Quasi-Parallel Corpora for Less-Resourced Languages: Parallelized Translations of Plato´s *Faidon* in Basque and Finnish.

Koldo Garai
UEF, Finnland
kgarai@uef.fi

EHU, Basqueland
koldo.garai@ehu.eus

Paper Abstract

As at the time Director-General of UNESCO Irina Bokova put it, "Language loss entails an impoverishment of humanity in countless ways. Each language – large or small – captures and organizes reality in a distinctive manner; to lose even one closes off potential discoveries about human cognition and the mind" (Bokova, Irina, 2010). The Foreword by Jordi Solé (Rehm & Way, 2023) also reflects upon languages considered not only as pure communication tools or even as vectors of culture but also as factors of identity; multilingualism is an expression of the identity of Europe.

There are 24 EU official languages, 11 additional official languages, and 54 Regional and Minority Languages (RML), protected by the European Charter for Regional or Minority Languages (ECERML) since 1992, and the Charter of Fundamental Rights of the EU. Some are Indo-European languages and some are non-Indo-European, but both share the task of defining the identity of Europe.

Out of these 90 European languages, more than half have either poor or no technological support; for instance, consider Figure 1 comparing Finnish and Spanish, and keeping in mind that Spanish has half of the technological resources of English (only European context, in both cases), and compare this against Figure 2, technological resources for Finnish, Basque, and Karelian languages (created ad hoc from the web page: *Atlas of the World's Languages in Danger - UNESCO Digital Library*, n.d.).

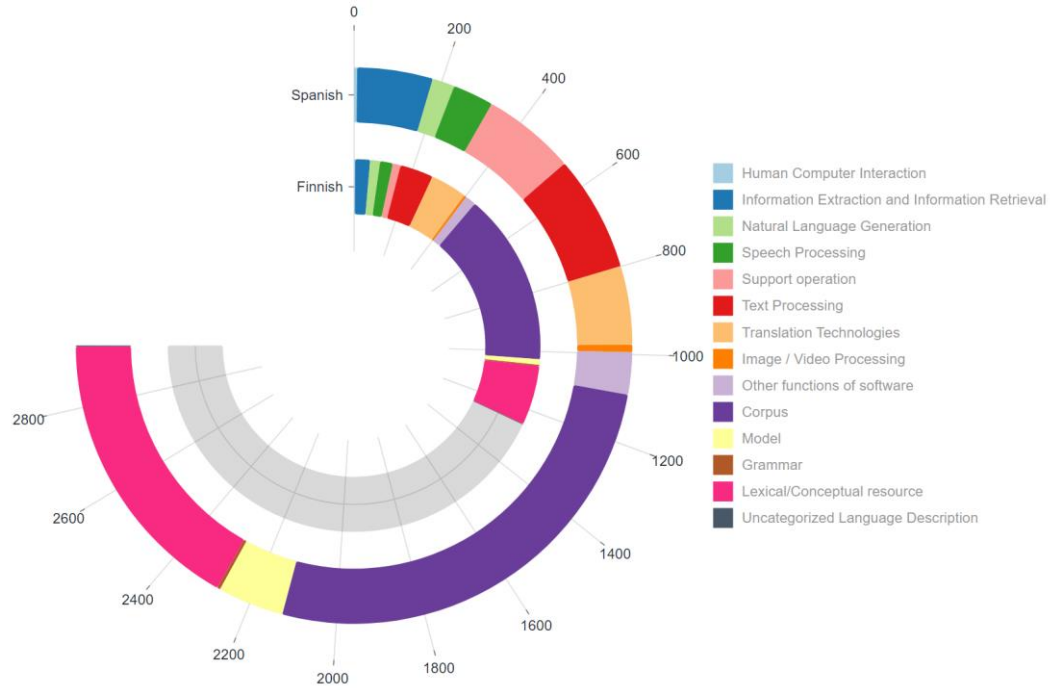

Figure: 1. Technological Factors in Spanish and Finnish

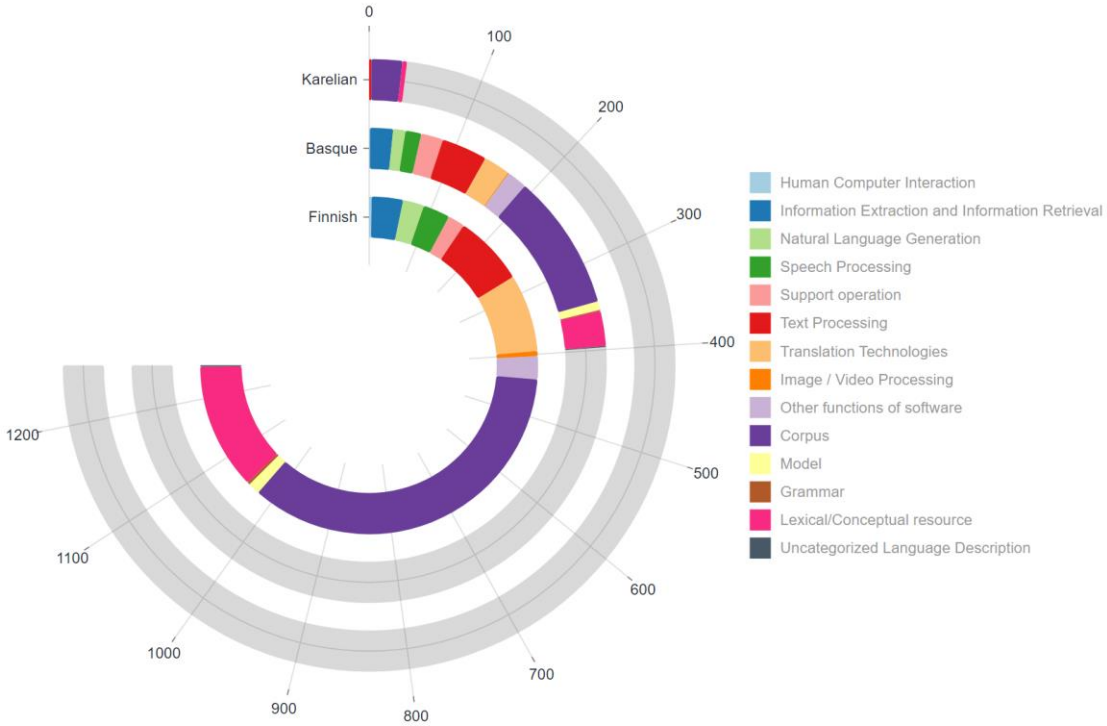

Figure: 2. Technological factors in Finnish, Basque, and Karelian languages

In the spirit of ELE, we present the first aligned Basque-Finnish corpus, both non-Indo-European languages. On the one hand, it is a finished project with the four steps for building a text-aligned corpus, and the description of the procedure can be used as a best practices manual for further prospects (Garai, 2024). On the other hand, it could be seen as a forerunner of a larger desideratum project of building a multilingual aligned corpus comprising all the European non-Indo-European languages to be used for both contrastive linguistic studies and a testbed for shared strategies and approaches to Language Technologies, given some typological convergences such as their postpositional nature or their rich morphology.

Whilst comparable corpora are made of comparable texts following some given criteria, be they from the same language or different languages, parallel refers to translations of a given text (McEnery & Xiao, 2018). Rather, here we coin the term "quasi-parallel" because one is not the direct translation of the other, but both are translations of the same omega text; in this case, Plato's Faidon (Plato, 2006 & Plato, 1978), one translated by Calamnius and the other by Zaitegi. Using already extant translations, and parallelizing them is the cheap path we are proposing for creating linguistic technologies for less-resourced languages.

As a finished study, this work travels through all four stages of building a corpus: (a) from printed text to machine-readable, (b) the standardization of the Basque text to erase graphemic idiosyncrasies to facilitate the next two steps, (c) the alignment, and (d) the automatic annotation following the Universal Dependencies (de Marneffe et al., 2021). The access to the actual outcomes will be shortly available in a repository to be announced.

Keywords: Annotation Universal Dependencies, Less-resourced languages, Parallel corpora, Plato: Finnish-Basque, Text alignment

REFERENCES

Atlas of the world's languages in danger—UNESCO Digital Library. (n.d.). Retrieved May 3, 2022, from https://unesdoc.unesco.org/ark:/48223/pf0000187026

Bokova, Irina. (2010). Preface. In C. Moseley, A. Nicolas, & Unesco (Eds.), *Atlas of the world's languages in danger* (3rd ed., entirely rev., enlarged and updated, p. 4). Unesco Paris.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Garai, K. (2024). *Quasi-Parallel Corpora for Less-Resourced Languages: Parallelized Translations of Plato´s Faidon in Basque and Finnish*. (2024009) [Master's thesis, UEF]. https://erepo.uef.fi/handle/123456789/31151

McEnery, T., & Xiao, R. (2018). *Parallel and Comparable Corpora: What is Happening?* (pp. 18–31). Multilingual Matters. https://doi.org/10.21832/9781853599873-005

Plato. (1978). *Platon. IV., Kriton eta Faidon* (I. Zaitegi Plazaola, Trans.). Euskaltzaindia.

Plato, 427? BCE-347? BCE. (2006). *Faidoni Platonin keskustelma Sokrateen viimeisistä hetkistä jasielun kuolemattomuudesta* (J. W. (Johan W. Calamnius, Trans.). https://www.gutenberg.org/ebooks/19210

Rehm, G., & Way, A. (Eds.). (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7