Nexus Linguarum

Balancing the digital presence of languages in and for technological development

A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud

Authors: *Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, Maxim Ionov, Jorge Gracia, Liudmila Rychkova, Giedre Valunaite Oleskeviciene, Christian Chiarcos, Thierry Declerck, Milan Dojchinovski*

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Balancing the digital presence of languages in and for technological development

A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud

Authors: *Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, Maxim Ionov, Jorge Gracia, Liudmila Rychkova, Giedre Valunaite Oleskeviciene, Christian Chiarcos, Thierry Declerck, Milan Dojchinovski*

THE PROBLEM

In the digital single market, the role of technology in overcoming the current language barriers to make Web content available for users independently of their language is beyond doubt. Language technology development, in turn, hinges greatly on the existence of **language resources**[1], including lexica, corpora, databases, etc. However, there are **dramatic differences in the availability of such resources across languages**, with only a few of them, such as English, featuring prominently in the landscape of discoverable and reusable data, while others, such as Greek or Serbian, with scarce technological support.

WHAT ARE UNDER-RESOURCED LANGUAGES?

"[Under-resourced languages are] languages that have a small or economically disadvantaged user base, that are therefore typically ignored by the commercial world and that are technologically underdeveloped due to limited human, financial and linguistic/language resources (LRs)[2]"

- In the CLARIN Virtual Language Observatory[3], 99% of all the textual data by volume (i.e., number of tokens) is in English, German, and Dutch languages. The rest is divided between other languages including Italian, Polish, Spanish and some truly low-resourced languages like Frisian, Selkup, and Dolgan. Calculated by the number of resources, English, German, and Dutch add up to 58.7%, which is not as extreme as when calculated by volume but still quite a high number, especially given that this does not include many major European languages. Technological support varies drastically across languages: according to the META-NET white papers, only English has good support for different language technologies, with languages like French and Spanish having only partial support[4]. There are different factors that contribute to a language being under-resourced: social, political, and technological.

TRULY MULTILINGUAL EUROPE?

TECHNOLOGICAL BARRIERS

- **Compromised universal access to information.** Current technological advancements based on deep learning largely depend on the availability of large amounts of data, which deepens the gap between 'resourced' and 'under-resourced' languages. Failure to provide language technology support will also directly affect the universal access to information of speakers of under-resourced languages[5].
- **Vicious cycle of unequal technological support.** Adapting current systems to other languages might fail to capture their linguistic richness in comparison to tools supporting them from scratch, which leads to a negative impact on the quality of results and a wider gap still.
- **Isolated development.** Even when some resources do exist, they were developed and are used in isolation from others, thus hindering their enrichment and integration with external data, and their visibility.

CULTURAL AND SOCIO-ECONOMIC BARRIERS

- **Linguistic identity discrimination.** With language being an important factor in the identity of communities, technology should be able to support them in using their own languages in a variety of contexts, and thus help reduce socio-economic inequalities associated with communication deficiencies. To this end, the availability of language resources in their own languages is essential.

---

1 *"encompassing (a) data sets (textual, multimodal/multimedia and lexical data, [...]) in machine readable form, and (b) tools/technologies/services used for their processing and management"* (Meta-Share Ontology, http://w3id.org/meta-share/meta-share/).

2 Pretorius, L. (2014). The multilingual semantic web as virtual knowledge commons: The case of the under-resourced South African languages. In Towards the multilingual semantic web (pp. 49-66). Springer, Berlin, Heidelberg.

3 https://vlo.clarin.eu

4 http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

5 Soria, C., & Russo, I. (2018). The Digital Language Diversity Project. In Technologies for Minority Languages Symposium: University of Groningen-Campus Fryslân. Ljouwert, the Netherlands.

- **Culture loss.** Linguistic discrimination poses barriers for cultural heritage maintenance. In addition, the low number of resources for historical languages encumbers the application of the latest advancements in technology to facilitate their analysis and preservation.
- **Education.** The scarce number of high quality and up-to-date teaching materials and educational resources, e.g. learner corpora, reference grammars, etc., negatively impacts integration of these languages into the global landscape.

CURRENT EFFORTS

- Considering this imbalance in the technological support for different languages, elaborated upon in the European Parliament Resolution "Language equality in the digital age[6]", there has been a number of initiatives that aim at the promotion of languages that are often under-resourced (e.g. ELEN[7]), as well as projects and events that are concerned with their inclusion in language technology development (ELE[8], LITHME[9], ELG[10], LT4All[11], ELRC[12], among others). In addition, a remarkable effort has been made in collecting rich datasets aimed at depicting linguistic diversity (e.g., ANU Database, AUTOTYP, STEDT, PHOIBLE)[13] among others such as Apertium[14] and the OpenMultilingualWordnet[15].

- However, even though such initiatives and datasets cover many under-resourced languages, the resulting data remain in project-specific formats, leading to insufficient data access, possibilities for sharing, and integration for query and comparison[16]. In order to address this scenario, there is a compelling need to focus on the **interoperability** of resources and tools with under-resourced languages in the spotlight.

THE ROLE OF (LINGUISTIC) LINKED DATA

- Linked data (LD) refers to a series of best practices and principles for "exposing, sharing, and connecting data on the Web", by identifying "things" on the Web with Unique Resource Identifiers (URIs) and using other standards to define useful information about them and links to other "things", thus enabling interoperability across datasets and systems. Such "things" or resources can be literally anything, therefore also linguistic information (for instance: words, translations, etc.), leading to the cloud of Linguistic Linked Data (LLD)[17].

Some of the advantages of exposing and sharing under-resourced languages data as LLD are the following:

- Sustaining the **technological development** of these languages.
- Preserving **cultural diversity** and indigenous knowledge systems, since their linguistic data is not isolated anymore but identified and exposed at a Web scale.
- Increased **discoverability** of their language resources through centralised repositories (e.g., LingHub).
- Language comparison and information integration through **conceptual interoperability**.
- Resources can be easily **linked** to other resources from other languages (under-resourced or not), thus the language data is put in a broader context and multilingual resources including under-resourced languages can be more easily built.

6 European Parliament (2018), P8_TA-PROV(2018)0332, "Language equality in the digital age", 11th of September 2018, https://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.pdf.

7 The European Language Equality Network, https://elen.ngo

8 European Language Equality, https://european-language-equality.eu/

9 COST Action CA19102 "Language In The Human-Machine Era", https://lithme.eu/

10 European Language Grid, https://www.european-language-grid.eu/

11 International Conference Language Technologies for All, 4-6 December 2019, Paris, France. https://en.unesco.org/LT4All

12 European Language Resource Coordination, https://www.lr-coordination.eu/

13 Moran, S., & Chiarcos, C. (2020). Linguistic Linked Open Data and Under-Resourced Languages: From Collection to Application. Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences, 39.

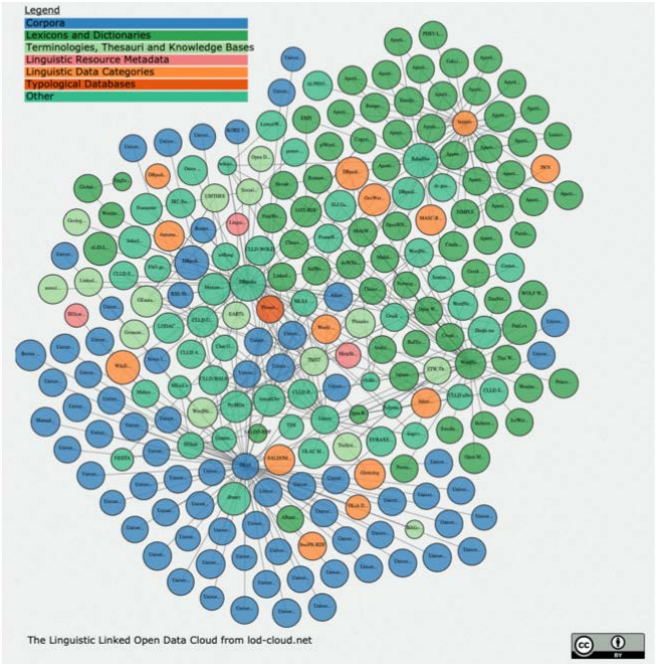
14 https://www.apertium.org/

15 http://compling.hss.ntu.edu.sg/omw/

16 Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts (pp. 205-227). IGI global.

17 Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). Linguistic linked open data cloud. In Linguistic Linked Data (pp. 29-41). Springer, Cham.

- Increased **portability** of LRs and Natural Language Processing (NLP) tools across closely related languages.
- Data published by following the **FAIR principles** (Findability, Accessibility, Interoperability, and Reusability of digital assets[18]).
- **Explainability** and interpretable reasoning — even in the absence of large amounts of data.



JOIN US IN THE EFFORT

There are different ways to get involved in the community to make this vision a reality: Joining activities in NexusLinguarum and related W3C Community Groups[19], or participating in their events, would be a starting point.

As a **researcher or developer**, you could...

- Develop tools and technologies that support under-resourced languages.
- Promote the inclusion of under-resourced languages in the LLOD cloud.
- Reach out to domain experts to access and convert their data to LL(O)D.

As a **language data provider**, you could…

- Participate in LOD training schools and provide data to be included in the LLOD cloud.
- Follow LD principles when publishing your data.
- Increase the visibility of under-resourced language data by linking it with existing language resources in the LLOD cloud.

As a **policy maker** or a **funding agency**, you could…

- Recommend LD as a strategic means to adhere to FAIR Principles and encourage publishing datasets in compliance with it.
- Promote and fund projects addressing under-resourced languages.
- Invest in the development and hosting of resources for these languages.

Acknowledgements: We would like to thank Penny Labropoulou, Sara Carvalho, and Fahad Khan for their careful review and valuable suggestions.

See more policy briefs at:
https://nexuslinguarum.eu/ results/policy-briefs

Date of 1st publication (v1): November 2021
Date of this version (v2): October 2022

Cite as: J. Bosque-Gil, V. Mititelu, H. Gonçalo-Oliveira, M. Ionov, J. Gracia, L. Rychkova, G. Valunaite-Oleskeviciene, C. Chiarcos, T. Declerck, and M. Dojchinovski. ***"Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud".***
DOI: 10.5281/zenodo.7142513.
NexusLinguarum CA18209 Cost Action. 2022

[18] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9. https://www.nature.com/articles/sdata201618

[19] Ontology-Lexica, https://www.w3.org/community/ontolex/; Linked Data for Language Technology, https://www.w3.org/community/ld4lt/, NexusLinguarum, https://nexuslinguarum.eu/ .