasilevich and Michael Wetzel

Abstract Creation and re-usability of language resources in accordance with Linked Data principles is a valuable asset in the modern data world. We describe the contributions made to extend the Linguistic Linked Open Data (LLOD) stack with a new resource, Coreon MKS, bringing together concept-oriented, language-agnostic terminology management and graph-based knowledge organisation. We dwell on our approach to mirroring of Coreon's original data structure to RDF and supplying it with a SPARQL endpoint. We integrate MKS into the existing ELG infrastructure, using it as a platform for making the published MKS discoverable and retrievable via a industry-standard interface. While we apply this approach to LLOD-ify Coreon MKS, it can also provide relevant input for standardisation bodies and interoperability communities, acting as a blueprint for similar integration activities.

1 Overview and Objectives of the Pilot Project

In a world depending on knowledge sharing, data-driven businesses and research communities are concerned with the creation, sharing, and use of language resources in accordance with Linked Data principles, which ensure better data discoverability, standardised structure, and cost savings for all parties involved in the creation of structured data. Robust, coherent, and multilingual information standards are needed to enable information exchange among public organisations, similar to standards that have been fostering technical interoperability for decades (Guijarro 2009).

We extend the Linguistic Linked Open Data (LLOD) stack with a new resource, Multilingual Knowledge System (MKS). MKS caters for the discovery, access, retrieval, and re-usability of terminologies and other interoperability assets organised in knowledge graphs (KG) in a taxonomic fashion. As a semantic knowledge repository, its main forte is the ability to exchange information among acting systems, ensuring that its precise *meaning* is understood and preserved among all parties, in any language. Injecting structure into the language data and expanding the result-

Alena Vasilevich · Michael Wetzel
Coreon GmbH, Germany, alena@coreon.com, michael@coreon.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_23

ing KG with multilingual terminologies, Coreon uses the European Language Grid (ELG) as a platform for making the published resources discoverable and retrievable through SPARQL, a protocol widely used for the retrieval of information from Semantic Web resources. While existing SPARQL tools enable users to query knowledge graphs, they are rarely used for termbases and other terminology resources, i. e., core data sources for translation and localisation (Stanković et al. 2014). This step makes Coreon integration into other systems tool-independent: instead of using the proprietary API, it relies on LLOD standards.

The goal of our contribution is to deliver MKS resources to the Semantic Web community, enabling it to query concept-oriented multilingual structured data with a well-established industry-standard syntax, and to promote the development of data multilingualism within the Semantic Web. In the long run, MKS as a LLOD resource can provide relevant input for standardisation bodies and interoperability communities: acting as a blueprint for similar integration activities, it can be viewed as a starting point for an international standard. We share our experience with ISO/TC37 SC3[1] working groups as a draft for a technical recommendation on how to represent TermBase eXchange (TBX) dialects as RDF.

2 Making Coreon Data Structure LLOD-compatible

Resource Description Framework (RDF) and Web Ontology Language (OWL) are standardised formats for representing Semantic Web data. They support data integration and offer a plethora of tools and methods for data access. SPARQL operates on RDF/OWL resources allowing users to retrieve structured responses to submitted queries. To express queries, it utilises triple patterns that are to be matched by RDF/OWL triples and filter conditions, imposing ranges for literals (Almendros-Jiménez and Becerra-Terón 2021). Despite the emerging interest in publishing terminological resources as linked data, the LLOD stack has not been heavily utilised for this purpose so far (Buono et al. 2020).

We implemented a solution for Coreon MKS, making termbases discoverable and accessible for LLOD systems (Chiarcos et al. 2013). Normally data owners deploy a technology like a RDF triple store for their terminology tool, often developing or setting up a tedious data-mirroring process. We go beyond the limits of RDF/knowledge graph editors, which tend to be good at relation modeling but have weaknesses when it comes to capturing linguistic information.

At the core of the MKS lies a language-independent KG. Unlike other popular solutions within terminology management, linking is performed *not* at the *term* but at the *concept* level; therefore, abstracting from terms, we can model structured knowledge for phenomena that reflect the non-deterministic nature of human language, such as word sense ambiguity, synonymy, and multilingualism. Linking *per concept* also ensures smooth maintenance of relations without additional data clutter:

[1] https://www.iso.org/committee/48136.html

relation edges are independent from labels, terms and their variants, and other metadata. Besides the mirroring process between the Coreon data model and an RDF graph, the RDF vocabulary was established, covering classes, relations, additional term-descriptive information, and administrative metadata. It binds elements into RDF triples. At this stage it was critical to identify information objects and mapping of predicates and literals.

```
{"created_at": "2021-04-20T13:04:59.816Z",
    "terms":[
            {"lang": "en",
             "value": "screen" ,
             "id": "607ed17b318e0c181786b549" ,
            "concept_id": "607ed17b318e0c181786b545",
            "properties": []},
            {"lang": "de",
             "value": "Bildschirm" ,
             "id": "607ed195318e0c181786b55e" ,
            "concept_id": "607ed17b318e0c181786b545",
            "properties": []}
    ],
     "id": "607ed17b318e0c181786b545" }
```

Listing 1 Excerpt of the Coreon data structure.

Listing 1 shows relevant lines within the original JSON data structure that represents the sample concept “screen”, with *concept* ID and individual *term* IDs and their values highlighted. To transform this data structure into an RDF graph, the concept and its two terms are bound together in statements, i. e., RDF triples. Each triple comprises a subject, a predicate and an object; in our case, the concept will act as the subject, the terms become objects and the required predicate is named `hasTerm`. The complete sample set of triples serialised in RDF/Turtle is provided in Listing 2, with highlighted lines 9-10 indicating that the resource with ID `606336dab4dbcf018ed99308` belongs to the OWL class *coreon:Concept* and contains a term with ID `606336dab4dbcf018ed99307`.

In RDF and LOD, data is stored in an atomic manner, with predicates and uniform resource identifiers (URIs) linking elements together. In our case, all instances represented as classes receive unique identifiers. Together with unique IDs, the namespace `coreon:` unambiguously identifies any given element, regardless of whether it is a concept, term, property or a concept relation. Table 1 lists our RDF vocabulary, derived from the original MKS data structure. During the Coreon-to-RDF conversion, there were obvious candidates for classes, like `Concept` and `Term`; yet mirroring descriptive information like `Definition` or `TermStatus` and mapping taxonomic and associative concept relations turned out to be challenging. For the predicates we had to specify what information can be used, defining `owl:range` and `owl:domain`;

```
coreon:607ed17b318e0c181786b547 a coreon:Edge;
  coreon:edgeSource coreon:606336dab4dbcf018ed99308;
  coreon:edgeTarget coreon:607ed17b318e0c181786b545;
  coreon:type "SUPERCONCEPT_OF" .

coreon:606336dab4dbcf018ed99307 a coreon:Term;
  coreon:value "peripheral device"@en .

coreon:606336dab4dbcf018ed99308 a coreon:Concept;
  coreon:hasTerm coreon:606336dab4dbcf018ed99307 .

coreon:607ed17b318e0c181786b545 a coreon:Concept;
  coreon:hasTerm coreon:607ed195318e0c181786b55e,
    coreon:607ed17b318e0c181786b549 .

coreon:607ed17b318e0c181786b549 a coreon:Term;
  coreon:value "screen"@en .

coreon:607ed195318e0c181786b55e a coreon:Term;
  coreon:value "Bildschirm"@de .
```

Listing 2 Triples serialised in RDF / Turtle

```
coreon:hasTerm
  rdf:type owl:ObjectProperty ;
  rdfs:comment "makes a term member of a concept" ;
  rdfs:domain coreon:Concept ;
  rdfs:label "has term" ;
  rdfs:range coreon:Term .
```

Listing 3 Specification of a predicate

e. g., the predicate `hasTerm` can only accept resources of type `coreon:Concept` as a subject (`owl:domain`). Listing 3 provides a full specification of this predicate.

	OWL Type	Coreon RDF Vocabulary
Classes	owl:Class	coreon:Admin, coreon:Edge, coreon:Concept, coreon:Flagset, coreon:Property, coreon:Term
Predicates	owl:ObjectProperty	coreon:hasAdmin, coreon:hasFlagset, coreon:hasProperty, coreon:hasTerm
Values	owl:AnnotationProperty	coreon:edgeSource, coreon:edgeTarget, coreon:id, coreon:name, coreon:type, coreon:value

Table 1 Derived Coreon RDF vocabulary

3 Real-Time Data Access via a SPARQL Endpoint

With the vocabulary defined, we equipped Coreon's export engine with a RDF publication mechanism, including the export in relevant syntax flavours (Turtle, N3, JSON-LD). The Coreon cloud service was supplied with a real-time accessible SPARQL endpoint via Apache Jena Fuseki.[2] It conforms to all published standards and tracks revisions and updates in the under-developed areas of the standard. Running as a secondary index in parallel with the repository's data store, Fuseki catches any changes made by data maintainers, updating the state of the repository in real time. Listing 4 demonstrates a sample SPARQL query over a MKS that deals with wine varieties: here, we want to return all terms, including the values of the *Usage* flag in case the terms have them.

```
SELECT ?t ?termvalue ?usagevalue
    WHERE { ?t rdf:type coreon:Term .
            ?t coreon:value ?termvalue .
            OPTIONAL {  ?t coreon:hasProperty ?p .
                        ?p coreon:key "Usage" .
                        ?p coreon:value ?usagevalue .
            }
    }
```

Listing 4 Sample SPARQL query over MKS

Table 2 shows a subset of the linked data structures returned by this query, i. e., a term's URI, its value, and usage recommendation if available.

[t]	termvalue	usagevalue
http://www.coreon.com/coreon-rdf#[...]8b8aa	Riesling	
http://www.coreon.com/coreon-rdf#[...]8b8bb	Cabernet Sauvignon	Preferred
http://www.coreon.com/coreon-rdf#[...]8b8be	CS	Not allowed
http://www.coreon.com/coreon-rdf#[...]8b8c2	Merlot	

Table 2 Results of the sample SPARQL query (Listing 4): returned grape varieties

4 Conclusions and Results of the Pilot Project

We developed a pipeline to make MKS resources LLOD-compatible, mapping Coreon data structure to RDF, conceiving the Coreon-RDF vocabulary and publishing MKS resources via ELG. Besides making the SPARQL endpoint available

[2] https://jena.apache.org

through ELG, we implemented a productised piece of software, providing TermBase eXchange-like terminology resources in the RDF and Semantic Web context; a set of demo repositories is accessible via the endpoint through ELG. Beyond establishing structural interoperability, the implemented interface bridges Coreon with other Semantic Web systems, enabling querying of elaborate multilingual terminologies. Our mirroring approach can act as a blueprint for similar conversion and integration activities, viewed as a starting point for an international standard. Deployed through ELG, Coreon's SPARQL interface enables the Semantic Web community to query rich heterogeneous MKS data with a familiar, industry-standard syntax, promoting data accessibility and contributing to the development of multilingual resources within the Semantic Web.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Almendros-Jiménez, Jesús Manuel and Antonio Becerra-Terón (2021). "Discovery and diagnosis of wrong SPARQL queries with ontology and constraint reasoning". In: *Expert Systems with Applications* 165, p. 113772. DOI: 10.1016/j.eswa.2020.113772.

Buono, Maria Pia Di, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm (2020). "Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data". In: *Proc. of the 7th Workshop on Linked Data in Linguistics, LDL@LREC 2020, Marseille, France, May 2020*. Ed. by Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia. ELRA, pp. 28–35.

Chiarcos, Christian, Philipp Cimiano, Thierry Declerck, and John P. McCrae (2013). "Linguistic Linked Open Data. Introduction and Overview". In: *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*. Pisa, Italy: ACL, pp. i–xi.

Guijarro, Luis (2009). "Semantic interoperability in eGovernment initiatives". In: *Computer Standards & Interfaces* 31.1, pp. 174–180. DOI: 10.1016/j.csi.2007.11.011.

Stanković, Ranka, Ivan Obradović, and Miloš Utvić (2014). "Developing Termbases for Expert Terminology under the TBX Standard". In: *Natural Language Processing for Serbian-Resources and Applications*, pp. 12–26.

Chapter 24
Open Translation Models, Tools and Services

Jörg Tiedemann, Mikko Aulamo, Sam Hardwick, and Tommi Nieminen

Abstract The ambition of the Open Translation Models, Tools and Services (OPUS-MT) project is to develop state-of-the art neural machine translation (NMT) models that can freely be distributed and applied in research as well as professional applications. The goal is to pre-train translation models on a large scale on openly available parallel data and to create a catalogue of such resources for streamlined integration and deployment. For the latter we also implement and improve web services and computer-assisted translation (CAT) tools that can be used in on-line interfaces and professional workflows. Furthermore, we want to enable the re-use of models to avoid repeating costly training procedures from scratch and with this contribute to a reduction of the carbon footprint in MT research and development. The ELG pilot project focused on European minority languages and improved translation quality in low resource settings and the integration of MT services in the ELG infrastructure.

1 Overview and Objectives of the Pilot Project

OPUS-MT (Tiedemann and Thottingal 2020) provides ready-made server solutions that can be deployed on regular desktop machines to run translations using any NMT model that has been released through the project.[1] The service is powered by Marian-NMT[2] (Junczys-Dowmunt et al. 2018), an efficient open-source framework written in pure C++ with implementations of state-of-the-art neural machine translation architectures. OPUS-MT provides two implementations that can be deployed on regular Ubuntu servers or through containerised solutions using docker images. Both solutions can easily be configured using JSON and can be deployed with a wide range of OPUS-MT models. Multiple translation services and nodes can be combined in one access point through a lightweight API. The coverage is constantly growing and

Jörg Tiedemann · Mikko Aulamo · Sam Hardwick · Tommi Nieminen
University of Helsinki, Finland, jorg.tiedemann@helsinki.fi, mikko.aulamo@helsinki.fi, sam.hardwick@helsinki.fi, tommi.nieminen@helsinki.fi

[1] https://github.com/Helsinki-NLP/Opus-MT

[2] https://marian-nmt.github.io

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_24

improved models are continuously released through our repository as a result of our on-going model training efforts.

A dockerised web app is implemented using the Tornado Python framework, which we adapted for the integration into the European Language Grid environment providing an interface that can seamlessly be deployed in the ELG infrastructure. The essential metadata records for the ELG service catalogue are generated from pre-defined templates using information available from released translation models. The routines support bilingual as well as multilingual models and can also be used to set up access points that serve several translation services. Appropriate docker images are compiled using installation recipes and scripts. We host them on Docker Hub from where they can be pulled by ELG requests to serve translation requests directly through the online APIs. Detailed deployment documentation is available from the repository.[3]

At the time of writing, OPUS-MT provides 89 registered MT services within ELG including a wide variety of bilingual and multilingual models. Registered services can be tested online and can also be accessed through the web API and ELG Python SDK. The translation runs on regular CPUs with minimal resource requirements thanks to the efficient decoder implementation in Marian-NMT. Multilingual models are handled in a special way: multiple source languages can be handled by a single access point whereas multiple target languages require separate access points. Metadata records include the relevant information to describe the service provided.

We also developed plugins for professional translation workflows under the label of OPUS-CAT[4] (Nieminen 2021). Our tools include a local MT engine that can run on regular desktop machines making MT available without the security and confidentiality risks associated with online services. OPUS-CAT integrates with popular translation software such as Trados Studio, memoQ, OmegaT and Memsource. It also provides an integrated fine-tuning procedure for domain adaptation. All OPUS-MT models can be downloaded and used locally with the MT engine, some of the plugins can also fetch translations directly from the OPUS-MT services in ELG.

2 Increasing Language Coverage

The general goal of OPUS-MT is to increase language coverage of freely available machine translation solutions. The project already provides over a thousand pre-trained translation models covering hundreds of languages in various translation directions. The ongoing effort is documented by public repositories and regular updates and we omit further details here as this is a quickly moving target.

Within our ELG pilot project, we further developed our pipelines and recipes to systematically train additional NMT models. The effort resulted in the model de-

[3] https://github.com/Helsinki-NLP/Opus-MT/tree/master/elg

[4] https://helsinki-nlp.github.io/OPUS-CAT/

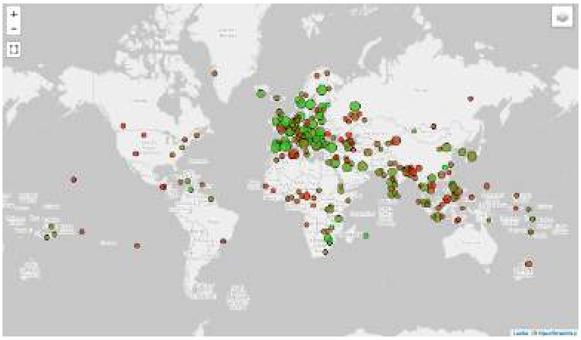


Fig. 1 OPUS-MT map: A visualisation of language coverage and model quality according to automatic evaluation metrics and the Tatoeba MT challenge benchmarks; here: models that translate from a source language mapped on their glottolog location to English; larger circles indicate bigger benchmark test sets and the color scale goes from green (high quality) to red (poor quality)

velopment framework OPUS-MT-train[5] with support for bilingual and multilingual models that can be trained on data provided by OPUS[6] and the Tatoeba translation challenge[7] (Tiedemann 2020).

In order to keep track of the development, we heavily rely on the Tatoeba benchmarks and we implemented an interactive tool to visualize the current state of our released models. Figure 1 shows an example screenshot.

The geographic distribution of released models is an appealing way to uncover blind spots in the NLP landscape. The lack of appropriate data resources is one of the major bottlenecks that block the development of proper MT solutions for most language pairs of the world. Another issue is the narrow focus of research that typically overemphasises well established tasks for reasons of comparability and measurable success. OPUS-MT does not have a strict state-of-the-art development focus based on major benchmarks but rather emphasises language coverage and the focus on under-researched translation directions. The OPUS-MT map and the Tatoeba MT challenge try to make this work visible and more attractive.

The main strategy to tackle issues with *limited data resources* is to apply transfer learning and some type of data augmentation. In OPUS-MT we are constantly facing the problem of limited training data and noise and the ELG pilot project specifically focused on low-resource scenarios and European minority languages.

The idea of transfer learning is based on the ability of models to pick up valuable knowledge from other tasks or languages. In MT, the main type of transfer learning is based on cross-lingual transfer where multilingual translation models can be used to push the performance in low-resource settings (Fan et al. 2021). The effect is typically pronounced with closely related languages where strong linguistic similarities can lead to big improvements across language boundaries (Tiedemann 2021).

[5] https://github.com/Helsinki-NLP/OPUS-MT-train

[6] https://opus.nlpl.eu

[7] https://github.com/Helsinki-NLP/Tatoeba-Challenge/

In OPUS-MT, we therefore focused on multilingual models of typologically related languages. In our setup, we rely on language groups and families established within the ISO 639-5 standard. A dedicated tool for mapping languages to language groups and connecting them with the hierarchical language tree has been developed to allow a systematic development of multilingual NMT models based on typological relationships.[8] The procedures have been integrated in the OPUS-MT training recipes and can be applied to arbitrary datasets from the Tataobea MT Challenge.

Table 1 illustrates the effect of cross-lingual transfer with multilingual models on the example of the Belarusian-English translation benchmark from the Tatoeba MT Challenge. All models apply the same generic transformer-based architecture (Vaswani et al. 2017) with identical hyper-parameters and training recipes.

NMT model	**Belarusian ⟶ English**	**English ⟶ Belarusian**
Belarusian – English	10.0	8.2
East Slavic – English	38.7	20.8
Slavic – English	**42.7**	**22.9**
Indo-European – English	41.7	18.1

Table 1 Machine translation between Belarusian and English with different NMT models; scores refer to BLEU scores measured on the Tatoeba MT Challenge benchmark

The bilingual baseline model is very poor due to the limited training data that is available from the Tatoeba dataset (157,524 sentence pairs). Augmenting the training data with closely related languages such as other (East) Slavic languages leads to significant improvements, which is not very surprising. The effect can be seen in both directions. Note that the multi-target models need to be augmented by language tokens to indicate the output language to be generated. The importance of systematic benchmarks is also shown in the table where we can see that Indo-European language model struggles and the effect of positive transfer diminishes due to the capacity issues of such a complex model setup.

Finally, we also tested a novel type of data augmentation using a rule-based system (RBMT) for back-translation (Sennrich et al. 2016) to produce additional data for the translation from Finnish to Northern Sámi (Aulamo et al. 2021). Our results revealed that knowledge from the RBMT system can effectively be injected into a neural MT model significantly boosting the performance as shown in Table 2.

We use two benchmarks in our evaluations: the UiT set[9], and the YLE set of 150 sentence pairs from news stories about Sámi culture.[10] Preliminary manual evaluation revealed that the NMT-based model was often unable to correctly translate proper names. Adding copies of monolingual data as suggested by Currey et al. (2017) helps to alleviate that issue. Furthermore, we also added experiments with subword regularisation (Kudo 2018) and data tagging (Caswell et al. 2019) to bet-

[8] https://github.com/Helsinki-NLP/LanguageCodes

[9] 2,000 sentence pairs sampled from the Giellatekno Free corpus https://giellatekno.uit.no

[10] Collected from https://yle.fi

	Training Data	UiT	YLE
Baseline	25,106	18.9	4.3
+ NMT-bt	422,596	34.0	9.8
+ RBMT-bt	378,567	36.3	15.5
+ NMT-bt + RBMT-bt	885,301	**40.1**	10.8
+ NMT-bt + copy	845,192	35.7	12.5
+ RBMT-bt + copy	757,134	35.7	**18.6**
+ NMT-bt + RBMT-bt + SR + TB	885,301	40.0	17.2

Table 2 Training data sizes (sentence pairs) and results (BLEU) for the Finnish-Northern Sámi translation models using original parallel data (Baseline), augmented data with back-translations from NMT and RBMT systems (NMT-bt, RBMT-bt), added monolingual data (copy), subword regularisation (SR) and tagged back-translations (TB) evaluated on the UiT and YLE test sets

ter exploit the distributions in the training data and to distinguish between sources with different noise levels. Preliminary results are encouraging and deserve further investigations. In future work, we plan to add pivot-based translation and multilingual models to further improve the performance of the system, to support additional input languages and to include other Sámi language varieties, too.

3 Conclusions and Results of the Pilot Project

OPUS-MT is an on-going effort to make MT widely available for open research and development with an extensive language coverage and well established deployment and integration procedures. Our ELG pilot project made it possible to strengthen the focus on minority languages and to further exploit transfer and data augmentation strategies to improve the quality of MT for under-resourced language pairs.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. We would also like to acknowledge the support by the FoTran project funded by the European Research Council (no. 771113) and CSC, the Finnish IT Center for Science, for computational resources.

References

Aulamo, Mikko, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann (2021). “Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation”. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland: Linköping University Electronic Press, pp. 351–356. URL: https://aclanthology.org/2021.nodalida-main.37.

Caswell, Isaac, Ciprian Chelba, and David Grangier (2019). “Tagged Back-Translation”. In: *Proc. of the Fourth Conf. on Machine Translation*, pp. 53–63.

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield (2017). "Copied Monolingual Data Improves Low-Resource Neural Machine Translation". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: ACL, pp. 148–156. DOI: 10.18653/v1/W17-4715. URL: https://aclanthology.org/W17-4715.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin (2021). "Beyond English-Centric Multilingual Machine Translation". In: *Journal of Machine Learning Research* 22.107, pp. 1–48. URL: http://jmlr.org/papers/v22/20-1307.html.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch (2018). "Marian: Fast Neural Machine Translation in C++". In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: ACL, pp. 116–121. URL: http://www.aclweb.org/anthology/P18-4020.

Kudo, Taku (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 66–75.

Nieminen, Tommi (2021). "OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 288–294. DOI: 10.18653/v1/2021.eacl-demos.34. URL: https://aclanthology.org/2021.eacl-demos.34.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96.

Tiedemann, Jörg (2020). "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT". In: *Proceedings of the Fifth Conference on Machine Translation (WMT)*. ACL, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

Tiedemann, Jörg (2021). "The Development of a Comprehensive Data Set for Systematic Studies of Machine Translation". In: *Multilingual Facilitation*. Ed. by Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. Finland: University of Helsinki, pp. 248–262. DOI: 10.31885/9789515150257.

Tiedemann, Jörg and Santhosh Thottingal (2020). "OPUS-MT – Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. URL: https://helda.helsinki.fi/bitstream/handle/10138/327852/2020.eamt_1_499.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

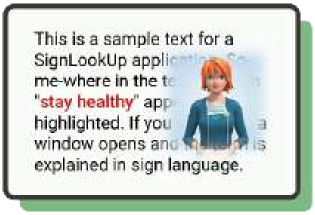
Chapter 25
Sign Language Explanations for Terms in a Text

Helmut Ludwar and Julia Schuster

Abstract The ELG pilot project SignLookUp serves the goal of developing a function that makes text documents easier to comprehend for deaf people. This is important as many of them are functional illiterates.

1 Overview and Objectives of the Pilot Project

The ELG (Rehm et al. 2021) pilot project SignLookUp aims to make texts easier to comprehend for deaf people. Deaf people have a difficult access to texts (Luckner et al. 2005). Learning a written language is a challenge with a hearing impairment (Harris et al. 2017). Therefore, about 75 percent of deaf people are functional illiterates.



This is a sample text for a
SignLookUp appli
me-where in the te
"stay healthy" app
highlighted. If you a
window opens and is
explained in sign language.

Fig. 1 LookApp visualisation

The ideal form of accessibility for the deaf would be the complete translation of texts into sign language. However, this is usually not possible due to limited resources and budgets. The LookApp technology is an intermediate solution and serves the goal of making texts easier to understand for the deaf.

SignLookUp is a technology that links texts to a sign language encyclopedia. Deaf people thus have the possibility to click on difficult or unknown terms in a text and

Helmut Ludwar · Julia Schuster
Sign Time GmbH, Austria, helmut.ludwar@signtime.media, julia.schuster@signtime.media

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_25

immediately receive the explanation or description of the word in their sign language which is displayed adjacent to the text. Using mouseover or clicking on the term, a window pops up and a sign language video is played. Often the explanation of a word or term in sign language is sufficient to make a whole sentence understandable.

SignLookUp starts with two sign languages, but is developed in such a way that it can be easily expanded. The product will be licensed for companies and is free for the end-user (deaf people). This technology thus supports the deaf in accessing and making sense of text information on the internet and at the same time promotes the integration of this marginalised group in our society.

2 Methodology

Selecting the terms that are most important for deaf people to better understand the whole text is a special challenge. On the one hand, it must of course be those that are of central importance, but on the other hand, consideration must also be given to how deaf people experience and understand facts. Last but not least, linguistic peculiarities such as idiomatic expressions, onomatopoeic terms and language images must also be taken into account when finding terms.

Therefore, for the creation of the sign language explanations of an item within a text on a website the following method is used:

1. *Determine the target audience or readers for the website*, e. g., language competence, relevant prior knowledge, thematic interest, age, gender, education.
2. *Perform word analysis* (Egle 2020):

    a. Does the author paraphrase or avoid certain terms in a noticeable manner (euphemisms, taboos)?
    b. Does the text contain words and expressions that must be understood in a figurative sense (linguistic images, metaphors, similes)?
    c. What language-layers or language-uses can be identified?
    d. Does the text contain a foreign word or technical expressions?
    e. Are there words and phrases in the text that can be associated or connoted with other ideas (e. g., "She's feeling blue" → "She's feeling sad")?
    f. Do buzzwords, empty phrases, or other stereotypes occur (e. g., "low-hanging fruit")?
    g. Do certain words acquire a special meaning when the context is taken into account (broadening or narrowing of meaning, emotional coloring)?
    h. From what time do the words used originate? Are they already obsolete (archaism) or newly formed (neologism)? What is their purpose?
    i. Can certain words be assigned to a specific area (e. g., technology, art, sports)? What is the effect?
    j. Are there exaggerations/understatements?
    k. Is only a part of a whole addressed: synecdoche (e. g., pars pro toto)?
    l. Are synonyms (different terms but describing the same in context) used?

3. *Analysis of the text and selection of items:* An automatic analysis of the text to show the comprehensibility and complexity of the text and individual words are used as a starting point, e. g., creation of the readability index (LIX, W. Lenhard and A. Lenhard 2011).[1]
   Thereafter a specialist who is fluent in the languages, e. g., a deaf person or an interpreter, checks whether the passages and terms are understandable for deaf people and selects the candidates for explanation based on the following criteria:
   a. Which terms are of central importance to the content?
   b. Special meaning, e. g., opposite of what is written (irony)
   c. Special words from item 2
4. *Providing the following (meta) information:* Concept (named entity), lemma, context, web link, text language, sign language, version.
5. *Term explanation (for each term):*
   a. Explanation of the term in simple language using the guidelines (Netzwerk Leichte Sprache 2013).
   b. It must not exceed 30 words and must be as brief as possible.
   c. Must be universal and general so that it is suitable for all uses in a text with the same context.
   d. Begins with a relationship to a higher-level or more general term.
   e. Includes the typical features of the term, using semes (the smallest unit of meaning) for this purpose.
   f. Add examples

   As a reference for the creation of explanations available sources may be used, e. .g., medicine DGS[2], medicine ÖGS[3].
6. *Translation into sign language:*
   a. If there is a common sign for the item, it must be used at the beginning, followed by the signed explanation.
   b. Translation into sign language glosses
   c. Transfer into sign language animations
   d. Producing a sign language explanation video
7. *Quality assurance according to the four-eyes-principle:* The draft version of an entry including sign language videos must be checked by a hearing sign language interpreter for completeness and correctness of content. In this way, native speaker competencies of both languages, written and sign language, are included.

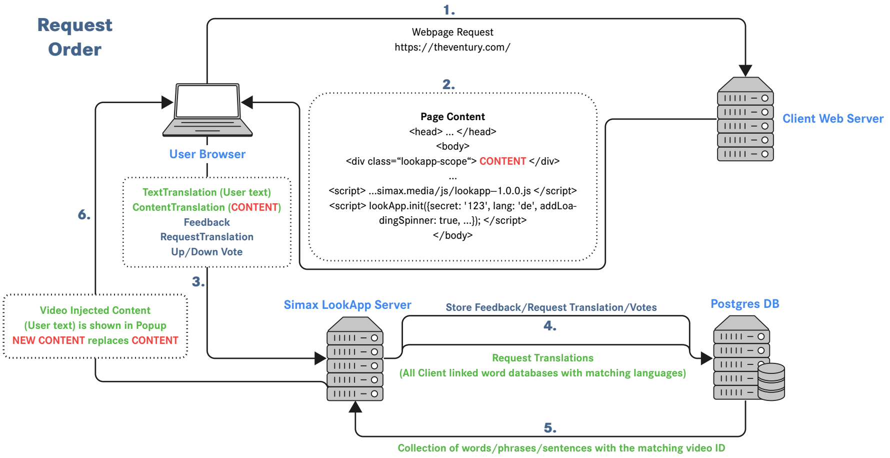
[1] https://wortliga.de/textanalyse/

[2] https://www.sign-lang.uni-hamburg.de/glex/intro/inhalt.html

[3] https://www.equalizent.com

3 Implementation

The beta-version of LookApp (preliminary product name) is implemented in JavaScript on the server where the respective website to be analysed is located. The workflow described below is also shown in Figure 2:


Request Order
1.
Webpage Request
https://theventury.com/
2.
Client Web Server
User Browser
TextTranslation (User text)
ContentTranslation (CONTENT)
Feedback
RequestTranslation
Up/Down Vote
6.
3.
Video Injected Content
(User text) is shown in Popup
NEW CONTENT replaces CONTENT
Simax LookApp Server
Store Feedback/Request Translation/Votes
4.
Postgres DB
Request Translations
(All Client linked word databases with matching languages)
5.
Collection of words/phrases/sentences with the matching video ID

```
Page Content
<head> ... </head>
<body>
<div class="lookapp-scope"> CONTENT </div>
...
<script> ...simax.media/js/lookapp–1.0.0.js </script>
<script> lookApp.init({secret: '123', lang: 'de', addLoa-
dingSpinner: true, ...}); </script>
</body>
```

Fig. 2 LookApp concept

1. End user goes to a website that offers LookApp.
2. The web server returns the content of the page which includes:

    a. Parts of the content with the LookApp-scope class
    b. The LookApp JavaScript is fetched from the LookApp server or served in a static way.
    c. The script is initialised with certain parameters.

3. The request 3 actually represents multiple calls between browser and client

    a. At first the "custom options" are loaded
    b. The client-specific CSS file is loaded
    c. Any LookApp action

4. Depending on the action

    a. Store feedback, requested translations, votes in the database → workflow ends here
    b. Query a list of translations belonging to the client side and corresponding to the passed parameter lang

5. Collection of words and explanations

   a. The server then replaces found words with an icon
   b. JavaScript will interpret as hover or clickable video translations

6. The page content is sent back and replaced by JavaScript.

4 Evaluation

In order to verify the usefulness of the application, a preliminary study was conducted. This involved providing a website with LookApp to a small group of deaf people and then performing a qualitative survey through sign language interpreters.

The results show consistently positive feedback regarding assistance for understanding as well as the avatar used. In order to be able to make reliable statements, however, a survey with a larger test group that represents the deaf community must be carried out.

5 Conclusions and Results of the Pilot Project

As part of the pilot project, a beta version of LookApp was created, which is already being used on early adopter websites, which is why it is evident that the concept and implementation can be used with a positive benefit. Further development of the functions (e. g., use of NLP methods) and the creation of high quality explanations of as many terms as possible are planned next.

It has already been shown in this phase of development that there are multiple advantages. Deaf people have better access to information that cannot be fully translated into sign language due to time or resource constraints. Although our reading aid does not provide the convenience of a full sign language translation, it supports text comprehension in a significant way. Customers who provide large amounts of information or whose content is updated frequently cannot translate all of their content into sign language due to time and economic constraints. With LookApp, even such content can be made much more accessible. Existing and future customers can thus be offered hybrid solutions. In addition to summaries of a website's content in sign language videos according to "Accessibility of websites and mobile applications" (European Parliament, Council of the European Union 2016), LookApp can be implemented for the entire content of the website. Implementing LookApp in a specific website requires only a small financial and organisational effort on the side of the customer but can produce great effects on the side of deaf users.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Egle, Gert (2020). *Leitfragen zur sprachlichen Analyse*. URL: http://teachsam.de/deutsch/d_schreibf/schr_schule/txtanal/txtanal_6_1a.htm.

European Parliament, Council of the European Union (2016). *Directive 2016/2102 of 26 Oct. 2016 on the Accessibility of the Websites and Mobile Applications of Public Sector Bodies*. URL: http://data.europa.eu/eli/dir/2016/2102/oj/eng.

Harris, Margaret, Emmanouela Terlektsi, and Fiona E. Kyle (2017). “Literacy Outcomes for Primary School Children Who Are Deaf and Hard of Hearing: A Cohort Comparison Study”. In: vol. 60. American Speech-Language-Hearing Association. URL: https://pubs.asha.org/doi/pdf/10.1044/2016_JSLHR-H-15-0403.

Lenhard, Wolfgang and Alexandra Lenhard (2011). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. URL: http://rgdoi.net/10.13140/RG.2.1.1512.3447.

Luckner, John, Ann Sebald, John Cooney, John Young, and Sheryl Muir (2005). “An Examination of the Evidence-Based Literacy Research in Deaf Education”. In: *American Annals of the Deaf* 150, p. 443.

Netzwerk Leichte Sprache (2013). *Regeln für Leichte Sprache*. URL: https://www.leichte-sprache.org/wp-content/uploads/2017/11/Regeln_Leichte_Sprache.pdf.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). “European Language Grid: A Joint Platform for the European Language Technology Community”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Chapter 26
Streaming Language Processing in Manufacturing

Patrick Wiener and Steffen Thoma

Abstract Often underestimated, (semi-)structured textual data sources are an important cornerstone in the manufacturing sector for product and process quality tracking. The ELG pilot project SLAPMAN develops novel methods for industrial text analytics in the form of scalable, reusable, and potentially stateful microservices, which can be easily orchestrated by domain experts in order to define quality anomaly patterns, e. g., by analysing machine states and error logs. The results are fully available as open source and integrated into the IIoT toolbox Apache StreamPipes.

1 Overview and Objectives of the Pilot Project

Continuous process and product quality monitoring is a critical task in the manufacturing sector for early detection of anomalies, e. g., gathering insights on potential machine failures, breakouts or performance degradation. Often underestimated, a large part of data sources that are able to provide insights to quality deviations are textual data sources. This includes machine status data and error data, but also production plans. Such information is very important for tracking anomalies and an important source to shop floor workers and other domain experts for identifying potentially critical situations and root causes. While the analysis of real-time measurements is well explored, the automated analysis of textual data is underexplored and hindered by language barriers and often confusing text codes specific to companies or domains. The goal of the SLAPMAN project is the development and integration of streaming language technology (LT) modules from the European Language Grid (ELG, Rehm et al. 2021) to process, analyse and exploit non-structured or semi-structured manufacturing process data. These modules have been integrated into the open-source IIoT toolbox Apache StreamPipes. StreamPipes provides services for self-service data analytics by pursuing a graphical flow-based modeling approach. This allows the description of stream processing applications in the form of processing pipelines composed of multiple, interconnected pipeline elements. This sig-

Patrick Wiener · Steffen Thoma
FZI Research Center for Information Technology, Germany, wiener@fzi.de, thoma@fzi.de

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_26

nificantly lowers rather high technological entry barriers towards making streaming language processing in particular, and LT in general, accessible for non-technical domain experts. SLAPMAN developed novel extensions to Apache StreamPipes that can be easily added to StreamPipes in the form of modular standalone services, e. g., streaming adapters to quickly connect textual data sources (e. g., production plans from MES systems), or pipeline elements for NLP including named entity recognition (NER), tokenising, word embeddings or translation.

2 Graphical, Flow-based Modeling with Apache StreamPipes

Apache StreamPipes[1] is an incubator project of the Apache Software Foundation, that provides a reusable toolbox to easily connect, analyse and exploit a variety of IIoT-related data streams without any programming skills. It leverages different technologies especially from the fields of stream processing, distributed computing, and the semantic web. Riemer et al. (2014) proposed a methodology for semantics-based management of event streams based on the dataflow programming paradigm which is the foundations of StreamPipes. In this regard, StreamPipes allows modelling stream processing applications in the form of processing pipelines. Pipelines comprise a sequence of pipeline elements provided by arbitrary event-driven microservices from an extensible toolbox. Such event-driven microservices are operated in a distributed environment consisting of multiple, potentially heterogeneous runtime implementations. In doing so, this facilitates the distributed execution of pipeline elements to account for business or application-specific requirements. Figure 1 gives a rudimentary overview of a basic named entity recognition pipeline in StreamPipes. The pipeline consists of three pipeline elements, a textual quality report data source for a group of flow rate sensors, a named entity recognition processor based on an ELG service, and a dashboard sink to visualise results.

The decomposition of complex analytical challenges into smaller function blocks allows StreamPipes to mitigate the problem of committing to a single stream processing technology. On top, it uses semantics to guide non-technical domain experts throughout the pipeline creation process. In recent years, several profound extensions to the knowledge base of StreamPipes were implemented to improve and extend existing capabilities. This includes StreamPipes Connect (Zehnder et al. 2020), a semantics-based adapter model and edge transformation functions, and StreamPipes Edge Extensions (Wiener et al. 2020), a methodology for geo-distributed pipeline deployment and operation. Besides StreamPipes, other solutions for low-code dataflow programming exist, e. g., Apache Nifi[2], or Node-RED[3].

[1] https://streampipes.apache.org

[2] https://nifi.apache.org

[3] https://nodered.org

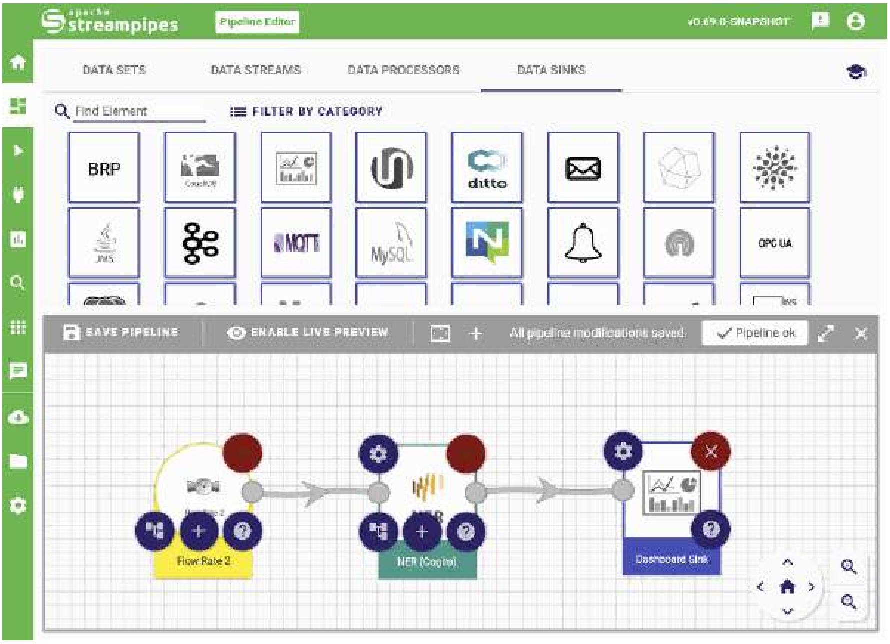


Fig. 1 Example pipeline in StreamPipes

3 Architecture

From an architectural point of view, SLAPMAN follows the microservice architecture of StreamPipes and provides a seamless integration with LT services offered by the ELG platform as shown in Figure 2. In general, the ELG platform provides various LT services that allow to perform language processing and LT-related operation. From a technical perspective, LT services are remotely accessible via REST over HTTP. As such, requests comprising textual data are issued against corresponding LT services that process the incoming call and in return provide the analysis results. For instance, using a machine translation service allows to translate quality defect reports from various plants in different source languages into a common target language, e. g., English, in order to globally investigate certain defect patterns.

In this context, StreamPipes allows to design and develop arbitrary pipeline elements using an SDK. Therefore, arbitrary LT services available on the ELG platform can be wrapped as specific pipeline elements providing language processing capabilities to domain experts to be leveraged in a reusable and self-service manner. Once a user models and deploys a pipeline using one of the LT pipeline elements, textual data is continuously transferred between participating pipeline elements in an event-driven manner by means of a topic-based publish/subscribe pattern. As such, output events from preceding pipeline elements are published to a message broker proto-

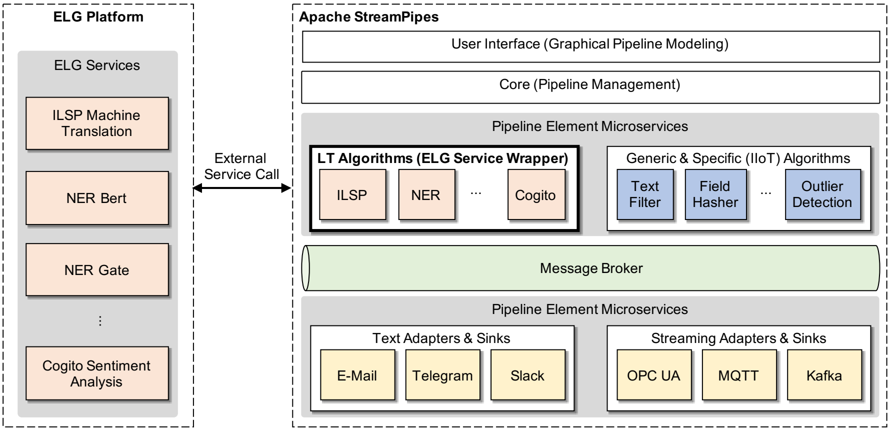

ELG Platform
ELG Services
ILSP Machine Translation
NER Bert
NER Gate
⋮
Cogito Sentiment Analysis
External Service Call
Apache StreamPipes
User Interface (Graphical Pipeline Modeling)
Core (Pipeline Management)
Pipeline Element Microservices
LT Algorithms (ELG Service Wrapper)
ILSP
NER
...
Cogito
Generic & Specific (IIoT) Algorithms
Text Filter
Field Hasher
...
Outlier Detection
Message Broker
Pipeline Element Microservices
Text Adapters & Sinks
E-Mail
Telegram
Slack
Streaming Adapters & Sinks
OPC UA
MQTT
Kafka

Fig. 2 Architecture: ELG platform and StreamPipes integration

col, e. g., Apache Kafka[4]. Succeeding pipeline elements subscribe to relevant topics in order to retrieve the previously published events. The complete life cycle of the event-driven application is internally managed by the core of StreamPipes which is responsible for the pipeline management. This includes pipeline element compatibility based on semantic verification to provide user support and guidance throughout the pipeline modeling process. In addition, this incorporates message broker protocol negotiation including system-side topic management of the publish/subscribe pattern. At run-time, streaming textual events subscribed by LT pipeline elements of deployed pipelines issue REST calls to remote LT services on the ELG platform to perform the essential processing tasks. Results are sent back and published again to the corresponding message broker protocol for further usage. The architectural design of standalone pipeline element microservices facilitates to extend StreamPipes with additional LT components.

4 Implementation

The main activities in SLAPMAN were focused on the development of new extensions for Apache StreamPipes related to language technology. As such, the extensions were focused on i) wrapping and integrating existing services from the ELG platform (e. g., NER, rumour veracity, sentiment analysis, machine translation), ii) developing new data processors and data sinks for Apache StreamPipes related to LT (e. g., chunker, language detection, part-of-speech-tagger, sentence detection, tokeniser), iii) developing additional adapters to connect text-focused data sources (e. g., Telegram, Slack, Manual Input) and iv) developing technical extensions to the

[4] https://kafka.apache.org

toolbox itself to ease the integration of new NLP models along with general usability improvements (e. g., file management, word cloud visualization).

In addition, a new Client API was developed which allows to adapt existing pipelines and to configure pipeline elements from external applications. This enables users to easily update trained language models using a convenient Java client. Moreover, from a deployment and orchestration perspective, StreamPipes relies on Docker as its default installation option. To further alleviate the integration into the ELG platform based on Kubernetes, a helm chart[5] for StreamPipes was developed which is available for public use. This helm chart paired with the general extensibility of StreamPipes to install new pipeline elements providing LT capabilities at run-time allows to integrate additional LT algorithms as demands change.

5 Conclusions and Results of the Pilot Project

In the future, we plan on pursuing the following key activities resulting from lessons learned along the way. In order to better facilitate the integration in existing enterprise architectures, StreamPipes is planned to support standard identity and access management systems such as Keycloak to complement the existing user management. This will also be beneficial for a smoother interaction with the ELG platform itself. In addition, the work on the StreamPipes Python wrapper to simplify the development of new pipeline elements and especially the integration of ELG services is continued. Similarly, the work on the Client API for external pipeline control from code is planned to be pursued.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

[5] https://helm.sh

References

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Riemer, Dominik, Ljiljana Stojanovic, and Nenad Stojanovic (2014). "SEPP: Semantics-Based Management of Fast Data Streams". In: *Proceedings of IEEE 7th International Conference on Service-Oriented Computing and Applications, SOCA 2014*. IEEE, pp. 113–118. DOI: 10.1109/SOCA.2014.52. URL: http://ieeexplore.ieee.org/document/6978598/.

Wiener, Patrick, Philipp Zehnder, and Dominik Riemer (2020). "Managing Geo-Distributed Stream Processing Pipelines for the IIoT with StreamPipes Edge Extensions". In: *Proceedings of the 14th ACM International Conference on Distributed and Event-Based Systems*. DEBS '20. Montreal, Quebec, Canada: ACM, pp. 165–176. DOI: 10.1145/3401025.3401764. URL: https://doi.org/10.1145/3401025.3401764.

Zehnder, Philipp, Patrick Wiener, Tim Straub, and Dominik Riemer (2020). "StreamPipes Connect: Semantics-Based Edge Adapters for the IIoT". In: *The Semantic Web*. Ed. by Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez. Cham: Springer, pp. 665–680.

Chapter 27
Textual Paraphrase Dataset for Deep Language Modelling

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and Otto Tarkka

Abstract The Turku Paraphrase Corpus is a dataset of over 100,000 Finnish paraphrase pairs. During the corpus creation, we strived to gather challenging paraphrase pairs, more suitable to test the capabilities of natural language understanding models. The paraphrases are both selected and classified manually, so as to minimise lexical overlap, and provide examples that are structurally and lexically different to the maximum extent. An important distinguishing feature of the corpus is that most of the paraphrase pairs are extracted and distributed in their native document context, rather than in isolation. The primary application for the dataset is the development and evaluation of deep language models, and representation learning in general.

1 Overview and Objectives of the Pilot Project

Natural language processing research focuses increasingly more at a deeper understanding of language meaning, which is the enabling factor for the next generation of language technology applications. Of especially recent interest are neural meaning representations that are robust to non-trivial re-phrasing of statements with equivalent or near-equivalent meaning. While deep learning methods have effectively solved many supervised learning tasks where large amounts of task-specific training data are available, their performance in representation learning tasks is much weaker (Glockner et al. 2018; Tsuchiya 2018; McCoy et al. 2019). In practical terms, we do not yet have well-proven general methods that, given arbitrary statements with the same contextual meaning but very different wording, would reliably produce highly similar representations for the statements. The fundamental limitation has been the lack of appropriate training data and learning procedures that are able to infer the projection from observable surface forms to faithful semantic representations.

In this ELG pilot project, we set out to address this limitation by building a fully manually annotated paraphrase corpus for Finnish, the Turku Paraphrase Corpus. In

Jenna Kanerva · Filip Ginter · Li-Hsin Chang · Valtteri Skantsi · Jemina Kilpeläinen · Hanna-Mari Kupari · Aurora Piirto · Jenna Saarni · Maija Sevón · Otto Tarkka
University of Turku, Finland, jmnybl@utu.fi, lhchan@utu.fi, figint@utu.fi

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_27

addition to building this resource, we also gathered experience and data regarding how such a resource can be built efficiently and what human resources are needed, built initial models based on the new resource, and produced baseline results.

2 Methodology

The primary distinguishing feature of our corpus compared to other related efforts is its fully manual annotation (as opposed to automatic candidate generation), resulting in paraphrase pairs that are non-trivial and challenging in not being highly lexically related. In other words, an important objective was to avoid bias due to automatic candidate selection so as to obtain a more realistic estimate of the performance of machine learning models on natural language understanding tasks. To this end, we gather source documents that are potentially rich in paraphrases for fully manual paraphrase candidate extraction. These documents include alternative translations of movie subtitles, news headings and articles reporting the same event, discussion forum messages with identical titles and topics, alternative student translations from translation course assignments, and student essays answering the same prompts.

Along with the manual extraction, all paraphrase candidates are manually classified into categories of paraphrases and non-paraphrases according to the developed annotation scheme. The design of the annotation scheme strives to capture varying levels of paraphrasability of candidate paraphrase pairs. We use a scale of four base labels, 1–4, similar to those used in some other paraphrase corpora (Creutz 2018). We define the four base labels as label *1* unrelated sentences, label *2* related but not paraphrases, label *3* paraphrases in the given context but not universally so, and label *4* universal paraphrases. In addition, label *4* paraphrases can be marked with optional flags $>$ or $<$ for subsumption, *s* for style, and *i* for minor deviations. These flags mark properties of the paraphrases that do not fulfill the strict universality criteria of the label *4* due to one of several defined reasons. The subsumption flag means that the paraphrasability is directional; one sentence can be universally substituted by the other, but not the other way around. The style flag means that the paraphrases convey the same meaning, but may have differing tones or registers, which make them not interchangeable in certain circumstances. The minor deviation flag marks minimal differences in meaning (for example, “this” vs. “that”), or grammatical number, person, tense, etc. that can be trivially identified automatically. These flags are independent of each other and thus one label *4* paraphrase pair can have multiple flags, disregarding the directional subsumption flags. More detailed description of the labels together with example annotations is given in the annotation guidelines (Kanerva et al. 2021a).

3 Implementation

The annotation work was carried out by six main annotators, each being a native Finnish speaker with a strong background in language studies by having completed or ongoing studies in a field related to languages or linguistics. Each annotator worked 5–9 months either full or part time in a strong collaboration with a broader project team including supportive roles in the annotation work.

An annotator starts the process by going through the automatically aligned source document pair presented side-by-side in a custom annotation tool[1] developed for the paraphrase extraction, and extracts all interesting paraphrase candidates by selecting the corresponding text passages from both documents. While saving the candidate, together with the text passage pair the tool also saves the actual position of the text passage in the original document, therefore supporting studying the paraphrase pairs in their original document context. To our knowledge, this is the first paraphrase corpus that includes the document context for the released paraphrase pairs. After extracting all interesting paraphrase candidates from the source document pair, the annotator marks the document finished and moves on to the next one.

The extracted paraphrase candidates are automatically transferred to a separate annotation tool[2] developed specifically for paraphrase labeling. In this tool, each pair of paraphrase candidates is shown separately, and the annotator can see the original contexts if necessary. The annotator labels the original paraphrase pair, and has the option to copy the original text and rewrite the texts into full paraphrases (label *4* without flags). In cases where the annotator decided to provide a rewritten pair, two or more pairs of paraphrases are obtained for the corpus: the original pair, and the rewritten pair(s). The annotators are instructed to rewrite the paraphrase candidates in cases where a simple edit, such as word deletion, insertion or synonym replacement, can be naturally constructed and does not require too much effort.

4 Evaluation

The paraphrase label annotation was guided using a shared annotation manual, daily meetings, and regularly assigned double annotation batches in order to ensure annotation consistency between the six annotators. The manual paraphrase extraction did not involve a similarly careful annotator training or consistency monitoring throughout the project. Instead of ensuring each annotator extracting the same segments if given the same text, the objective is to collect a diverse set of different paraphrase candidates, where minor deviations in the personal extraction habits only creates more diversity to the data. In order to study the extraction behaviour of the annotators, we measure the average number of paraphrase pairs extracted from one docu-

[1] https://github.com/TurkuNLP/pick-para-anno

[2] https://github.com/TurkuNLP/rew-para-anno

ment pair, indicating how eager the annotator was to include or exclude borderline uninteresting, extremely difficult or otherwise debatable pairs from the corpus.

While the data sources used in the paraphrase extraction step have distinct characteristics in terms of extraction ratios, we use the subset originating from the alternative subtitles (approx. 80% of the full corpus) for this study in order to account for differing source text proportions between the annotators. We measure the average number of paraphrases extracted from one subtitle document pair (about 15 minutes worth of the subtitled program's runtime), while taking into account all document pairs where the extraction and labeling was carried out by the same annotator, and the document pair resulted at least one extracted paraphrase. The statistics are shown in Table 1, the individual extraction rates falling between 13 and 50 pairs indicating some amount of diversity between the annotators. When measuring the mean lexical similarity of the extracted paraphrase pairs (together with standard deviation) as well as annotated paraphrase label distribution for each annotator, we do not notice any significant difference between annotators oriented towards higher or lower extraction rates. The label distributions are visualised in Figure 1. Finally, in Table 1 we measure the proportion of extracted paraphrase pairs each annotator chose to rewrite during the label annotation (row *Rewritten*), showing large differences among the annotators, between 1.4% and 29.5% of rewritten paraphrase pairs.

	Ann1	**Ann2**	**Ann3**	**Ann4**	**Ann5**	**Ann6**
Extracted pairs	28,685	18,908	9,553	7,713	6,359	1,897
Total extracted (%)	39.1	25.8	13.0	10.5	8.7	2.6
Extracted/doc	23.4	13.2	13.4	22.0	48.9	23.4
Rewritten (%)	6.8	23.4	1.3	29.5	14.9	1.4

Table 1 Comparison of the six annotators in terms of the average number of paraphrase pairs extracted from one 15-min subtitle pair (Extracted/doc), as well as the percentage of paraphrase pairs, where the annotator provided a rewrite (Rewritten); in addition to these two metrics, we also illustrate the total amount of the paraphrase pairs extracted by the annotator (both raw count and percentage); note that the number of extracted paraphrases does not sum up to the total corpus size as the comparison is done on the subtitle subset only (approx. 80% of the full corpus)

In order to ensure the consistency of the label annotation, approx. 2% of the paraphrase pairs are double annotated, where two different annotators annotate the labels independently from one another for the same paraphrase candidates. The two individual annotations are merged and conflicting labels resolved together with the annotation team, resulting in a consolidated subset of consensus annotation. The overall accuracy of the individual annotations against the consensus labels is around 70%, on the full set of labels permitted in the annotation scheme. The level of agreement is on par with similar numbers reported in other paraphrase studies (Dolan and Brockett 2005; Creutz 2018). The agreement measures when calculated separately for each annotator vary between 64% and 76%, the most common disagreements being between the semantically nearest labels (i. e., labels *3* and *4</>*, or labels *4</>* and *4*), or whether to include or not include the rare additional flags *s* or *i*.

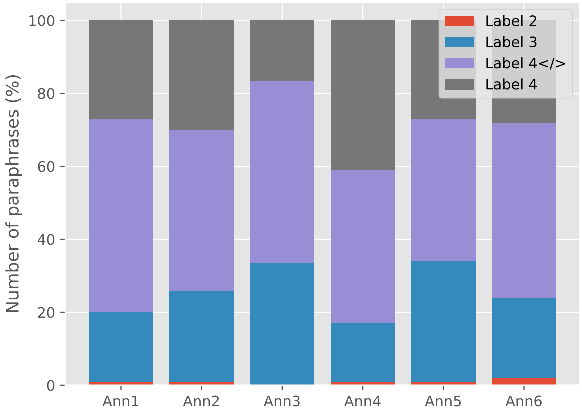


Fig. 1 Label frequencies illustrated separately for the six annotators using the same subtitle subset of the corpus as in Table 1

5 Conclusions and Results of the Pilot Project

The project resulted in a high quality corpus of Finnish paraphrases including a total of 104,645 manually classified pairs, 91,604 being naturally occurring pairs directly extracted from the source documents, while 13,041 are produced through manual rewriting. The manual extraction method presented in the article both skews the label distribution towards true paraphrases ensuring efficient use of human resources (98% being labeled positive) as well as preserves the original document context, making this the first released corpus of paraphrasing in context. The contextual information is used in Kanerva et al. (2021b), where we present a novel approach to paraphrase detection by framing the task as detecting the target paraphrase span from the given document, a similar setting as used in question answering. In addition to the actual corpus, the project also provided models trained for paraphrase classification and fine-tuned sentence representations.

All resources presented in this article are available through the European Language Grid[3] and also on the TurkuNLP website[4] under the CC-BY-SA license.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. In addition, this work was supported by the Academy of Finland and the Digicampus project. Computational resources were provided by CSC – IT Center for Science.

[3] https://live.european-language-grid.eu/catalogue/corpus/7754

[4] https://turkunlp.org/paraphrase.html

References

Creutz, Mathias (2018). “Open Subtitles Paraphrase Corpus for Six Languages”. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 1364–1369.

Dolan, William B. and Chris Brockett (2005). “Automatically Constructing a Corpus of Sentential Paraphrases”. In: *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, pp. 9–16.

Glockner, Max, Vered Shwartz, and Yoav Goldberg (2018). “Breaking NLI Systems with Sentences that Require Simple Lexical Inferences”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 650–655. DOI: 10.18653/v1/P18-2103. URL: https://aclanthology.org/P18-2103.

Kanerva, Jenna, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, et al. (2021a). “Annotation Guidelines for the Turku Paraphrase Corpus”. In: *arXiv preprint arXiv:2108.07499*.

Kanerva, Jenna, Hanna Kitti, Li-Hsin Chang, Teemu Vahtola, Mathias Creutz, and Filip Ginter (2021b). “Semantic Search as Extractive Paraphrase Span Detection”. In: *arXiv preprint arXiv:2112.04886*.

McCoy, Tom, Ellie Pavlick, and Tal Linzen (2019). “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334.

Tsuchiya, Masatoshi (2018). “Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment”. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA, pp. 1506–1511. URL: https://aclanthology.org/L18-1239.

Chapter 28
Universal Semantic Annotator

Roberto Navigli, Riccardo Orlando, Cesare Campagnano, and Simone Conia

Abstract Explicit semantic knowledge has often been considered a necessary ingredient to enable the development of intelligent systems. However, current state-of-the-art tools for the automatic extraction of such knowledge often require expert understanding of the complex techniques used in lexical and sentence-level semantics and their linguistic theories. To overcome this limitation and lower the barrier to entry, we present the Universal Semantic Annotator (USeA) ELG pilot project, which offers a transparent way to automatically provide high-quality semantic annotations in 100 languages through state-of-the-art models, making it easy to exploit semantic knowledge in real-world applications.

1 Overview and Objectives of the Pilot Project

Natural Language Processing (NLP) is the field of Artificial Intelligence (AI) which aims at enabling computers to process, understand and generate text in the same way as we humans do. Although AI systems are nowadays able to process massive amounts of text, they are still far from achieving true Natural Language Understanding (NLU). Indeed, current systems still struggle in explicitly identifying and extracting the meaning or semantics conveyed by a text of interest. Nonetheless, the integration of explicit semantics has already been successfully exploited in a wide array of downstream tasks that span multiple areas of AI from NLP with information retrieval, question answering, text summarisation, and machine translation, to computer vision with visual semantic role labeling and situation recognition. Unfortunately, expert knowledge of lexical semantics, sentence-level semantics and complex deep learning techniques often becomes a roadblock in the integration of explicit semantic information into downstream tasks and real-world applications, especially in multilingual scenarios. To lower the entry point for semantic knowledge integration into multilingual applications, we present the Universal Semantic Anno-

Roberto Navigli · Riccardo Orlando · Cesare Campagnano · Simone Conia
Sapienza University of Rome, Italy, navigli@diag.uniroma1.it, orlando@diag.uniroma1.it, campagnano@di.uniroma1.it, conia@di.uniroma1.it

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_28

tator (USeA) project, the first unified API for three core tasks in NLU: Word Sense Disambiguation (WSD), Semantic Role Labeling (SRL), and Abstract Meaning Representation (AMR) parsing. With USeA, we offer a simple yet efficient way to use state-of-the-art multilingual models within a single framework accessible via REST API, browsers, and programmatically. This will ease the integration of NLU models in NLP pipelines (also for low-resource languages), allowing them to exploit explicit semantic information to improve their performance.

2 Methodology

USeA is the first unified set of APIs for high-performance multilingual NLU, supporting 100 languages. USeA employs state-of-the-art multilingual neural networks to provide automatic semantic annotations for WSD, SRL and AMR Parsing.

Word Sense Disambiguation (WSD) is the task of associating a word in context with its most appropriate sense from a sense inventory (Bevilacqua et al. 2021b). USeA provides word sense labels using an improved version of the state-of-the-art WSD model proposed by Conia and Navigli (2021), which, differently from other ready-to-use tools for WSD based on graph-based heuristics (Moro et al. 2014; Scozzafava et al. 2020) or non-neural models (Papandrea et al. 2017), is built on top of a Transformer encoder. Crucially, thanks to BabelNet 5 (Navigli et al. 2021), a multilingual encyclopedic dictionary, USeA is able to disambiguate text in 100 languages.

Semantic Role Labeling (SRL) is the task of answering the question "Who did What, to Whom, Where, When, and How?" (Màrquez et al. 2008), providing a structured and explicit representation of the underlying semantics of a sentence. Differently from other available SRL systems, USeA encapsulates an improved version of the neural model introduced by Conia et al. (2021a), which performs state-of-the-art cross-lingual SRL with heterogeneous linguistic inventories.

Abstract Meaning Representation (AMR) parsing is the task of capturing the semantics of a sentence through a rooted directed acyclic graph, with nodes representing concepts and edges representing their relations (Banarescu et al. 2013). USeA offers a multilingual version of SPRING (Bevilacqua et al. 2021a), a recent state-of-the-art, end-to-end system for Text-to-AMR generation.

3 Implementation

The USeA pipeline is organised in five self-contained modules that are transparent to the end user, as shown in Figure 1.

Orchestrator Module. The Orchestrator Module is the core of USeA and serves as an entry point for the semantic API. Being an end-to-end system, the end user

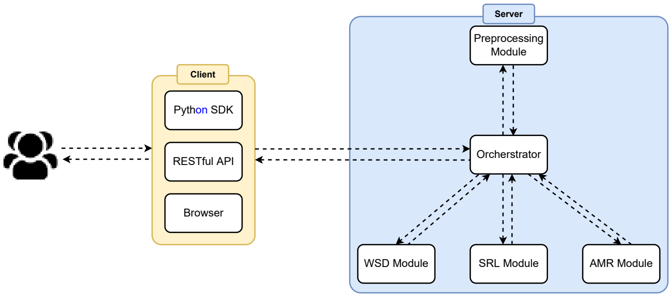


Fig. 1 USeA architecture: a user sends text to the USeA server and receives semantic information; in the server, the orchestrator processes the input using task-specific modules

is only required to send raw text to our service. The input text is then processed by the Preprocessing Module and the result sent to the WSD, SRL and AMR Parsing modules. In particular, since the SRL and AMR Parsing tasks are more demanding, we offload the WSD module to CPU and run SRL and AMR Parsing requests on GPU to optimise hardware usage. The responses from the three semantic modules are then combined and sent back to the end user.

Preprocessing Module. The preprocessing module takes care of producing the preprocessing information that is usually needed by NLP systems, i. e., language identification, document splitting, tokenisation, lemmatisation, and part-of-speech tagging. In order to support as many languages as possible while keeping low hardware requirements, the preprocessing module is built around Trankit (Nguyen et al. 2021) and supports 100 languages with a single model.

WSD Module. We developed AMuSE-WSD (Orlando et al. 2021) as our WSD module. Its neural architecture is based on XLM-RoBERTa (Conneau et al. 2020), a multilingual Transformer model. More specifically, given a word in context, the WSD module i) builds a contextualised representation of the word using the hidden states of XLM-RoBERTa, ii) applies a non-linear transformation to obtain a sense-specific representation, and iii) computes the output score distribution over all the possible senses of the input word.

SRL Module. InVeRo-XL (Conia et al. 2021b) is the SRL system we developed for USeA. Similarly to the WSD module, the SRL module is also based on XLM-RoBERTa. In particular, given an input sentence, the SRL module i) builds a sequence of contextualised word representations using the hidden states of XLM-RoBERTa, ii) identifies and disambiguates each predicate in the sentence, and iii) for each predicate, produces its arguments and their semantic roles.

AMR Parsing Module. The AMR Parsing Module is heavily based on SPRING (Blloshmi et al. 2021), which we extended to support multiple languages. SPRING is a sequence-to-sequence Transformer model that operates as a parser by "translating" an input sentence into a linearised AMR graph. We extend SPRING to support 100 languages by replacing BART with the multilingual version of T5.

	English datasets						Multilingual datasets		
	Se2	Se3	Se07	Se13	Se15	All	Se13	Se15	Xl-Wsd
Moro et al. (2014)	67.0	63.5	51.6	66.4	70.3	65.5	65.6	–	52.9
Papandrea et al. (2017)	73.8	70.8	64.2	67.2	71.5	–	–	–	–
Scozzafava et al. (2020)	71.6	72.0	59.3	72.2	75.8	71.7	73.2	66.2	57.7
USeA (WSD)	**77.8**	**76.0**	**72.1**	**77.7**	**81.5**	**77.5**	**76.8**	**73.0**	**66.2**

Table 1 English WSD results in F1 scores on Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), and the concatenation of the datasets (ALL); we also include results on multilingual WSD in SemEval-2013 (DE, ES, FR, IT), SemEval-2015 (IT, ES), and XL-WSD (average over 17 languages, English excluded)

	Catalan	Czech	German	English	Spanish	Chinese
AllenNLP's SRL demo	–	–	–	86.5	–	–
InVeRo	–	–	–	86.2	–	–
USeA (SRL)	**83.3**	**85.9**	**87.0**	**86.8**	**81.8**	**84.9**

Table 2 Comparison between USeA and other recent automatic tools for SRL; F1 scores on argument labeling with pre-identified predicates on the CoNLL-2012 English test set and the CoNLL-2009 test sets converted from dependency-based to span-based

4 Evaluation

USeA offers state-of-the-art models for multilingual WSD, SRL and AMR Parsing. Here, we report its results on standard gold benchmarks for each task.

Results in WSD. We evaluate our WSD Module against other disambiguation tools on gold standard benchmarks for English and multilingual WSD, covering 17 languages. The results (Table 1) show that USeA outperforms its competitors by a wide margin, especially in multilingual WSD (+8.5% in F1 Score on XL-WSD).

Results in SRL. We report the performance of our SRL Module on two gold standard benchmarks for SRL, CoNLL-2009[1] and CoNLL-2012, covering six languages. USeA is the first package to provide annotations in languages other than English while also outperforming its competitors in English (Table 2).

Results in AMR Parsing. Finally, we examine the performance of our AMR Parsing Module on AMR 3.0[2], which is currently the largest AMR-annotated corpus. Even though USeA supports 100 languages, it is still competitive with other recently proposed English-only AMR parsing systems (Table 3).

[1] The CoNLL-2009 dataset was originally intended for dependency-based SRL. We convert dependency-based annotations to span-based annotations using the gold syntactic trees.

[2] https://catalog.ldc.upenn.edu/LDC2020T02

	SMATCH
Lyu et al. (2021)	75.8
Zhou et al. (2021)	81.2
SPRING (Bevilacqua et al. 2021a)	83.0
USeA (AMR-Parsing)	80.9

Table 3 SMATCH score obtained by USeA compared with recent literature on AMR 3.0 (English)

5 Conclusions and Results of the Pilot Project

We presented the USeA project, providing an overview on its objectives and on how we worked towards achieving them. We hope that USeA will represent a useful tool for the integration of explicit semantic knowledge – word meanings, semantic role labels, and graph-like semantic representations – into real-world applications.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (2013). “Abstract Meaning Representation for Sembanking”. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186. URL: https://aclanthology.org/W13-2322.

Bevilacqua, Michele, Rexhina Blloshmi, and Roberto Navigli (2021a). “One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline”. In: *Proc. of AAAI* 35.14, pp. 12564–12573. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17489.

Bevilacqua, Michele, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli (2021b). “Recent Trends in Word Sense Disambiguation: A Survey”. In: *Proc. of IJCAI-21*, pp. 4330–4338. DOI: 10.24963/ijcai.2021/593.

Blloshmi, Rexhina, Michele Bevilacqua, Edoardo Fabiano, Valentina Caruso, and Roberto Navigli (2021). “SPRING Goes Online: End-to-End AMR Parsing and Generation”. In: *Proceedings of EMNLP*, pp. 134–142. URL: https://aclanthology.org/2021.emnlp-demo.16.

Conia, Simone, Andrea Bacciu, and Roberto Navigli (2021a). “Unifying Cross-Lingual Semantic Role Labeling with Heterogeneous Linguistic Resources”. In: *Proceedings of NAACL*, pp. 338–351. URL: https://www.aclweb.org/anthology/2021.naacl-main.31.

Conia, Simone and Roberto Navigli (2021). “Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration”. In: *Proceedings of EACL*, pp. 3269–3275. URL: https://www.aclweb.org/anthology/2021.eacl-main.286.

Conia, Simone, Riccardo Orlando, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli (2021b). “InVeRo-XL: Making Cross-Lingual Semantic Role Labeling Accessible with Intelligible Verbs and Roles”. In: *Proceedings of EMNLP*, pp. 319–328. URL: https://aclanthology.org/2021.emnlp-demo.36/.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. ACL, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://www.aclweb.org/anthology/2020.acl-main.747.

Lyu, Chunchuan, Shay B. Cohen, and Ivan Titov (2021). "A Differentiable Relaxation of Graph Segmentation and Alignment for AMR Parsing". In: *Proc. of EMNLP*, pp. 9075–9091. URL: https://aclanthology.org/2021.emnlp-main.714.

Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson (2008). "Semantic Role Labeling: An Introduction to the Special Issue". In: *Comp. Linguistics* 34.2, pp. 145–159. URL: https://aclanthology.org/J08-2001.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli (2014). "Entity Linking meets Word Sense Disambiguation: A Unified Approach". In: *TACL* 2, pp. 231–244. URL: https://aclanthology.org/Q14-1019.

Navigli, Roberto, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi (2021). "Ten Years of BabelNet: A Survey". In: *Proc. of IJCAI-21*, pp. 4559–4567. DOI: 10.24963/ijcai.2021/620.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). "Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 80–90. DOI: 10.18653/v1/2021.eacl-demos.10. URL: https://aclanthology.org/2021.eacl-demos.10.

Orlando, Riccardo, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli (2021). "AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 298–307. DOI: 10.18653/v1/2021.emnlp-demo.34. URL: https://aclanthology.org/2021.emnlp-demo.34.

Papandrea, Simone, Alessandro Raganato, and Claudio Delli Bovi (2017). "SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: ACL, pp. 103–108. DOI: 10.18653/v1/D17-2018. URL: https://www.aclweb.org/anthology/D17-2018.

Scozzafava, Federico, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli (2020). "Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 37–46. DOI: 10.18653/v1/2020.acl-demos.6.

Zhou, Jiawei, Tahira Naseem, Ramón Fernandez Astudillo, and Radu Florian (2021). "AMR Parsing with Action-Pointer Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 5585–5598. DOI: 10.18653/v1/2021.naacl-main.443. URL: https://aclanthology.org/2021.naacl-main.443.

Chapter 29
Virtual Personal Assistant Prototype YouTwinDi

Franz Weber and Gregor Jarisch

Abstract YouTwinDi is the next step in a digitised world in which the digital twin evolves and interacts with other digital twins and makes autonomous decisions in the interest of its human twin. In this scenario, security and digital ethics assure ethical decisions and IT specialists concur on improving the digital landscape with ethical models. This vision also includes overcoming language barriers. A continuous match of supply and demand as well as tailored searches help human twins to improve their lives in all respects. YouTwinDi uses the most advanced translation and language analysis technologies, allowing the user and its digital twin to interact with all European citizens without being blocked by language barriers.

1 Overview and Objectives of the Pilot Project

The goal of this ELG (Rehm et al. 2021) pilot project was to build the prototype of a personal virtual assistant, which can be installed on a small device or integrated in an ELG-compatible container. We wanted to demonstrate that this can be accomplished using ELG language resources and technologies while keeping highest security standards. We use the open source software EDDI which is running in a docker container for the natural language interface. This prototype is the basis for the development of a minimum viable product ready for market launch.

We believe that conversational AI applications are well suited to support interactions between people that speak different languages due to their real-time nature and the ability to create personalised customer experiences at scale.

In line with the broader ELG principle that "with 24 official EU and many more additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies", the YouTwinDi[1] solution was developed on top of our existing technology and integrated into the European Language Grid. We use APIs to translate text input (or speech input, via speech-

Franz Weber · Gregor Jarisch
Labs.ai, Austria, franz@labs.ai, gregor@labs.ai

[1] https://www.youtwindi.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_29

to-text technologies) and to recognise intents to query specific data sources and to provide feedback in the language spoken by the user either in written or spoken form (via text-to-speech technologies). YouTwinDi uses these features to add translations of web audio and video streams and to convert the channels into text streams – two appropriate examples are the automatic translation of the European Commission's LinkedIn broadcast events or the automatic translation of local radio stations. Through the integration of ELG APIs we can also integrate technologies such as sentiment analysis into YouTwinDi. Such features are fundamental especially for public institutions to better support citizens.

2 Methodology

The basis for the Digital Twin prototype is our open source chatbot framework EDDI (Enhanced Dialogue Driven Intelligence).[2] This solution has several features that simplify the integration of and with the available ELG resources.

Our software development process is based on the agile software development approach, in particular on Scrum. All product features are listed and prioritised in a product backlog, which consists of what needs to be done to successfully deliver a working software system, including bug fixes and non-functional requirements.

Cross-functional teams estimate and sign up to deliver potentially shipable increments of software during successive sprints, typically lasting 30 days. Once a sprint's backlog is committed, no further functionality can be added to the sprint except by the team. Once a sprint has been delivered, the product backlog is analysed and re-prioritised, if necessary, and the next set of deliverables is selected for the next sprint. From the lean product development best practices we have adopted the concept of minimum viable product (MVP) as a strategy to avoid building products that customers do not need or want, realising often the product with the agreed number of features and the minimum level of quality that can be easily verified by senior users. We develop our solution keeping in mind the ability to interface with external services and resources via APIs and building software development kits. This allows us to integrate fast and to test the integration with available ELG building blocks.

Each feature under development was monitored in terms of costs (human resources and hardware as well as software resources) and in terms of delivery. Acceptance tests were linked to use cases and test criteria. Integration has always been important for us as an open source solution provider, which is why all our software features are available at the API level. Modern concepts as Graph API and authentication and authorisation security are at the core of our software development methodology, allowing for easy testing and integration with existing systems.

Our development strictly follows the Service Oriented Architecture (SOA) concept, removing the bottleneck of dependencies and permitting the usage of independent layers to achieve the development goals. We also subscribe to the concept of

[2] https://www.eddi.labs.ai

microservices (already adopted by ELG), which allows us to easily embed our solution in the ELG ecosystem. Our goal was to develop a portable solution that can run on a small hardware solution (e. g., Raspberry Pi) and that can also be interfaced with the ELG platform or directly embedded in ELG as a container.

We value change management and have documented all steps to integrate our solution using “how to” documents and guidelines.

2.1 Use Case 1: Automated Translation of local News

The Newbly[3] use case relates to the delivery of local news in foreign languages (see Figure 1). In this use case, the user interacts via text or voice with YouTwinDi.

- The automated translation translates the topic expressed in the search query into the local language (set in the configuration).
- YouTwinDi initiates a look up for the topic in local news and social media in the local language.
- YouTwinDi checks if the news is categorised as fake, in which case the user is alerted and asked if they want to proceed anyway.

If the news is not categorised as fake, the user is presented with the news and the news is stored in order to be periodically checked against the fake news database, which case YouTwinDi will notify the user accordingly.

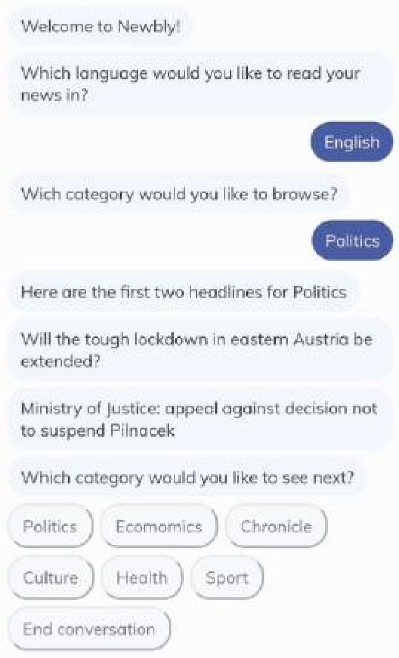
2.2 Use Case 2: Secure Communication between Virtual Assistants

The second use case revolves around communication between *multiple* virtual assistants. Imagine a friend has a wish list on an ecommerce platform – you could ask your friend for access to this list, but that would make your friend anticipate the present. One solution for this challenge can be personal assistants negotiating for a piece of information. Your bot could ask your friend’s bot what to gift the friend based on the online wish list, which, in the case of Amazon, is provided by Alexa. As your and your friend’s virtual assistants are “friends” themselves (trusted domain), they are allowed to communicate such information without your friend receiving a notification.

3 Implementation

The pilot project consisted of five work packages. Work Package 1 was dedicated to the research of potential suitable hardware to be used for the prototype. In addi-

[3] https://newb.ly



Welcome to Newbly!

Which language would you like to read your news in?

English

Wich category would you like to browse?

Politics

Here are the first two headlines for Politics

Will the tough lockdown in eastern Austria be extended?

Ministry of Justice: appeal against decision not to suspend Pilnacek

Which category would you like to see next?

Fig. 1 YouTwinDi use case 1: automated translation of local news

tion we verified if running a containerised version of EDDI would be possible on the shortlisted hardware. For the prototype we decided to use a standard Android smartphone. We also specified the use cases (see Sections 2.1 and 2.2).

Work Package 2 focused upon the integration of EDDI into the ELG platform and setting up the needed containers. We implemented the two use cases, as defined in WP1, on the Android phone. The first use case is defined as translating news from the German language into other languages, such as Romanian or Croatian using machine translation tools available in ELG. The second use case concentrates on the communication between two virtual assistants where one wants to obtain a birthday wish list from the other assistant's owner.

Work Package 3 concentrated on preparing the hardware and installing the software including the use cases on the selected Android smartphone running in a container. In order to accomplish this some modifications had to be applied to the operating system. Afterwards we could easily install EDDI running in a container, however, we came to realise that the ELG language technology tools would be too large to run on the smartphone in a container. From a security point of view our goal was to have all technologies on the device in order to provide maximum security and privacy to users. As this was not possible, we decided for the prototype to be able to call remote services.

Work Package 4 was dedicated to finalising and testing the prototype. In addition, we created a presentation and documented which compromises we had to engage in compared to the initial specification in WP 1 and WP 2.

Work Package 5 took care of all dissemination activities. This was an ongoing process from the beginning to the end of the pilot project. We set up a project website[4] which was updated on a regular basis with updates and news about the pilot project. We also posted updates on social media, such as LinkedIn and Twitter. The audience reached with the project website was, on average, 145 unique users per month. In total, users were reached from Austria, USA, Czech Republic, China, Netherlands, Canada, Germany, United Arab Emirates, Switzerland and Croatia.

4 Conclusions and Results of the Pilot Project

The main technology achievements of our pilot project can be summarised as follows. We could successfully demonstrate that Docker containers can run on small devices such as Android smartphones and that applications such as EDDI and databases such as MongoDB can run within these containers. We could also show that peer-to-peer networks for communication between virtual assistants are possible with both a public and a private section of the accessible data and a handshake and identity check mechanisms to verify both users of the virtual assistants with key exchange and end-to-end encryption in order to achieve highest security standards. Based on the research work during the pilot project and the implemented prototype, we plan to develop the software further to a minimum viable product.

YouTwinDi is the next step in a digitised world in which the digital twin evolves and interacts with other digital twins and makes autonomous decisions in the interest of its human twin. In this scenario, security and digital ethics assure ethical decisions and IT specialists concur on improving the digital landscape with ethical models. This vision also includes overcoming language barriers. A continuous match of supply and demand as well as tailored searches help human twins to improve their lives in all respects. YouTwinDi uses the most advanced translation and language analysis technologies, allowing the user and its digital twin to interact with all European citizens without being blocked by language barriers.

We use our existing open source software EDDI which is running in a docker container for the natural language interface. This prototype is the basis for the development of a minimum viable product ready for market launch.

The ELG pilot project YouTwinDi had two major innovation aspects:

Technical innovation For the first time an AI application runs within a Docker container on a small hardware device without any technical limitations.

Creative-economical innovation The creative-economical innovation relates to the idea that the digital twin interacts with other digital twins and makes autonomous decisions in the interest of its human twin.

[4] https://youtwindi.com

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.