Cuadernos Europeos de Deusto

Núm. Especial 04 (Julio 2022)

DOI: https://doi.org/10.18543/ced042022

ESTUDOS

Natural Language Processing and Language Technologies for the Basque Language

Procesamiento del lenguaje natural y tecnologías del lenguaje para el euskera

Itziar Gonzalez-Dios, Begoña Altuna

doi: https://doi.org/10.18543/ced.2477

Recibido el 5 de abril de 2022 • Aceptado el 20 de mayo de 2022 • Publicado en línea: julio de 2022

Derechos de autoría (©)

Los derechos de autor (para la distribución, comunicación pública, reproducción e inclusión en bases de datos de indexación y repositorios institucionales) de esta publicación (*Cuadernos Europeos de Deusto, CED*) pertenecen a la editorial Universidad de Deusto. El acceso al contenido digital de cualquier número de *Cuadernos Europeos de Deusto* es gratuito inmediatamente después de su publicación. Los trabajos podrán leerse, descargarse, copiar y difundir en cualquier medio sin fines comerciales y según lo previsto por la ley; sin la previa autorización de la Editorial (Universidad de Deusto) o el autor. Así mismo, los trabajos editados en CED pueden ser publicados con posterioridad en otros medios o revistas, siempre que el autor indique con claridad y en la primera nota a pie de página que el trabajo se publicó por primera vez en CED, con indicación del número, año, páginas y DOI (si procede). Cualquier otro uso de su contenido en cualquier medio o formato, ahora conocido o desarrollado en el futuro, requiere el permiso previo por escrito del titular de los derechos de autor.

Copyright (©)

Copyright (for distribution, public communication, reproduction and inclusion in indexation databases and institutional repositories) of this publication (*Cuadernos Europeos de Deusto, CED*) belongs to the publisher University of Deusto. Access to the digital content of any Issue of *Cuadernos Europeos de Deusto* is free upon its publication. The content can be read, downloaded, copied, and distributed freely in any medium only for non-commercial purposes and in accordance with any applicable copyright legislation, without prior permission from the copyright holder (University of Deusto) or the author. Thus, the content of CED can be subsequently published in other media or journals, as long as the author clearly indicates in the first footnote that the work was published in CED for the first time, indicating the Issue number, year, pages, and DOI (if applicable). Any other use of its content in any medium or format, now known or developed in the future, requires prior written permission of the copyright holder.

Natural Language Processing and Language Technologies for the Basque Language

Procesamiento del lenguaje natural y tecnologías del lenguaje para el euskera

Itziar Gonzalez-Dios
HiTZ Basque Center for Language Technologies-Ixa NLP Group,
University of the Basque Country UPV/EHU
itziar.gonzalezd@ehu.eus

Begoña Altuna
HiTZ Basque Center for Language Technologies-Ixa NLP Group,
University of the Basque Country UPV/EHU
begona.altuna@ehu.eus

doi: https://doi.org/10.18543/ced.2477

Recibido el 5 de abril de 2022
Aceptado el 20 de mayo de 2022
Publicado en línea: julio de 2022

Summary: I. Introduction.—II. Current Scenario. 1. Language diversity in Europe 2. Basque Language. 3. Presence of Basque in the digital world. 4. Natural language and speech processing.—III. End-user NLP products for Basque.—IV. NLP technologies behind the products. 1. Lexico-semantic resources. 2. Tools. 3. Corpora.—V. Discussion: initiatives and lessons learnt.—VI. Conclusion.

Abstract: The presence of a language in the digital domain is crucial for its survival, as online communication and digital language resources have become the standard in the last decades and will gain more importance in the coming years. In order to develop advanced systems that are considered the basics for an efficient digital communication (e.g. machine translation systems, text-to-speech and speech-to-text converters and digital assistants), it is necessary to digitalise linguistic resources and create tools. In the case of Basque, scholars have studied the creation of digital linguistic resources and the tools that allow the development of those systems for the last forty years. In this paper, we present an overview of the natural language processing and language technology resources developed for Basque, their impact in the process of making Basque a "digital language" and the applications and challenges in multilingual communication. More precisely, we present the well-known products for Basque, the basic tools and the resources that are behind the products we use every day. Likewise, we would like that this survey serves as a guide for other minority languages that are making their way to digitalisation.

Keywords: Basque, natural language processing, language technologies, and language digitalisation.

Resumen: *Que una lengua tenga presencia en el ámbito digital es hoy en día crucial para su supervivencia, ya que en las últimas décadas la comunicación en línea y los recursos lingüísticos digitales se han convertido en parte de nuestra vida cotidiana y se utilizarán más en los próximos años. Para desarrollar sistemas que se consideran necesarios para una comunicación digital eficiente (por ejemplo, sistemas de traducción automática, conversores de texto a voz y de voz a texto o asistentes digitales) es necesario digitalizar recursos lingüísticos y crear herramientas adecuadas. En el caso del euskera, desarrollar estos recursos y sistemas ha tenido un interés primordial entre los académicos durante los últimos cuarenta años. En este artículo, presentamos una visión general de los recursos de procesamiento del lenguaje natural y de tecnología lingüística que se han desarrollado para el euskera, su impacto en el proceso de hacer del euskera una «lengua digital» y las aplicaciones y retos en los escenarios de comunicación multilingüe. En concreto, presentamos los productos más conocidos y las herramientas y recursos básicos que los soportan. Asimismo, queremos que este estudio sirva de guía a otras lenguas minoritarias que están realizando su camino a la digitalización.*

Palabras clave: *Euskara, procesamiento de lenguaje natural, tecnologías del lenguaje, y digitalización de lenguas.*

I. **Introduction**

In recent years in our Information and Communication Technology (ICT) society, interactions between people, and more recently also between humans and machines, are more frequently held in the digital domain. People have fast gotten used to interacting with other people and reaching resources online, but to achieve the most satisfactory communication results, all the participants need to be able to express their queries, messages and requests in the most comfortable and accurate way, i.e. in a language they master.

Nowadays, English is by far the predominant language in the digital world, followed closely by languages such as Spanish and French, and a certain knowledge of these languages is required in order to be able to efficiently navigate the digital resources. People, however, might not be proficient in those languages and also own the right of speaking and of being understood in the language of their choice. Hence, if more successful communication is to be achieved, languages other than English need to be working languages in the digital field too. There is no limit to the creation of digital resources in any language, but resource development requires large and language-guided efforts, which need to be backed by proportional investments. To attain this, not only the speaker community, the target audience, needs to show interest, but also all the stakeholders (governments, researchers and developers) must take part.

In this paper, we analyse the Language Technologies (LT) used in the process of making Basque a fully functional language for the digital domain and in the revitalisation and preservation of the language. Basque, a minority language of Europe surrounded by largely spoken languages, does not get the deserved attention due to the fact that it is not a national language, that it is not even official in all the areas in which it is spoken, and because most of its speakers are bilingual or plurilingual. This also gets its reflection in the digital domain as the lack of available resources in Basque make the speakers switch into other languages[1]. This overview of the efforts of the digitalisation of Basque may serve as a reminder of the work done towards revitalisation and preservation of the language, as the basis for setting the next steps and as inspiration for the work for other minority and less-resourced languages that are struggling to find their place in the digital era.

[1] Antton Gurrutxaga and Klara Ceberio, "Basque-a Digital Language?", in *Reports on Digital Language Diversity in Europe,* ed. Claudia Soria, Irene Russo, and Valeria Quoqui (The Digital Language Diversity Project, 2017), http://www.dldp.eu/sites/default/files/documents/DLDP_Basque-Report.pdf

II. **Current Scenario**

In this section, we present the scenario in which the resources and tools for Basque have been developed. More precisely, we focus on the situation of the Basque language in the local and in the international context, as well as on the technologies that have allowed the development of the digital tools and resources used in the process of revitalisation and standardisation of Basque.

1. *Language diversity in Europe*

Europe is a land of 10,180,000 km^2 that is home to circa 745 million people. The population is divided into 87 ethnic groups, of which 33 are part the majority population in at least one of the 50 sovereign states, while the remaining 54 are considered ethnic minorities[2]. Europeans speak circa 225 indigenous languages[3]. 48 of them are official languages and the rest are considered minority languages, although some of the last —such is the case of Basque— have been awarded a certain status that offers them protection and confers to the speakers certain language rights, normally circumscribed to the region in which the minority language is spoken.

If the focus is centred on the countries that form the European Union, it is to point out that 24 languages hold the status of official languages of the Union, but other 60 languages are part of the heritage of the different EU countries. Adding the languages spoken by migrants, around 175 languages are spoken in the area nowadays. Therefore, the European Union has largely been determined to preserve the language rights of its citizens as well as to protect the European languages. In Article 22 of the European Charter of Fundamental Rights,[4] it is stated that the EU respects linguistic diversity and in Article 3 of the Treaty of European Union,[5] it is declared that the EU should ensure that Europe's cultural heritage is safeguarded and enhanced. This commitment to the preservation of the languages in Europe has been put into practice through a series of

[2] Christoph Pan and Beate Sibylle Pfeil, *Minderheitenrechte in Europa: Handbuch der europäischen Volksgruppen, Band 2* (Wien: Braumüller, 2002).

[3] https://edl.ecml.at/Facts/LanguageFacts/tabid/1859/language/Default.aspx

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

[5] https://eur-lex.europa.eu/resource.html?uri=cellar:2bf140bf-a3f8-4ab2-b506-fd71826e6da6.0023.02/DOC_1&format=PDF

initiatives such as The Digital Language Diversity Project[6] in the digital domain, or Horizon Europe call "Safeguarding endangered languages in Europe"[7]. The interested reader is referred to works by Ferreira and Bouda[8] and Olko and Sallabank[9] for more related work about language revitalization and conservation works.

Basque is a co-official language of Spain and a regional language in France (See the following "Basque language" section) and, thus, is a heritage language of two countries of the European Union and benefits from the protection and enhancement initiatives of the EU. Unfortunately, this is not the case of all European languages.

2. *Basque language*

Basque is a non-Indo-European language spoken in the area of the Bay of Biscay in the autonomous community of the Basque Country and the chartered community of Navarre in Spain and in the Basque Municipal Community in the department of the Atlantic Pyrenees in France. These three territories host a population of around three million people of which over one million can speak Basque and around 460,000 can understand it (Table 1).

The official status of Basque varies in the three administrative circumscriptions. Basque is official alongside Spanish in the three provinces of Basque Autonomous Community (Alava, Biscay and Gipuzkoa) since 1982 and official in some parts of Navarre since 1986. In Navarre, some municipalities have been granted access to the Basque speaking area in recent years, increasing the number of towns in which rights for Basque speakers are observed, but that is not the case in most of the central and southern areas of Navarre. In the case of the Northern Basque Country, it is not an official language in any of the three Northern provinces (Labourd, Low Navarre and Soule), although it has had a declaration of officiality by the Basque Municipal Community since 2018.

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

[7] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl2-2022-heritage-01-01

[8] Vera Ferreira and Peter Bouda, *Language Documentation and Conservation in Europe (whole volume)* (Honolulu: University of Hawai'i Press, 2016), https://nflrc.hawaii.edu/ldc/sp09-language-documentation-conservation-europe/

[9] Justyna Olko and Julia Sallabank, eds,, *Revitalizing Endangered Languages: A practical guide* (Cambridge: Cambridge University Press, 2021), https://doi.org/10.1017/9781108641142

Table 1

Population of the Basque Country and speakers of Basque

Region	Total population	Basque Speakers	Understand Basque
Autonomous Community of the Basque Country	2.220.504 (2020)	895,942 (2016)*	391,897 (2016)
Chartered community of Navarre	647.554 (2019)	75,810 (2018)**, ***	42,994 (2018)
Basque Municipal Community	315.349 (2019)	51,000 (2016)****	23,000 (2016)

* https://eu.eustat.eus/elementos/ele0014600/41euskal-aeko-2-urteko-eta-gehiagoko-biztanleria-bizi-zonari-eta-euskara-maila-globalari-jarraiki-lurralde-historiko-eta-sexuaren-arabera/tbl0014684_e.html

** https://gobiernoabierto.navarra.es/sites/default/files/azterketa_soziolinguistikoa_2018.pdf

*** These numbers have been calculated on the population who are over 16 years of age.

**** https://www.mintzaira.fr/fileadmin/documents/Enquete_sociolinguistique/Sintesia_2016_euskaraz.pdf

Despite a handful of historical attempts of writing for a larger public, the systematic standardisation of the language is a rather recent process. The standardisation, promoted by Euskaltzaindia[10], the Royal Academy of the Basque language, started officially in 1968 and it is now in a consolidation phase. The extension of Basque education and the need for acquiring Basque language certificates along with the multiplication of available media and resources in the standard variety have boosted the spreading of the new conventions. Nonetheless, even if orthography and many morphological and syntactic rules have been coded by 2000, non-standard variants are still used in colloquial daily communication. One should not either forget the prodigious effort on the creation of specialised terminology, which has allowed for the use of Basque in high level teaching, industry, medicine and law making among others.

[10] https://www.euskaltzaindia.eus/en/

3. *Presence of Basque in the digital world*

Around 90% of the Basque population has access to digital resources through the internet connection[11,12]. However, the use of Basque on the internet is rather small. Less than 0,1% of the websites that are analysed by the World Wide Web Technology Surveys[13] use Basque. Moreover, as of 2019[14], only 42% of the Basque speakers visit websites in Basque or interact in social media in Basque. To expand these initial insights, we have summarised three surveys that have been carried out to assess the status of Basque as a digital language in the last ten years.

Basque language is high-risk according to the META-NET White Paper Series[15], where the language technology support of 30 European languages was assessed. Although Basque has a number of products, technologies and resources, tools for speech synthesis, speech recognition, spelling correction, and grammar checking and applications for automatic translation, the conclusion of this report is that Basque still needs research in order to ensure that language technology solutions are truly effective and ready for everyday use. Nevertheless, the authors of the study indicate that the case of Basque can be considered with cautious optimism, since there is a viable research community and the language technology industry is well established.

A recent report on the situation of Basque in the digital era[16] shows that the situation has been stable for the last decade. As of December 2021, there are 317 resources available for Basque in the ELG and ELE collections combined: 145 corpora, 10 language models, 1 grammar, 40 lexical or conceptual resources and 121 tools or services. These and other resources that are not indexed in those collections show that Basque possesses state-of-the-art technology and robust, broad-coverage natural language processing (NLP) resources. Nonetheless, although the available resources are top-notch,

[11] https://eu.eustat.eus/elementos/ele0014600/41euskal-aeko-2-urteko-eta-gehiagoko-biztanleria-bizi-zonari-eta-euskara-maila-globalari-jarraiki-lurralde-historiko-eta-sexuaren-arabera/tbl0014684_e.html

[12] https://administracionelectronica.navarra.es/GN.InstitutoEstadistica.Web/DescargaFichero.aspx?Fichero=\web\informes\ciencia_tecnologia\esi\eticce_20_21_hogares_nastat.pdf

[13] https://w3techs.com/technologies/details/cl-eu-

[14] Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, *Euskal Herriko parte-hartze kulturalari buruzko inkesta 2019. Emaitzen txostena*. (Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, 2019), https://www.euskadi.eus/contenidos/informacion/keb_argit_ohiturak_eh_2018/eu_def/adjuntos/estatistika-parte-hartze-kulturala-eh-2019.pdf

[15] Imna Hernáez *et al.*, *Euskara Aro Digitalean-The Basque Language in the Digital Age* (Springer, 2012), https://link.springer.com/book/10.1007/978-3-642-30796-6

[16] Kepa Sarasola *et al.*, *D1.4 Report on the Basque Language* (ELE Consortium, 2022), https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_4__Language_Report_Basque_.pdf

Language Technologies for Basque still need work and commitment as they can play a crucial role in the revitalisation and preservation of the language.

Regarding the usage of Basque in the web, a study carried out by the Digital Language Diversity Project in 2016 revealed that Basque is a digitally fit language and that it is actively used in comparison to the other languages studied in the project (Breton, Karelian and Sardinian). However, there is a lack of entertainment products in Basque and finding and using tools in Spanish is easier[17] for Basque speakers.

4. *Natural Language and Speech Processing*

Many of the resources contributing to the use of a certain language in the digital domain (e.g. machine translation, speech recognition, or question answering systems) benefit from, or are built on, different Natural Language Processing and Speech Processing efforts that provide the necessary automatic information extraction and analysis that is then used to create the final product for the users.

Natural Language Processing (NLP) is a multidisciplinary field that studies how computers can process natural language and texts, in other words, how to teach computers to *learn* and *understand* language. Moreover, NLP is closely related to computational linguistics, which is also concerned with study of language from a computational perspective: modelling language or answering linguistic questions based on computational approaches. Language analysis and processing have developed side by side since the beginning as NLP has allowed for the mass analysis of language and that analysis has offered the linguistic generalisations needed for language processing.

NLP started back in 1950 and since them it has had three generations (Figure 1): i) symbolic NLP, where hand-written rule-based systems were created; ii) statistical NLP, where machine learning algorithms were introduced due to the increasing computing power and increasing data; and, iii) neural NLP, where deep-learning-based machine learning techniques are used benefiting from the vast amounts of data available and high computation power. Basque Language has had approaches in all of them. In NLP most of the work is carried out for English, and later, techniques are applied for other languages, but these languages are usually limited to Western European Languages (particularly German, French, and Spanish) and to a lesser extent to Chinese, Japanese, and Arabic[18].

[17] Gurrutxaga and Ceberio, "Basque-a Digital Language?"

[18] Damián Blasi, Antonios Anastasopoulos and Graham Neubig, "Systematic Inequalities in Language Technology Performance across the World's Languages", *arXiv preprint arXiv:2110.06733,* (2021), https://arxiv.org/pdf/2110.06733.pdf

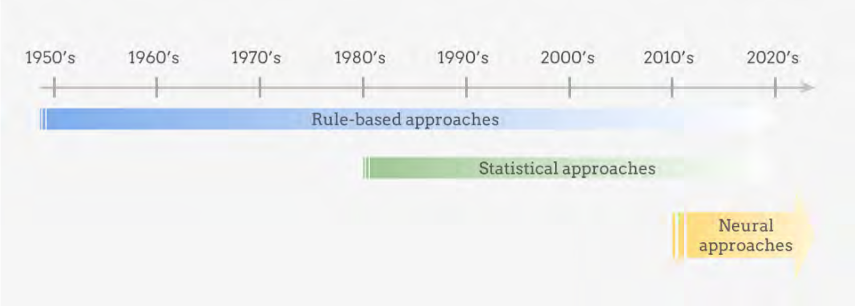

Figure 1

A representation of the NLP approaches in history

On the other hand, Speech Processing (SP) aims at processing human speech by computers and, to that end, digital signal processing techniques are applied to speech signals. Its main research areas are text to speech conversion, speech and speaker recognition, and speech synthesis. SP also dates back to the 1950's and early attempts focused on the recognition of simple phonetic elements. The use of Hidden Markov Models in the mid 80's brought a big improvement due to the power of this statistical approach, and nowadays mainly artificial neural networks are used.

NLP and SP are the core elements of Language centric Artificial Intelligence expected to be supporting multilingual Europe as an efficient way to break down the language barriers. To understand the importance of NLP and SP in the recent and the coming years, the interested reader is referred to the survey by Rehm *et al*.[19] for an overview of the activities, actions, funding programmes and challenges in different European Countries about Languages Technologies. For an overview of the language data, we suggest the ELRC White Paper[20].

[19] Georg Rehm *et al*., "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* (European Language Resources Association, 2020), 3322-3332, https://aclanthology.org/2020.lrec-1.407/

[20] European Language Resource Coordination, *ELRC WHITE PAPER. Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters* (ELRC Consortium, 2019), ISBN: 978-3-943853-05-6 https://lr-coordination.eu/sites/default/files/Reports%202021/ELRCWhitePaper.pdf

III. End user NLP products for Basque

Natural Language Processing and Speech Processing products have played a key role in the expansion of the usage of Basque as well as on its normalisation. Some daily-use linguistic products and resources strongly rely on NLP and SP and provide the necessary tools for consultation and efficient communication in Basque.

The first and most successful outcome of the research on NLP was the spell checker Xuxen[21]. Since the very first prototype, Xuxen has been adapted, both to include the new words accepted in Basque by Euskaltzaindia, and to be used in different softwares (text editors, web browsers and operating systems…) and their corresponding updates. That is, Xuxen is an evolving product that needs to be updated both at linguistic level and at engineering level. As in the Information and communications technology (ICT) society changes happen faster than a product can be successfully adapted to the latest version, Xuxen has overcome this problem by offering its services (plus additional features) in the web service http://xuxen.eus. The latest versions of Xuxen also include some grammatical corrections. Indeed, a grammatical checker called Xuxeng[22] was released, but due to the high cost of adapting to new versions and lack of funding, the project halted. However, due to the new advances in NLP, this project has been resumed[23] and Xuxen is nowadays maintained by the Elhuyar Foundation and Ixa group.

Xuxen has been and still is the most used spell checker for Basque and has undoubtedly contributed to the standardisation of Basque, since it has helped users with orthographic doubts and choosing among standard and non-standard words. As an example of its popularity, the Xuxen web service has made 1,382,557 corrections in the last two years (2020-2021) and the MSOffice, Libreoffice, Chrome and Firefox extensions are downloaded on average more than 3,300 times per year.

Digital and online lexical resources are also very popular digital applications, although they do not always rely on NLP technologies. One that does include them is the most popular general bilingual dictionary is Elhuyar Hiztegia[24], which includes a lemmatizer to help with the queries.

[21] Eneko Agirre *et al.*, "XUXEN: A Spelling Checker/Corrector for Basque based on Two-Level Morphology", in *proceedings of the third conference on applied natural language processing* (1992), 119-125, https://aclanthology.org/A92-1016.pdf

[22] http://ixa.si.ehu.es/node/7622

[23] Zuhaitz Beloki *et al.*, "Grammatical Error Correction for Basque through a Seq2seq Neural Architecture and Synthetic Examples", *Procesamiento del Lenguaje Natural*, 65 (2020): 13-20, http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6271

[24] https://hiztegiak.elhuyar.eus/

GalNet[25], on its part, is an application developed in Galicia that includes the semantic networks (wordnets, concept-based dictionaries) of the Iberian Peninsula official languages, English and Chinese. Nonetheless, the most remarkable product is EuskalBar[26]. EuskalBar is a Firefox[27] and Chrome[28] extension that makes it possible to query simultaneously many dictionaries and corpora. It includes all the Basque updated digital dictionaries (monolingual, bilingual), terminological databases as well as some corpora and it can be customised according to users' preferences. All these products can be consulted online, but most of them are not open source and, therefore, they cannot be embedded in other applications. The open source alternatives are The WikiHiztegia[29], the multilingual dictionary developed as a Wikimedia product.

Another popular NLP application is machine translation used for professional translation or in a more casual way for little translation needs while surfing the net. Machine translation has experienced an overwhelming improvement in the last years making it an essential tool when writing or editing texts, even if it is not perfect yet as it produces errors that are more or less frequent depending on the language pairs involved. Human post-editing of the translated texts is still necessary, although it has been proven that human translators tend to be much more efficient when relying on machine translation. This leads to the reduction of costs and times of the translation work, which opens the opportunity to easily multiply the resources available in different languages and translation to less-spoken languages should not be considered a waste of effort anymore. In the case of Basque, a list of high quality neural translators (Elia.eus[30], itzuli+[31], and batua.eus[32]) and speech interpreting systems (Interprest[33]) are available, allowing the Basque speakers to interact with the speakers that do not know the language: Basque content can be reached for non-speakers and Basque speakers can get content in Basque.

As language technologies evolve, more sophisticated products are created. One of these products is Talaia[34] and it is related to the monitoring

[25] https://play.google.com/store/apps/details?id=gal.sli.digalnet&hl=es_MX&gl=US

[26] https://github.com/euskalbar/euskalbar

[27] https://addons.mozilla.org/eu/firefox/addon/euskalbar/

[28] https://chrome.google.com/webstore/detail/euskalbar/jfemlmedfkiadfaihdcdkcoooleihhfn?hl=es

[29] https://eu.wiktionary.org/wiki/Azala

[30] https://elia.eus/itzultzailea

[31] https://www.euskadi.eus/itzuliplus/

[32] https://www.batua.eus/

[33] https://interprest.io/

[34] http://talaia.elhuyar.eus/

of social media and digital press through the analysis of the sentiment and polarity of the texts[35]. This was put into practice in the context of the European Capital Culture Donostia 2016 with the Behagunea[36], where the tweets about the event were classified as positive, neutral or negative. It was also used to track the elections to the Basque Parliament in 2016. Talaia is maintained by the Elhuyar foundation.

The products so far mentioned are mainly headed for non-specialised communication situations and registers, but there are also other products that have helped in the elaboration and the development of the specialised communication. As Zabala[37] points out, there has been an overlap between the codification and elaboration of specialised registers in Basque and sometimes the experts were using terms that were later rejected by Euskaltzaindia in the normative dictionary or in the official terminological database Euskalterm. In order to describe the real terminology, that is, terms used by the experts, there are two products: the corpus Garaterm[38,39] and the terminological database TZOS[40,41]. The Garaterm corpus contains texts written in Basque by the lectures of the University of the Basque Country from different specialisations, and TZOS stores the Basque terms together with their equivalents in Spanish, English and French. These products are based on a combination of NLP technologies together with human effort and should serve as a basis for the stabilisation of terms.

Finally, the speech technologies for Basque are to be mentioned. Although it might seem that automatic speech recognition and text-to-speech systems rely only on the ability of matching sounds and their transcriptions, NLP plays a crucial role in that process. For example, being able to predict which word follows improves the quality of closed captions

[35] Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri, "Real Time Monitoring of Social Media and Digital Press", *arXiv e-prints, arXiv-1810* (2018), https://arxiv.org/pdf/1810.00647.pdf

[36] http://behagune.elhuyar.eus/

[37] Igone Zabala, "The Elaboration of Basque in Academic and Professional Domain" *Linguistic Minorities in Europe Online*, eds. Miren Lourdes Oñederra and Iván Igartua, (De Gruyter, 2019) ISSN 2510-5361 https://www.degruyter.com/database/LME/entry/lme.9612443/html

[38] http://garaterm-corpusa.ixa.eus/

[39] Igone Zabala *et al.*, "GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako laningurunea", in *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begir*a, eds. Xabier Alberdi and Pello Salaburu (Bilbao: Publishing Service of the UPV/EHU, 2013), 98-114, https://www.ehu.eus/documents/2430735/2730483/LIBURUAehuei13.pdf

[40] https://tzos.ehu.eus/

[41] Xabier Arregi *et al.*, "TZOS: An On-line System for Terminology Service", in *Actualizaciones en Comunicación Social*. (Centro de Lingüística Aplicada, 2013), 400-404 http://ixa.si.ehu.eus/sites/default/files/dokumentuak/3988/actas_I_TZOS.pdf

in what refers to speech-to-text environments, while word sense disambiguation is a compulsory step to achieve natural intonation in text-to-speech tasks.

In respect to the contributions of speech technologies to the revitalisation process of Basque, the ones related to accessibility are the most interesting. In the case of automatic speech recognition subtitling initiatives, these have given access to information in Basque to both Basque learners and hearing-impaired Basque speakers. Aditu.eus[42] provides a tool for the automatic transcription of videos in Basque. The ultimate goal of text-to-speech applications, on the other hand, is twofold: it can be of great help to give access to contents written in Basque to visually-impaired people, but it can also achieve a more ambitious objective, that is, preserving the voice of the people who have lost it. AhoTTS[43,44] provides automatically generated locutions of texts in Basque and AhoMyTTS[45] is a voice synthesiser in which the voice of the own users is employed.

In addition to these products, there are many research prototypes that can be tested as a demo. With necessary funding, some of these prototypes can be in the near future applications which can be used everyday. Here we mention some promising prototypes:

— Maria chatbot: a chatbot that answers questions about any topic on Wikipedia in 3 languages: Basque, Spanish and English[46].
— Ihardetsi: a Question-Answering system[47] for the area of Science and Technology[48].
— MultiAzterTest: an open source NLP tool which analyses texts on more than 125 linguistic and stylistic features for English, Spanish and Basque[49] and assesses the complexity level (readability)[50].

[42] https://aditu.eus/

[43] https://aholab.ehu.eus/tts/

[44] Imna Hernaez *et al.*, “Description of the AHOTTS System for the Basque Language.”, in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesi*s, (2001), https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.9152&rep=rep1&type=pdf

[45] https://aholab.ehu.eus/ahomytts/

[46] http://ixa2.si.ehu.eus/convai/maria-bot/index.html

[47] http://ixa2.si.ehu.eus/IhardetsiWebDemo/IhardetsiBezeroa.jsp

[48] Olatz Ansa *et al.*, “Ihardetsi: a Basque Question Answering System at QA@ CLEF 2008”, in *Workshop of the Cross-Language Evaluation Forum for European Languages,* (Berlin, Heidelberg: Springer, 2008), 369-376, http://ixa.si.ehu.es/sites/default/files/dokumentuak/3765/pdf.pdf

[49] Kepa Bengoetxea and Itziar Gonzalez-Dios, “MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment”, *arXiv preprint arXiv:2109.04870*, (2021), https://arxiv.org/pdf/2109.04870.pdf

[50] http://ixa2.si.ehu.eus/aztertest/

IV. **NLP technologies behind the products**

The products mentioned in the previous section are the *end product* users receive, but behind these products there are many NLP resources and technologies that need to be developed until the close product can be commercialised. NLP technologies are usually divided into resources (grammars, lexico-semantic resources and corpora) and tools. The tools are normally used in a pipeline in which each module deals with a different linguistic feature (meaning, morphology, syntax, etc.) and the resources provide the linguistic knowledge needed in each step. In this section, we will deal with the most important and used in the processing of Basque.

1. *Lexico-semantic resources*

The first resource that was created for the processing of Basque was the lexical database *Euskararen Datu Base Lexikala* (EDBL)[51], which was created in 1992 and published in 1995. The main aim of EDBL was to give lexical support for the construction of a general morphological analyser and the spell checker Xuxen[52]. That is, EDBL is the linguistic input of Xuxen and all the linguistic updates that are included in Xuxen should be done by updating this database.

EDBL is organised into three axes: 1) standard versus non-standard entries, 2) dictionary entries according to their parts-of-speech and other entries (non-independent morphemes, irregularly inflected forms, inflections of verb auxiliaries) and 3) single-word entries versus multiword lexical units. Moreover, in the case of the one-word lexical units it is described how morphemes are linked to other morphemes in order to constitute a word form (in the morphotactics). Thanks to this feature of EDBL, inflected word forms can be generated (all the Basque word forms can be created based on this information) instead of storing all of them. Due to this design, the different inflected forms can be properly identified and

[51] Izaskun Aldezabal *et al.*, “EDBL: A General Lexical Basis for the Automatic Processing of Basque”, in *proceedings of the IRCS Workshop on linguistic databases* , (IRCS Workshop on linguistic databases, 2001). https://www.ixa.eus/sites/default/files/dokumentuak/3301/2001-IRCS.pdf

[52] Eneko Agirre *et al.*, “XUXEN: A Spelling Checker/Corrector for Basque based on Two-Level Morphology”, in *proceedings of the third conference on applied natural language processing* (1992), 119-125, https://aclanthology.org/A92-1016.pdf

processed and Xuxen is able to correct all the possible inflections of a word and its standardisation status.

In 2010, EDBL was populated with dictionaries from the Elhuyar foundation and the UZEI lexicographic centre and it was renamed to *Euskararen datu-base lexikala-Lexikoaren Behatokiaren datu-base lexikala* (EDBL-LBDBL). Since then, EDBL-LBDBL is updated when *Euskaltzaindia* releases a new version of the *Euskaltzaindiaren Hiztegia* dictionary[53], the normative dictionary. At the writing of this paper (February 2022[54]), EDBL-LBDBL has 126,342 entries, out of them 105,381 are lexical entries.

As we saw, EDBL-LBDBL contains lexical and morphological information, but when a word has more than one sense EDBL just contains one entry. In order to represent semantic information wordnets are used. Wordnets are semantic networks organised around the notion of *synset*. Synsets are sets of cognitive synonyms that represent a concept e.g. the words included in the synset {car, auto automobile, machine, motorcar} represent the concept ‘a motor vehicle with four wheels; usually propelled by an internal combustion engine’. Each synset is connected to other synsets via semantic relations such as hypernymy-hyponymy, meronymy or antonimy to mention a few. So, contrary to traditional dictionaries that are organised alphabetically, wordnets are organised at concept level and the concept can be considered as the *entry*. In the case of ambiguous words such as *bank*, it has as many synsets as concepts it represents (river bank, bench, economic institution…).

The first wordnet was created for English (Princeton WordNet) in 1985[55,56] , but since then, wordnets have been created for other languages. The Basque WordNet[57] contains nowadays more than 30,600 synsets and 50,700 words (mainly nouns and verbs) and it serves as the basis for word-sense disambiguation tool for Basque UKB[58]. The Basque Wordnet is

[53] https://www.euskaltzaindia.eus/index.php?option=com_hiztegianbilatu&Itemid=410

[54] Date of the retrieval, Feb 15, 2022.

[55] Miller, George A. VWordNet: A Lexical Database for English», *Communications of the ACM,* 38, No. 11 (1995): 39-41. https://dl.acm.org/doi/pdf/10.1145/219717.219748

[56] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database* (Cambridge, MA: MIT Press, 1998). https://ieeexplore.ieee.org/book/6267389

[57] Eli Pociello, Eneko Agirre, and Izaskun Aldezabal, “Methodology and Construction of the Basque WordNet”, *Language resources and evaluation,* 45, 2 (2011): 121-142. https://doi.org/10.1007/s10579-010-9131-y

[58] Eneko Agirre and Aitor Soroa, “Personalizing PageRank for Word Sense Disambiguation”, in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, (Association for Computational Linguistics, 2009), 33-41, https://aclanthology.org/E09-1005.pdf

integrated in the Multilingual Central Repository (MCR)[59,60], a framework that integrates all the open-licenced wordnets of the official languages in the Iberian peninsula and English. This way, Basque is connected to all the open wordnets, being part of one of the biggest multilingual resources[61]. The MCR is also one of the linguistic knowledge bases of the GalNet application.

EDBL-LBDBL and Basque WordNet are resources that need to be constantly updated as language evolves, so that the tools can use the most accurate lexico-semantic information, but updating them is costly. Moreover, it is not trivial to decide which words should be added as an entry or synset. Linguists are needed to determine the appropriate lexicalization and representation of the concepts[62].

2. *Tools*

Tools are used to process the texts since computers cannot understand language: for them written texts are but strings, lines of continuous characters. In classical NLP, different tools are used in order to add relevant information that can be processed by the computers. These tools are programmed taking into account rules written by linguists (rule-based), statistical information derived from corpora (data-driven) or a mix of both approaches (hybrid). In contemporanean NLP (since~2016) the tendency has changed and texts are processed based on tools that learn the so-called language models from large corpora.

In the following lines, we describe the main tasks in text processing in classical NLP and the tools that have been created to carry out that process in Basque. A graphic summary of the main tools is presented in Figure 2. When available, we provide the links to the demos of the tools.

[59] Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. “Multilingual Central Repository version 3.0.”, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC’12), (European Language Resources Association, 2012), https://aclanthology.org/L12-1128/

[60] Xabier Gomez Guinovart *et al.*, “Multilingual Central Repository: a Cross-lingual Framework for Developing Wordnets”, *arXiv preprint arXiv:2107.0033*3, (2021) https://arxiv.org/pdf/2107.00333.pdf

[61] http://compling.hss.ntu.edu.sg/omw/

[62] Izaskun Aldezabal *et al.*, “Basque e-lexicographic Resources: Linguistic Basis, Development, and Future Perspectives”, in *Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence*, (2018), https://ixa.si.ehu.es/sites/default/files/dokumentuak/12709/Elexis18_abstract_def.pdf

- Morphosyntactic analysis: after doing the initial tokenisation (separating the text string in tokens/words) and segmentation (segmentation of tokens in lexemes and morphemes), Morfeus[63] carries out the morphosyntactic analysis of the words (tokens) and determines all the possible Parts of Speech a word can have[64].
- Multi-word item identification: words that should be analysed together are recognised[65].
- Lemmatization and syntactic function identification: Eustagger[66] determines the lemma of the words and identifies the syntactic functions of the words[67].
- Named entity recognition and classification: Eihera recognises the named entities in a text and classifies them according to their type: person, location, organisation or other[68].
- Shallow parsing: syntactic functions are disambiguated and noun and verb chains (chunks) are identified by Ixati[69,70].
- Dependency parsing: deep parsing is carried out following the dependency grammar formalism and three parsers have been created for Basque: a linguistic rule-based approach (EDGK)[71],

[63] http://ixa2.si.ehu.eus/demo/analisianali.jsp

[64] Iñaki Alegria *et al.*, "Robustness and Customisation in an Analyser/lemmatiser for Basque", in *LREC-2002 Customizing knowledge in NLP applications workshop* (European Language Resources Association, 2002), 1-6, https://www.ixa.eus/sites/default/files/dokumentuak/3340/robust2.pdf

[65] Jose Mari Arriola *et al.*, "Reusing the CG-2 Grammar for Processing Basque Complex Postpositions" in *Actas del XXIX Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2013)*, (2013), 20-27, http://ixa.si.ehu.es/sites/default/files/dokumentuak/3197/Reusing%20the%20CG-2%20Grammar.pdf

[66] http://ixa2.si.ehu.eus/demo/analisimorf.jsp

[67] Itziar Aduriz el al., "Finite State Applications for Basque", in *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing* (Association for Computational Linguistics, 2003), 3-11, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.7468&rep=rep1&type=pdf

[68] Iñaki Alegria *et al.*, "Design and Development of a Named Entity Recognizer for an Agglutinative Language", in *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*, (Berlin, Heidelberg: Springer, 2004), http://ixa.si.ehu.es/sites/default/files/dokumentuak/3794/IJCNLP04.pdf

[69] http://ixa2.si.ehu.eus/demo/zatiak.jsp

[70] Itziar Aduriz *et al.*, "A Cascaded Syntactic Analyser for Basque", in *International Conference on Intelligent Text Processing and Computational Linguistics*, Berlin, Heidelberg: Springer, 2004), 124-134, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.6003&rep=rep1&type=pdf

[71] María Jesús Aranzabe, "Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala" (Doctoral dissertation, University of the Basque Country (UPV/EHU), 2008). https://dialnet.unirioja.es/servlet/tesis?codigo=177705

and statistical approach[72] (Maltixa[73]) and a hybrid system (ASKHi)[74].

- Sentence and clause boundary detection: sentences and clause boundaries are detected[75].
- Apposition detection: appositions are detected and classified according to their type[76].
- Word sense disambiguation: UKB[77] assigns the most probable sense of a word based on the senses listed in Basque WordNet[78].
- Semantic role labelling: bRol labels the arguments with the corresponding semantic role[79].
- Coreference resolution: EUSKOR identifies textual expressions and determines which of them refer to the same entity[80].

Some of these tools, however, due to the period that they were created, rely on proprietary software and formalism. This implies that the user needs to have the adequate licences for them, which is not always possible.

[72] Kepa Bengoetxea, “Estaldura zabaleko euskararako analizatzaile sintaktiko estatistikoa”, (Doctoral dissertation, University of the Basque Country (UPV/EHU), 2014), http://ixa.si.ehu.es/sites/default/files/dokumentuak/4131/Tesi%20txostena.pdf

[73] http://ixa2.si.ehu.eus/maltixa/index.jsp

[74] Iakes Goenaga, “ASKHi: Analisi sintaktiko konputazional hibridoa paradigma esberdinen konbinazioan oinarrituta”, (Doctoral dissertation, University of the Basque Country (UPV/EHU), 2017) http://hdl.handle.net/10810/21586

[75] María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios, I. (2013). “Transforming Complex Sentences using Dependency trees for Automatic Text Simplification in Basque”, *Procesamiento del lenguaje natural,* 50 (2013): 61-68. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4660/2762

[76] Itziar Gonzalez-Dios *et al.*, “Detecting Apposition for Text Simplification in Basque”, in *International Conference on Intelligent Text Processing and Computational Linguistics,* (Berlin, Heidelberg: Springer, 2013), 513-524, https://link.springer.com/chapter/10.1007/978-3-642-37256-8_42

[77] Eneko Agirre and Aitor Soroa, “Personalizing PageRank for Word Sense Disambiguation”, in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, (Association for Computational Linguistics, 2009), 33-41, https://aclanthology.org/E09-1005.pdf

[78] http://ixa2.si.ehu.eus/wsd-demo/

[79] Haritz Salaberri, Olatz Arregi, and Beñat Zapirain, “bRol: The Parser of Syntactic and Semantic Dependencies for Basque”, in Proceedings of the International Conference Recent Advances in Natural Language Processing (INCOMA Ltd. Shoumen, 2015), 555-56,2 https://aclanthology.org/R15-1072.pdf

[80] Ander Soraluze *et al.*, “EUSKOR: End-to-end Coreference Resolution System for Basque”, Plos one, 14 (2019): e0221801, https://doi.org/10.1371/journal.pone.0221801

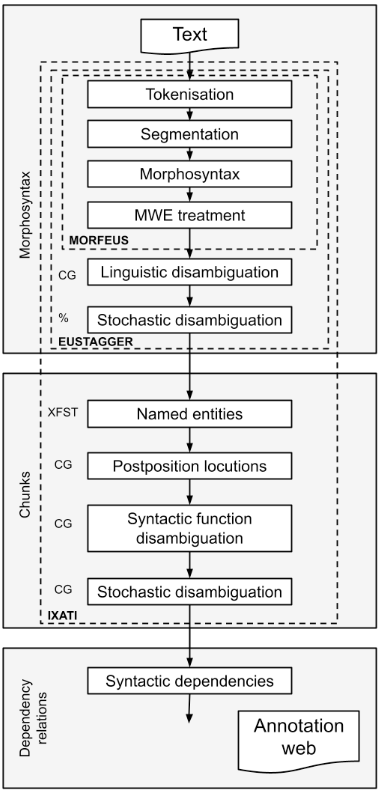

Figure 2

Schema of the basic annotation chain for Basque

In order to make the basic processing easier and free, open source tools have been released in the subsequent years: IXA pipes[81,82] and ixaKat[83,84]. Ixa pipes are a set of data-driven tools for different languages with open

[81] https://ixa2.si.ehu.eus/ixa-pipes/

[82] Rodrigo Agerri, Josu, Bermudez, and German Rigau, G. (2014). "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), (European Language Resources Association , 2014), 3823-3828, https://aclanthology.org/L14-1605/

[83] http://ixa2.si.ehu.eus/ixakat/

[84] Arantza Otegi *et al.*, "A Modular Chain of NLP Tools for Basque", in *International Conference on Text, Speech, and Dialogue*, (Springer, 2016), 93-100, https://ixa.ehu.eus/sites/default/files/dokumentuak/8328/tsd750.pdf

licence (Apache Licence 2.0). In the case of Basque, the tools available are the tokenizer, the PoS tagger, named entity recognition tagger and the probabilistic chunker. ixaKat, on its part, is a modular chain of four ready-to-use and freely available tools (GPL v3 free licence) that include the following processes: morphological analysis and PoS tagging, dependency parsing, semantic role labelling and coreference resolution. As both IXA pipes and ixaKat use an XML-based format, the ixaKat chain output can straightforwardly be processed with IXA pipes tools such as the Named Entity Disambiguation tool and the Wikification tool, both based on UKB. These tools offer the necessary information to process Basque texts effectively and without licence problems, making it possible for software developers to integrate them in end-user products. IxaKat tools have been downloaded more than 1900 times since 2017, reflecting the great interest created by the chain among scholars and developers and the importance of free licensing of basic NLP resources.

As mentioned, nowadays NLP has evolved and relies more on data-driven approaches than on linguistic rules. In fact, the last wave in NLP has been dominated by tools that make abstract representations of languages based on the information in huge corpora without little or any human supervision. Based on pure data-driven approaches, three kinds of resources can be distinguished:

- Static word embeddings (aka word embeddings, or embeddings): word embeddings are vectorial representations of words, that is, each word is converted to a multi-dimensional numerical representation in which the linguistic and non-linguistic features are different dimensions and the words gets a certain value for each of them. In Basque, there are two resources freely available: the ones created by Goikoetxea *et al*.[85,86] and the fastText embeddings[87,88].
- Contextual embeddings: static word embeddings have only one vectorial representation for each word. This means that ambiguous words such as *bank* share the same representation. In order to overcome this problem, contextual word embeddings include information from the contexts in which words appear in text in their representations. This

[85] https://ixa2.si.ehu.eus/ukb/bilingual_embeddings.html

[86] Josu Goikoetxea, Aitor Soroa, and Agirre, "Bilingual Embeddings with Random Walks over Multilingual Wordnets", *Knowledge-Based Systems*, 150, (2018): 218-230, https://doi.org/10.1016/j.knosys.2018.03.017

[87] https://fasttext.cc/

[88] Piotr Bojanowski *et al*., "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146, https://aclanthology.org/Q17-1010.pdf

way, different senses of words are coded. The available contextual word embeddings for Basque are distributed in FLAIR[89,90].

— Pre-trained language models: language models are probability distributions over sequences of words (vectorial representations). That is to say, in language models, sequences of words have attached a probability value of occurring together, based on their occurrence in vast amounts of data. As training them is expensive, time consuming and harmful for the environment[91,92] due to the size of the training data and the complexity of the calculations, pre-trained language models are used. That is, language models are usually trained by big technological companies that have the necessary computing power and then they are adapted to other tasks and purposes (fine-tuning). Those pre-trained language models are ready to be distributed and they are uploaded to HuggingFace[93], a company that stores comunity- and company-built language models and resources and that aims at democratising and advancing in machine learning. For Basque, at the writing of this paper (February 2022), we find three monolingual pretrained language models: Berteus[94,95], RoBasquERTa[96], byt5-basque[97], and another four roberta-based models just released in March 2022[98] for text processing and Wav2Vec2-Large-XLSR-Basque[99], Wav2vec2-large-xls-r-300m-

[89] https://github.com/flairNLP/flair

[90] Alan Akbik, Duncan Blythe and Rolang Vollgraf, "Contextual String Embeddings for Sequence Labeling", in *Proceedings of the 27th international conference on computational linguistics,* (Association for Computational Linguistics, 2018), 1638-1649, https://aclanthology.org/C18-1139.pdf

[91] Emma Strubell, Ananya Ganesh, and Andrew McCallum, A. (2019). "Energy and Policy Considerations for Deep Learning in NLP", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, 2019), 3645-3650, https://aclanthology.org/P19-1355.pdf

[92] Emily M. Bender, *et al.*, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* (Association for Computing Machinery, 2021) 610-623, https://dl.acm.org/doi/10.1145/3442188.3445922

[93] https://huggingface.co/

[94] https://huggingface.co/ixa-ehu/berteus-base-cased

[95] Rodrigo Agerri *et al.*, "Give your Text Representation Models some Love: the Case for Basque", in *Proceedings of the 12th Language Resources and Evaluation Conference*, (European Language Resources Association, 2020), 4781-4788. https://aclanthology.org/2020.lrec-1.588/

[96] https://huggingface.co/mrm8488/RoBasquERTa

[97] https://huggingface.co/monsoon-nlp/byt5-basque

[98] Mikel Artetxe *et al.*, "Does Corpus Quality Really Matter for Low-Resource Languages?2" *arXiv preprint arXiv:2203.0811*1, (2022), https://arxiv.org/pdf/2203.08111.pdf

[99] https://huggingface.co/cahya/wav2vec2-large-xlsr-basque

basque[100] for speech among others. The most important multilingual pre-trained LMs for Basque that process text are IXAmBERT[101,102] and the multilingual Bert[103,104]. At this moment, a large multilingual language model created by the BigScience workshop[105] is being trained and it also includes Basque[106].

As we see in the latest months, the quantity of language models created is increasing and they are substituting the classical NLP pipeline in order to process the texts. Furthermore, as language models already provide the needed linguistic knowledge, creating linguistic processing tools that rely on them is a rather straightforward programming task.

Developing each tool of the classical NLP pipeline was a work that was carried out in the contexts of PhD theses, which take at least 4 years. Creating large language models, instead, is faster but requires large quantities of texts and enormous computational capabilities, which are not normally available in the contexts of the lesser resourced languages. As example of what the training of a larger language model takes, the training BigScience's multilingual model was launched on the 11th of March, 2022[107] and 15 days later, on the 24th of March, 2022[108], it had only reached 10% of its training.

3. *Corpora*

As we mentioned, language models are the basic processor nowadays in NLP, the engines of the system. But all the engines need fuel to work,

[100] https://huggingface.co/deepdml/wav2vec2-large-xls-r-300m-basque

[101] https://huggingface.co/ixa-ehu/ixambert-base-cased

[102] Arantza Otegi *et al.*, "Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque", in *Proceedings of The 12th Language Resources and Evaluation Conference* (European Language Resources Association, 2020), 436-442, https://aclanthology.org/2020.lrec-1.55.pdf

[103] https://huggingface.co/bert-base-multilingual-cased

[104] Jacob Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Association for Computational Linguistics, 2018), 4171-4186, https://aclanthology.org/N19-1423.pdf

[105] https://bigscience.huggingface.co/

[106] https://bigscience.huggingface.co/blog/building-a-tb-scale-multilingual-dataset-for-language-modeling

[107] https://bigscience.huggingface.co/blog/what-language-model-to-train-if-you-have-two-million-gpu-hours

[108] https://twitter.com/BigScienceLLM/status/1506933846830366722

and their fuel is corpora. There are more that 20 Basque corpora (a detailed list and explanation is gathered in the Basque Wikipedia[109]), but the main problem is that their availability is limited. Most of them can be consulted online without limitations, but they cannot be downloaded (with an appropriate licence) and others have closed licences. These facts hinder the progress in NLP since this fuel cannot be used.

In order to overcome this problem, EusCrawl[110] has been created. EusCrawl is the second largest available corpus for Basque, but it has been built following the tailored crawling approach, where websites with high-quality content are manually identified and then scrapped (i.e. the contents are downloaded automatically). After the results of manual evaluation of the data, EusCrawl has a much higher quality than other web-scrapped corpora, becoming the biggest high quality corpus for Basque.

V. Discussion: initiatives and lessons learnt

Speakers of lesser resourced and minority languages face three main problems in the digital world: i) the availability of technology, ii) usability of technology, when available and iii) how this technology has been developed for minority languages. Minority languages have less technologies available and using them entails in many cases flaws and difficulties e.g. lack of specific keyboards. Moreover, it has been corroborated that minority language speakers switch to the dominant language, because the technology is better or there are more services. And, finally, regarding the development of the technologies can be done by big companies with little or no involvement of the communities or by activists' approach[111]. An example of an activist approach was Codefest[112], a hackathon/summer lab that took place in the context of the European Culture Capital Donostia 2016 that aimed to revitalise resource scarce languages by providing the communities with effective tools and by teaching them how to use them.

Moreover, the use of open source initiatives and social media resources (blogs, Telegram, Twitter) are important to develop tools and resources for

[109] https://eu.wikipedia.org/wiki/Testu_corpus

[110] Mikel Artetxe *et al.*, "Does Corpus Quality Really Matter for Low-Resource Languages?2" *arXiv preprint arXiv:2203.0811*1, (2022), https://arxiv.org/pdf/2203.08111.pdf

[111] Claudia Soria, "Decolonizing Minority Language Technology", *State of the Internet's languages report,* (2022), https://internetlanguages.org/en/stories/decolonizing-minority-language/

[112] https://www.ehu.eus/ehusfera/ixa/2016/04/07/codefest-summer%C2%ADlab-aims-to-revitalise-resource-%C2%ADscarce-language-donostia-july-4-8/

the communities[113]. Having tons of texts with open licences is one of the key aspects nowadays in the development of the technologies. This way, the community itself, together with its language experts, can be an active agent of its tools, and not only mere users of the products created by the big tech companies such as the large language models. And developing tools with open licence helps in the progress of language technology.

NLP tools save money and time as they automatise many data extraction and generation processes, but funding is necessary to create appropriate tools for the community with the guarantee and the fullest possible cover. This implies that the institutions, which will be also beneficiaries of the process, need to invest money if they want to preserve their language. And this should not be done with the communities and open source.

Public investment for Basque NLP is not neglectable though, and it should be preserved in time. Spain has a well-funded plan for LT[114] alongside the Coordinated Plan on Artificial Intelligence,[115] and the Spanish strategy R+D+i for Artificial Intelligence,[116] from which Basque NLP gets some funding. Unfortunately, as of now, there is no equivalent plan in the Basque Autonomous Community or in Navarre, although the Basque Autonomous Community has promoted a list of NLP projects, through the Etortek and ElkarTek Industry Programmes, in which public research institutions and private entities have collaborated in the creation of Language Technology resources. In the French area of the Basque Country, the Basque Municipal Community promotes a digital agenda, as does the IKER research centre, the only laboratory in France that specialises in Basque Studies.

Backing of initiatives by public institutions is crucial, but the creation of infrastructures in which knowledge and expertise exchange is promoted is the best approach to the dissemination of the relevant efforts. In Europe two infrastructures are to be mentioned: CLARIN[117] and DARIAH[118]. Both support the sharing of tools and resources for research in the humanities and encourage the re-use of existing tools. Scholars and researchers in the Basque Humanities field are now setting up CLARIAH-EUS[119], the Basque node for CLARIN and

[113] Imna Hernáez *et al.*, *Euskara Aro Digitalean — The Basque Language in the Digital Age* (Springer, 2012), https://link.springer.com/book/10.1007/978-3-642-30796-6

[114] Plan de Impulso de las Tecnologías del Lenguaje, Ministerio de Turismo, Energia y Agenda Digital, 2015, http://www.ciencia.gob.es/portal/site/MICINN/menuitem.26172fcf4eb029fa6ec7da6901432ea0/?vgnextoid=70fcdb77ec929610VgnVCM1000001d04140aRCR

[115] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

[116] http://www.ciencia.gob.es/portal/site/MICINN/menuitem.26172fcf4eb029fa6ec7da6901432ea0/?vgnextoid=70fcdb77ec929610VgnVCM1000001d04140aRCR

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

[118] https://www.dariah.eu/

[119] http://ixa2.si.ehu.eus/clariah-eus/

DARIAH. Through CLARIAH-EUS, the Humanities community will be able to reach the NLP resources and tools in the platform and integrate them in their services fostering the interaction of Basque with other languages.

Therefore, our recommendation for other minority languages is that data and tools should be developed by the community with open licences, so they can be shared easily, and taking their needs into account, so they are genuinely useful to the community. Depending on the texts that they have available, they can start by creating a lexical database and follow the classical NLP pipeline, but, if they have a substantial amount of texts, they can try deep learning approaches. All in all, we hope that the Basque experience serves them to pave the way to the digital world and so that we can talk to advanced assistants like Siri in our languages[120].

VI. Conclusion

In a world in which interpersonal communication is held more and more frequently in the digital domain, speakers need the tools that will allow them to interact in the most effective way. In the case of minority language speakers, this implies developing the resources and tools that will ensure communication among the speakers of the minority languages as well as with a wider public. In order to address the needs of the speakers, not only the speakers need to get involved in the development of the tools, but also the administrations and research institutions need to contribute either financing the development of apparently less profitable resources or adding the knowledge needed for it.

Those resources and tools, on their side, need previous Natural Language Processing. The linguistic resources and tools offered to the end user rely on the linguistic knowledge extracted and processed in the linguistic analysis pipeline. Moreover, the advances in NLP now allow fast and largely comprehensive analysis and have notoriously reduced the effort needed for the different NLP tasks, making the investment in this previous step more productive. Hence, public administrations should be eager to promote this core research as a way to effectively give an answer to a wide range of people's needs. We would like to emphasise the importance of the involvement of the Basque local administration in the promotion of Language Technologies for Basque as Basque is a language in the area of their competence and due to the fact that local management should improve the process and make it more efficient.

[120] https://www.thejournal.ie/minority-languages-digital-equality-5725794-Mar2022/

Basque has come a long way, and might be inspiring for other low resourced languages, mostly for the ones that have complex morphosyntactic features, on their way to become a digital language. All languages can profit from the benefits of language technologies either when their starting point is the creation of a lexical database such as EDBL or by relying on multilingual resources created by international researchers or big tech companies. Having will and a community of potential users behind are the first steps. Language complexity or the lack of available data can be easily overcome if room for experimentation is granted. From Xuxen to Euscrawl, Basque has been paving its way to the highest levels of language technologies. However, this success, which is far from attaining the perfect situation, is due to the implication and the will of the community (users, researchers and stakeholders), the availability of a standard variant and the stake for open licence resources and technologies.

Acknowledgements

This survey has been partially funded by Basque Government post-doctoral grant POS 2021 2 0030. We also want to thank Gorka Labaka for providing us with the current statistics of EDBL-LBDBL, Zuhaitz Beloki with the statistics about Xuxen and Esther Miranda and Arantxa Otegi with the statistics about IxaKat. We also acknowledge the following projects: DeepText (KK-2020/00088), IXA group (Basque Government (excellence research group IT1343-19).

About the autors

Itziar Gonzalez-Dios (Pasai San Pedro, 1988) (Orcid: 0000-0003-1048-5403) is an assistant professor at the Faculty of Engineering in Bilbao in the department of Basque Language and Communication and researcher of the HiTZ center (Ixa group) from the University of the Basque Country (UPV/EHU). She received her PhD on Language Analysis and Processing (computational linguistics) in 2016, her M.A. on the same topic in 2011 and her B.A. on German Philology in 2010, all of them at the University of the Basque Country (UPV/EHU). She has published over 45 international peer-reviewed articles and conference papers in Natural Language Processing, mainly in the areas of readability assessment and automatic text simplification. Her research is also focused on developing lexical, semantic and terminological resources for less resourced languages. She has participated in national and international research projects. She has also served as reviewer in various international journals, conferences and workshops and has experience organizing international scientific conferences, workshops and hackathons. She speaks fluently Spanish, Basque, English, German, French and Italian.

Begoña Altuna (Bilbao, 1989) (Orcid: 0000-0002-4027-2014) is a postdoctoral researcher at the HiTZ center (Ixa group) of the University of the Basque Country (UPV/EHU). She has a PhD in Language Analysis and Processing (University of the Basque Country, 2018), as well as a degree in Basque Philology (University of Deusto, 2011) and a master's degree in Language Analysis and Processing (University of the Basque Country, 2013). She has done two stays as a visiting researcher at the Fondazione Bruno Kessler (Trento, Italy), one as a predoctoral researcher (2016) and the second as a postdoctoral researcher (2020-2022) as a beneficiary of the Basque Government postdoctoral fellowship. Her main line of research is the analysis of temporal information in Basque, the creation of annotated corpora and the development of tools for the extraction of temporal information. In addition, she collaborates in research on the analysis of neural networks and in the field of digital humanities. She is the author of more than 20 peer-reviewed publications in international journals and conferences in the field of natural language processing. She has participated in several national and international projects, having special responsibility in the European Clinical Case Corpus (E3C) project. In addition, she has organized seminars and workshops and is a reviewer at several international conferences.

Sobre las autoras

Itziar Gonzalez-Dios (Pasai San Pedro, 1988) (Orcid: 0000-0003-1048-5403) es profesora adjunta (ayudante doctor) del departamento de Lengua Vasca y Comunicación en la escuela de Ingeniería de Bilbao e investigadora del centro HiTZ (grupo Ixa) de la Universidad del País Vasco (UPV/EHU). Es doctora en Análisis y Procesamiento de Lenguaje (lingüística computacional) (2016), tiene un máster en la misma área de conocimiento (2011) y licenciada en filología alemana (2010), todos ellos por la UPV/EHU. Ha publicado más de 45 artículos internacionales revisados por pares y documentos de conferencias en procesamiento del lenguaje natural, principalmente en las áreas de evaluación de la complejidad y de la simplificación automática de textos. Su investigación también se centra en el desarrollo de recursos léxicos, semánticos y terminológicos. Ha participado en varios proyectos de investigación nacionales e internacionales y también ha sido revisora en varias revistas y conferencias. Tiene experiencia organizando workshops científicos y hackathones. Habla con fluidez español, euskera, inglés, alemán, francés e italiano.

Begoña Altuna (Bilbao, 1989) (Orcid: 0000-0002-4027-2014) es investigadora postdoctoral en el centro HiTZ (grupo Ixa) de la Universidad del País Vasco (UPV/EHU). Es doctora en Análisis y Procesmaiento del Lenguaje (Universidad del País Vasco, 2018), así como licenciada en Filología Vasca (Universidad de Deusto, 2011) y graduada del máster de Análisis y Procesamiento del Lenguaje (Universidad del País Vasco, 2013). Ha realizado estancias como investigadora visitante en la Fondazione Bruno Kessler (Trento, Italia), una como investigadora predoctoral (2016) y la segunda como investigadora postdoctoral (2020-2022) como beneficiaria de la beca de postdoctorado del Gobierno Vasco. Su principal línea de investigación es el análisis de la información temporal en euskera, la creación de corpus anotados y el desarrollo de herramientas para la extracción de la información temporal. Además, colabora en investigaciones sobre el análisis de las redes neuronales y en el ámbito de las humanidades digitales. Es autora de más de 20 publicaciones revisadas por pares en revistas y congresos internacionales en el ámbito del procesamiento del lenguaje natural. Ha participado en varios proyectos nacionales e internacionales, teniendo especial responsabilidad en el proyecto European Clinical Case Corpus (E3C). Además, ha organizado seminarios y workshops y es revisora en varios congresos internacionales.