EUROPEAN LANGUAGE EQUALITY

D1.4

Report on the Basque Language

Authors	Kepa Sarasola, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernaez, Eva Navas
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.4
Deliverable title	Report on the Basque Language
Type	Report
Number of pages	26
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Kepa Sarasola, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernaez, Eva Navas
Reviewers	Annika Grützner-Zahn, Maria Giagkou
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
DLPD	Digital Language Diversity Project
GPU	Graphics Processing Unit
HAC	Hizkuntzen Arteko Corpusa
HCI	Human Computer Interaction (see HMI)
HPC	High-Performance Computing
LR	Language Resource/Resources
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLG	Natural Language Generation
NLP	Natural Language Processing
POS	Part-of-Speech
SR	Speaker Recognition
TLD	Top-level Domain
TTS	Text-to-Speech

Abstract

This report on the Basque Language is part of a series of language deliverables developed within the framework of the European Language Equality (ELE) project. The series seeks to not only delineate the current state of affairs for each European language, but to additionally identify the gaps and factors that hinder further development in research and technology. The survey presented here focuses on the condition of Language Technology (LT) with regard to Basque, a language with 751,000 speakers within a territory that spans across part of Northern Spain and Southern France. Basque has been immersed in a process of revitalisation since 1968 that has faced formidable obstacles. Nonetheless, despite these challenges, significant progress has been made in numerous areas. Together, these advances have fostered the necessary conditions for the successful development and dissemination of LT, including two important prerequisites: a language community that can construct LT tools and a widely accepted standardised language that facilitates the effective uptake of novel Natural Language Processing (NLP) technology. This course of events is noteworthy because LT is now recognised as a powerful aid in enabling low-resource language communities to revitalise and support their languages. After thirty years of collaborative work among several groups, research has resulted in state-of-the-art technology and robust, broad-coverage NLP for Basque. However, a dramatic difference remains between Basque and other European languages in terms of both the maturity of research and the state of readiness with respect to language solutions.

While there are good reasons to be optimistic about Basque's future in this arena, the development of high-quality language technology for under-resourced languages is urgent and important for their preservation.

Laburpena

1968. urtetik aurrera Euskara etengabeko indarberritze prozesu batean murgildurik egon da, eta hainbat oztopo gainditu behar izan ditu. Horiek guztiak ahaztu gabe, aurrerapauso garrantzitsuak lortu dira arlo askotan. Esan genezake, aurrerapauso hauen atzean sei gako nagusi daudela: 1) Euskara Batuaren ezarpen eta onarpen ofiziala 1968. urtean; 2) euskara hezkuntza sisteman sartu izana; 3) euskarazko komunikabideen sorrera (irratia, egunkariak, telebista); 4) euskara ofiziala izateko marko legala eratu izana; 5) instituzioen eta gizarte eragileen arteko elkarlana; eta 6) alfabetatze kanpainak. Oro har, eta, era berean, elementu hauek guztiek hizkuntza teknologiaren garapena eta zabalkundea ahalbidetu dute, aurretik betetzen ziren bi baldintza ezinbestekoekin batera: hizkuntza-teknologiako tresnak sortzeko gai zen komunitate linguistikoa, eta hizkuntzaren prozesamendurako teknologia berri hau garatzea eta ezartzea errazten zuen hizkera estandar onartua. Hizkuntzaren prozesamenduan corpusa biltzea, online hiztegia, zuzentzaile ortografikoa, analizatzaile morfologikoa, corpusen etiketatzea, POS (*Part Of Speech*) etiketatzailea eta testuen meatzaritza dira besteak beste, aurre egin beharreko lehen urratsak, eta urrats horiek guztiak aurrera eramateko ezinbestekoa da gizarte-eragileek onartutako hizkera estandarra. Euskararen ibilbide hau aipagarria da oso, izan ere, gaur egun, hizkuntza-teknologia baliabide gutxiko hizkuntza-komunitateek euren hizkuntzak biziberritu eta babestu ahal izateko tresna garrantzitsua dela aitortzen baita.

Hogeita hamar urtez hainbat talderen artean elkarlanean aritu ondoren, ikerketak euskararentzako abangoardiako teknologia eta estaldura zabaleko hizkuntzaren prozesamendu sendoa eman du. Sortutako baliabideen artean honako hauek aipatuko genituzke: 48 milioitik gora hitz dituzten lau corpus elebakar (handienak 355 miloi ditu), Basque WordNet ontologia, morfologikoki eta sintaktikoki etiketatutako corpusak (ZT eta EPEC), hizkuntza-

ereduak (BERT eta T0), hizketaren hainbat datu-base (SpeechDat, ADITU, AhoSyn, AhoEmo), eta aplikazio desberdinak (Xuxen zuzentzaile ortografikoa, itzultzaile automatiko neuronalak, testuen prozesamendua egiteko hizkuntza naturalaren prozesamendurako Ixa-pipes katea, ahotsa ezagutzeko tresnak, testutik ahotsera pasatzeko tresnak, eta iritzien meatzaritza egiteko Behagunea izeneko tresna, besteak beste).

Hala ere, oraindik ere alde nabarmena dago euskararen eta Europako gainerako hizkuntzen artean, ikerketaren heldutasunari eta hizkuntza-irtenbideen inguruko prestakuntza-egoerari dagokienez. MC4 dataset eleaniztunak, adibidez, 10,401 Gb eskaintzen ditu ingeleserako, 1,613 Gb gaztelaniarako (6 aldiz gutxiago), eta 5 Gb bakarrik euskararako (2.000 aldiz gutxiago). Era berean, BERT hizkuntza-ereduaren ingeleserako jatorrizko bertsioa Google Books-en corpus bat erabiliz entrenatu zen. Corpus horrek 155.000 milioi hitz ditu Estatu Batuetako ingelesez eta 34.000 milioi hitz Britainia handiko ingelesez. Horrek esan nahi du corpus ingelesa bere euskal baliokidea (384 milioi hitz) baino ia 500 aldiz handiagoa zela 2020an. Hizkuntzen arteko alde hori hizketarako baliabideetan ere argi ikusten da. Common Voice enpresak, adibidez, 2015 baliozkotutako hizketa-ordu ematen ditu ingeleserako, 377 gaztelaniarako, eta 91 bakarrik euskararako.

Goiko adibide gutxi horietan ikusten den hizkuntzen arteko alde nabariak hizkuntza teknologian dagoen eten digital endemikoa azpimarratzen du. Hala ere, euskara bezalako baliabide gutxiko hizkuntzentzat puntu positiboa da aurrez prestatutako hizkuntza-eredu elebakarrek eta eleaniztunek nahiko emaitza onak ematen dituztela Hizkuntzaren Prozesamenduko ataza desberdinetan, baita entrenamendurako corpus askoz txikiagoak erabilita ere.

Beraz, duela hamarkada bat euskarari buruzko 2012ko META-NET liburu zuri aitzindarian adierazi zen bezala, egunerokoan erabiliko diren hizkuntza teknologiako irtenbide benetan eraginkorrak prest egon daitezen, euskarak, oraindik ere, ikerketa gehiago behar duten EBko hizkuntzen artean egon behar du. Euskarak arlo horretan izango duen etorkizunaz baikor izateko arrazoi onak dauden arren, baliabide gutxi dituzten hizkuntzetarako kalitate handiko hizkuntza-teknologia garatzea premiazkoa eta garrantzitsua da haien bizirautea bermatzeko. Hori nola egin ulertzeko, azken bost urteetan hizkuntza-teknologian eta IKTetan egin diren aurrerapenei begiratu besterik ez dago, Europako hizkuntza bakoitzerako ofizialtasun-koefizienteak ezartzea eta arrazoizko kostuan aplikatzea ahalbidetzen baitute. Nahiz eta ofizialtasunak ez duen nahitaez hizkuntza jakin bat dokumentu edo alor guztietan egon behar denik esan nahi, Europako hizkuntza-aniztasunaren ahal handia aberastuko luke ezbairik gabe.

1 Introduction

This study is part of a series that reports on the results of an investigation into the level of support European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, experts in more than 30 European languages, have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.

The report has been developed within the framework of the European Language Equality (ELE) project. With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initia-

tives, the ELE project develops a strategic research, innovation and implementation agenda as well as a road map for achieving full digital language equality in Europe by 2030.

2 The Basque Language in the Digital Age

2.1 General Facts

The Basque language is spoken by 28.4% (751,500) of Basques in a territory that spans across part of Northern Spain and Southern France. Of these, 93.2% (700,300) reside on the Spanish side of the Basque Country and the remaining 6.8% (51,200) live in the French region.[1] The standardisation of a language is a prerequisite for the successful use of its written form. However, Basque is an extremely fragmented language; a number of dialects and sub-dialects spread over an area of 10,000 km2. The dialectal split began in the early Middle Ages (Mitxelena, 1981) and, over the last few centuries, the linguistic distance between dialects has been increasing to the extent that today peripheral varieties are not mutually intelligible in oral speech by non-trained speakers. At present, we may distinguish between five main dialects of the Basque language: the Western dialect, also called "Bizkaian", the Central dialect, also known as "Gipuzkoan", the (High) Navarrese dialect, the (Low) Navarrese-Lapurdian dialect and the Zuberoan dialect (Zuazo, 2014).[2]

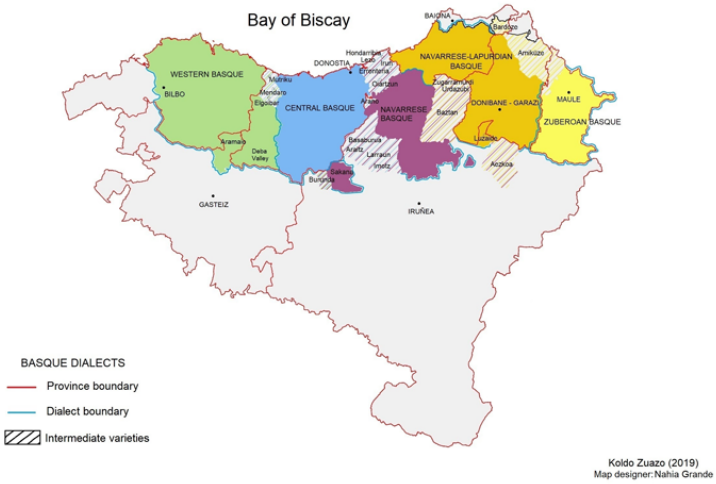

Figure 1: Five main dialects of the Basque Language.

These five dialects are noticeably distinct from each other and, while there were sporadic attempts in the early twentieth century to bring some uniformity to Basque, it was not until 1968 that the Royal Academy of the Basque Language (founded in 1919)[3] decided to standardise it. Standard Basque (*Batua*) is a literary variety constructed upon central dialects of the language. The basis of Standard Basque is formed by a spelling system, paradigms of noun and verb morphology, syntactic rules, and an official dictionary (*Euskaltzaindiaren Hiztegia*).[4] After some years of discussion, this Standard Basque became widely accepted and it

[1] VIème Enquête Sociolinguistique en Euskal herria (2016). https://www.mintzaira.fr/fileadmin/documents/Aktualitateak/015_VI_ENQUETE_PB__Fr.pdf
[2] http://euskalkiak.eus/en/ezaugarriak.php
[3] https://www.euskaltzaindia.eus/en/
[4] https://www.euskaltzaindia.eus/index.php?option=com_hiztegianbilatu&view=frontpage&Itemid=410&lang=eu

is now utilised in almost all formal contexts: school, university, administration, and official pages on the Internet. Similarly, Standard Basque is employed by academics and journalists across every type of media, including print, radio, and television.

There are six predominant factors behind this relatively positive sociolinguistic assimilation: 1) the aforementioned implementation and official acceptance of Standard Basque; 2) integration of Basque into the educational system; 3) creation of media in Basque; 4) construction of a legal framework; 5) collaboration between public institutions and grassroots organisations; and 6) campaigns for Basque language literacy (Agirrezabal, 2010). Even so, the successful societal acceptance of Standard Basque is remarkable given the fact that there is no administration common to all territories where Basque is spoken. Not only is the language transnational, spanning both France and Spain, separate administrative jurisdictions also exist within the latter and possess different legislation regarding the Basque language. Moreover, Basque speakers are almost always fully bilingual in either Spanish or French, so that the existence of a standard Basque language is not strictly required for communication beyond the local level.

The inconsistency of Basque's status as an official language adds a further dimension to this variegated terrain. The Basque Autonomous Community in Spain (provinces of Álava, Biscay, and Gipuzkoa) has established Basque as a co-official language. Spanish is the only official language for all of the Chartered Community of Navarre, but it grants co-official status to the Basque language in the Basque-speaking areas of northern Navarre. Basque has no official status in the French Basque Country. The same is true for the European Union, which limited the status of official European language to state languages and, thus, ensured the latter's compulsory and extensive use in European administration. This decision was based on the erroneous idea that there is only one language in each member state (state monolingualism). Although, in principle, Article II-82 of the Treaty establishing a Constitution recognises European linguistic diversity, the legal scope of this article is neither clear nor fully developed. Indeed, the measures for its application both to state languages and regional or minority languages have yet to be defined (Urrutia, 2004; Urrutia and Lasagabaster, 2007). Future recognition of Basque as an official European language would represent a significant stride towards realising Europe's desire for greater inclusion, even if official status did not guarantee the language's presence in all documents or spheres.

As a non-Indo-European language isolate, Basque grammar differs considerably from that of the languages surrounding it. Nevertheless, Basque has borrowed up to 40% of its vocabulary from Romance languages and the Latin script is used as the language's writing system. Basque is agglutinative, head-final and pro-drop and this can be challenging for computational processing. A declarative sentence in Basque contains a verb and its arguments, an aspect marker attached to the verb and a verbal inflection containing agreement morphemes, tense and modality. It can also contain other phrases, such as adverbials or postpositional phrases (Laka, 1996). The arguments of the verb can be identified by grammatical cases or postpositions. There are three grammatical cases in Basque: Ergative (k morpheme), Dative (i morpheme) and Absolutive (Ø morpheme). Basque has a strong tendency to place the heads of phrases at the end of the phrase. Rather than prepositions at the beginning of prepositional phrases, Basque has post-positions that appear at the end of postpositional phrases. Grammatical cases are no exception to this generalisation (Laka, 1996). Consequently, Basque is one of the so-called "rich morphology" languages, and this may cause issues in NLP that do not occur in other languages.

2.2 Basque in the Digital Sphere

The Basque language in the digital sphere can be measured according to several data provided by various institutions and projects. Data collected by the Basque Institute of Statistics

(EUSTAT),[5] shows that 84.9% of people aged fifteen and over in the Basque Autonomous Community (1,602,600 individuals) connected to the Internet between June and September 2021. This level of interaction was partially captured in the two latest reports published in 2016 and 2017 by the PuntuEUS Observatory,[6] which measures the presence of Basque on the Internet each year. According to the data provided by the Observatory, there are currently 12,470 websites with the Basque language code (.eus) as the top-level domains (TLD). In 2020, the number of websites with the Basque language code as the TLD that had content in Basque was 84.4%.

In 2017, the Digital Language Diversity Project (DLPD)[7] conducted a survey on digital use and usability of regional and minority languages. A high number of respondents stated that they used Basque regularly on the Internet. Websites, blogs, forums, and smartphone apps were found to be the most widespread digital media in Basque (Gurrutxaga and Ceberio, 2017). Indeed, a remarkable number of respondents indicated that they had a blog of their own. Internet television, streaming audio and video can also be found in Basque, but few speakers were aware they exist. Similarly, there is a Wikipedia available for Basque and nearly all respondents (97%) were aware of its existence. Only 11% of them, however, were active users and content producers. The majority (81%) browsed it, while a few (8%) made no use of it. Basque was actively employed in electronic communication by respondents of the same survey (97%) for writing emails, texting, chatting or other instant messaging such as Whatsapp, Google Chat, Snapchat, Skype, and Facebook Messenger. Instant messaging applications appeared to be the most utilised instruments for e-communication, followed by regular email and texting/SMS. Still, Basque was less prevalent on LinkedIn, the business and employment-oriented social networking service. This may reflect the situation of the Basque language in general, where use has grown in familiar, academic and informal settings, but where more work needs to be done within professional and formal contexts. It is also true that there is a demand for more entertainment products in Basque, especially for young people. Most individuals consume computer or mobile games in other languages because they are rarely available in Basque.(Gurrutxaga and Ceberio, 2017)

The results of the survey demonstrate that Basque is a digitally fit and actively utilised language online. Respondents demonstrated a high linguistic competence and good knowledge of existing digital tools and resources. This also applied to social media, especially to Facebook and Twitter, which showed significant activity in Basque. We may add that Twitter has been translated to Basque in a collaborative way. Despite this, and although translation of the Twitter interface was already underway in 2012, Twitter does not have Basque in its language detection for tweets, nor does it produce trending topics in Basque. Instead, the native application Umap.eus filters the content of Twitter in Basque and extracts the trending topics on a daily basis, in addition to providing a ranking of Basque speakers amongst Twitter users (Goñi, 2013).[8] According to data provided by Umap.eus, there were 10,542 active users on Twitter in March 2021 that tweeted in Basque. These users published 582,957 tweets per month and 40.7% of those tweets were in the Basque language.[9] Because these types of social networks are mostly associated with informal registers, their use corroborates that Basque is both used and useful for everyday spoken and written online communication, an undeniable sign of vitality. This is an important factor in the digital survival of a language and, significantly, respondents to the DLPD survey expressed a strong desire to be able to use Basque online as part of their everyday life. There was a high level of agreement that Basque was suited for use on the Internet and that it should not be considered appropriate

[5] https://www.eustat.eus/elementos/la-comunicacion-con-los-demas-y-la-busqueda-de-informacion-lo-mas-utilizado-por-la-poblacion-de-la-ca-de-euskadi-usuaria-de-internet-el-849-en-2021/not0019072_c.html
[6] https://www.domeinuak.eus/en/observatory/
[7] http://www.dldp.eu
[8] http://basquetribune.com/lost-in-translation/
[9] https://umap.eus/media/pdf/umap_2021_3.pdf

only as a spoken language. More varied, and of note, however, was the reply to the question on ease of use: nearly a third of respondents felt that using Spanish was easier than Basque.

3 What is Language Technology?

Natural language[10] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably Artificial Intelligence (AI)), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

LT is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing's renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e. systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase,

[10] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.

- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Recognition, i. e. the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e. the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction (HCI)** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). Popular applications within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name a few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[11]

4 Language Technology for Basque

After thirty years of collaborative work among several groups, research has resulted in state-of-the-art technology and robust, broad-coverage natural language processing for Basque.

[11] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

Resources include monolingual corpora (containing from 48 to 355 million-words), Basque Wordnet, morphologically and syntactically tagged corpora (ZT and EPEC), BERT and T0 language models, several speech databases (SpeechDat, ADITU, AhoSyn, AhoEmo) and applications, such as a spellchecker (Xuxen), neural machine translators, NLP pipelines for text processing (Ixa-pipes), speech recognition and text-to-speech applications, and an opinion-mining tool (Behagunea), among others. Yet, as highlighted in the META-NET White Paper on Basque (Hernáez et al., 2012) in 2012, Basque LT still requires further research and development to offer truly effective LT solutions for everyday use. Although the presence of Basque LT has grown over the past 10 years, the development of high-quality LT for under-resourced languages such as Basque is urgent and important for its preservation. This section presents a comprehensive review of the support that the Basque language receives from LT. The data presented in Sections 4.1 and 4.3 are based on a meticulous compilation of metadata aimed at collecting and documenting ideally all datasets, tools, services, components, repositories, companies and research groups relevant to Basque LT in the last decade. The resources identified through this process have been imported as metadata records into the Catalogue of the European Language Grid (ELG).[12]

4.1 Language Data

As of February 2022, more than 300 resources are available for Basque in ELG. Approximately half of them are corpora; more than 100 tools/services are listed for Basque, while the rest are some lexical resources, language models and grammars.

Monolingual text corpora

Most of the Basque monolingual corpora available in the ELG are annotated at some linguistic level (lemma, morphology, multi-word unit, syntax, etc.). The largest 4 corpora included in this annotated group contain between 48 and 355 million words:

- The ETC corpus (Egungo Testuen Corpusa: 21st Century Basque text corpus)[13] is the largest and contains 355 million words drawn from books, newspaper articles, Wikipedia, and TV transcripts published in Spain and France in the 21st century. It is annotated at the lemma level and can be consulted online.
- The Corpus of the Lexical Observatory (Lexikoaren Behatokia Corpusa),[14] containing 98 million words in 2021, was created with texts taken from the media. The project was launched by the Academy of the Basque Language (Euskaltzaindia) in 2008 for the purpose of monitoring the use of Basque. It is tagged in XML and follows the TEI standard. The corpus is available online under the CC-BY-SA-4.0 licence.
- The Dabilena website[15] offers a language corpus made up of texts collected from the Internet. It is composed of two parts: a monolingual Basque corpus with 300 million words and a bilingual Basque-Spanish corpus, mentioned below.
- The CorpEus service,[16] which enables consultations to be made on the Internet as if it were an immense Basque-language corpus.

[12] https://www.european-language-grid.eu
[13] https://www.ehu.eus/etc/
[14] http://lexikoarenbehatokia.euskaltzaindia.eus
[15] https://dabilena.elhuyar.eus
[16] http://corpeus.elhuyar.eus/cgi-bin/kontsulta.py

Bilingual and multilingual text corpora

Most of the Basque text corpora in ELG are bilingual or multilingual. These types of corpora are generally composed of comparable or parallel data, among which we may highlight Paracrawl, WikiMatrix and Opensubtitles. Paracrawl contains Basque-Spanish parallel data released as part of the ParaCrawl project, which provides 64 million source words. WikiMatrix offers parallel corpora from Wikimedia in 86 languages. The total number of sentences for language pairs in which Basque is part of is 1,699,000. Opensubtitles is a collection of translated movie subtitles that provides sentence alignments between distinctive language pairs in different formats and with diverse annotation types. For Basque, the corpus contains 230 thousand sentences. Also worth mentioning is HAC (Hizkuntzen Arteko Corpusa), a cross-lingual corpus that contains 629,916 translation units for four languages (Basque, Spanish, French and English) and the Basque-Spanish EiTB corpus of aligned comparable sentences with 564,625 translation units.

These resources are complimented by corpora comprising texts from the Internet, compiled with web-crawling techniques. There are different approaches to building this type of corpora. Apart from the already mentioned Paracrawl, the Dabilena website contains around 34 million words of bilingual parallel corpora (15 million words in Basque) automatically extracted from domains with bilingual content. In contrast, mC4 comprises natural text in multiple languages drawn from the public Common Crawl web scrape, while the OSCAR corpus is a multilingual corpus obtained by language classification and a filter of the Common Crawl corpus. The former offers 5 Gb for Basque and the latter contains 97 million Basque words.

Multimodal corpora

In comparison to text corpora, the amount of resources that include other modalities is relatively small. However, during the last decade several important databases for speech recognition, speech synthesis and speech-to-speech translation have been built, most within the context of publicly funded projects.

The majority of speech resources for Basque have been developed for ASR applications. Alongside earlier data, such as SpeechDat, some recent collections are available for ASR in Basque. For example, SLR76 is a crowd-sourced multispeaker high-quality speech dataset that contains about 14 hours of HI-FI recordings in Basque; Common Voice 7.0, part of the Mozilla Common Voice initiative, includes an additional 132 total hours of recorded speech (91 of them validated); and the dataset King-ASR-825, known as the Spanish Basque Speech Recognition Corpus, contains 50 hours of audio recordings in Basque for mobile platforms.

For high-quality speech synthesis, large datasets obtained from a single speaker (typically a professional speaker) are needed. There are currently no public datasets of this sort available for commercial use in Basque. However, smaller datasets developed by research groups at the UPV/EHU are on hand for research in the field of speech synthesis, such as Abiadura (a database of sentences recorded at slow, normal and fast speech-rates) and Ahoemo2/Ahoemo3 (two emotional speech databases with 500 and 700 sentences for each emotion recorded by a total of four speakers). Additionally, Ahoemo1 is an emotional database which includes also video recordings.

Speech-to-speech translation, a new research area that requires bilingual data, has made some inroads with respect to Basque. The Mintzai-ST corpus (Etchegoyhen et al., 2021), for example, is a bilingual Basque and Spanish dataset obtained from parliamentary sessions of the Basque Parliament over eight years. It contains parallel speech-text sentences both for the pair Spanish – Basque (around 180,000 sentences or 480 hours) and for the pair Basque-Spanish (around 88,000 sentences or 190 hours).

Lexical/conceptual resources

Among the approximately 40 lexical and conceptual resources identified, 37% correspond to lexicons, 20% to dictionaries or thesauri, 17% to ontologies and monolingual or multilingual wordnets and 12% to terminological resources. The Egungo Euskararen Hiztegia (Dictionary of Contemporary Basque) and the Orotariko Euskal Hiztegia (Basque General Dictionary) count as two of the most important Basque dictionaries. In addition, euLex and the Euskararen Datu-base Lexikala(Lexical Database for Basque) are two noteworthy lexical databases needed for the automatic treatment of Basque. Both offer lexical support of Basque spellers, morphological analysers and lemmatisers EUSLEM.

With regard to wordnets and ontologies, three variants of wordnet for Basque stand out: EusWordNet, the Multilingual Central Repository 3.0 and SLI Galnet (which includes Galician). To these may be added the Predicate Matrix, a new lexical resource resulting from the integration of multiple sources of predicate information, including FrameNet, VerbNet, PropBank, WordNet and ESO.

Various other databases with specific knowledge are also notable: SLI Termoteca (tems), Konbitzul (translation of Spanish – Basque Multiword Expressions), the Basque Verb Index, e-ROLda (a Basque predicate analysing tool), sentiment Lexicons for 81 Languages (Sentiment Polarity Lexicons), multi-languages stopwords and patterns of frequency in the Basque Lexicon (Euskal Hiztegiaren Maiztasun Egitura, EHME/PFBL).

Models and grammars

Some Basque language models and a grammar are featured in the ELG collection. The available language models may be divided into monolingual and multilingual. Among the former is BERTeus,[17] a Basque language model pretrained on crawled news articles from online newspapers and the Basque Wikipedia. BERTeus improved state-of-the-art results for PoS tagging, NER, sentiment analysis and topic classification (FastText and Flair embeddings are also provided using the same Basque Media Corpus). The latter include IXAmBERT,[18] a multilingual pretrained language model for English, Spanish and Basque. The training corpora are composed of the English, Spanish and Basque Wikipedias, together with crawled news articles from Basque online newspapers. In contrast, the language-agnostic BERT Sentence Encoder (LaBSE)[19] is a BERT-based model trained for sentence embedding for 109 languages.

4.2 Language Technologies and Tools

The available tools and services for Basque span a wide range of applications from spellcheckers to speech processing and translation technologies. However, no tools or services for information extraction and retrieval, language generation and summarisation or human-computer interaction (HCI) are listed. The most representative tools and services for each LT area are as follows:

Text Analysis

Various linguistic processors and tools for Basque are ready-to-use. For example, ixaKat[20] and IXA-pipes[21] are a modular set of NLP tools for Basque with an input/output format that is in a NAF format. Thus, interaction between tools from both sets in the same processing

[17] https://huggingface.co/ixa-ehu/berteus-base-cased
[18] https://huggingface.co/ixa-ehu/ixambert-base-cased
[19] https://huggingface.co/setu4993/LaBSE
[20] https://ixa2.si.ehu.eus/ixakat/index.php?lang=en
[21] https://ixa2.si.ehu.eus/ixa-pipes/

pipeline is feasible. As a result, it is possible to create a pipeline containing a tokeniser, morphological analyser and PoS tagger, dependency parser, semantic labeling tool, coreference resolution tool and NER tagger for Basque. Similarly, pipelines for Basque may be constructed utilising UDPipe, a trainable pipeline which performs sentence segmentation, tokenisation, PoS tagging, lemmatisation and dependency parsing in multiple languages. Other types of linguistic processing are also available. UKB,[22] for instance, offers a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a preexisting knowledge base, the RST partial parser for Basque[23] allows for the detection of a text's central units within the framework of rhetorical structure theory, and Analhitza[24] favours the use of linguistic information in Humanities research by a offering a tool to explore and extract linguistic information from large corpora.

Spellcheckers

Because of Basque's relatively late standardisation, spellcheckers have historically been crucial tools to facilitate its use. In this context, there are currently three spellcheckers: Xuxen,[25] Hobelex[26] and IDITE.[27]

Speech Processing

There are two major Text-to-Speech (TTS) engines developed in the Basque Country to read texts with high-quality synthetic voices either in Basque or Spanish: AhoTTS[28] and Aditu.[29] Regarding Automatic Speech Recognition, Elhuyar Fundazioa offers a speech recognition service for Basque.[30] Google's Cloud Speech-to-Text is available for Basque, but only with the default and the command and search models.[31] There are no additional enhanced models available for Basque as there are for English, French or Spanish. There is no option for using Google's Cloud Text-to-Speech in Basque. Amazon does not include Basque in their TTS service, Amazon Polly, or in their Automatic Speech Recognition (ASR) service, Amazon Transcribe.

Translation Technologies

Besides the well-known Google Translate, there are four locally developed neural systems that provide high-quality translation. Three of these (elia,[32] batua[33] and lingua[34]) are provided by three separate companies and one (itzuli[35]) by the Basque Government. All four technologies translate between Basque and Spanish. But, itzuli and batua translate between 4 languages (Basque, Spanish, English and French) and elia translates between 6 languages (Basque, Spanish, English, French, Catalan and Galician). These translation technologies are

[22] https://ixa2.si.ehu.eus/ukb/
[23] https://ixa2.si.ehu.eus/rst/tresnak/rstpartialparser/
[24] https://ixa2.si.ehu.eus/clarink/analhitza.php
[25] http://xuxen.eus/eu/home
[26] https://uzei.eus/online/hobelex/
[27] https://uzei.eus/online/idite/
[28] https://aholab.ehu.eus/tts/
[29] https://www.euskara.euskadi.eus/r59-4734/es/contenidos/informacion/ahotsaren_sintesia/es_8543/sintesis_voz.html
[30] https://aditu.eus/inicio
[31] https://cloud.google.com/speech-to-text/docs/languages
[32] https://elia.eus
[33] https://www.batua.eus
[34] https://lingua.eus
[35] https://www.euskadi.eus/itzuli/

not simply offered through a webpage; almost all provide extra functionalities as well, such as the option to access the technology via a toolbar, ready-to-use libraries, translation plugins, mobile applications and translating documents with different formats (for example Microsoft Office and documents).

Other tools and services

The following specialised resources are worth adding to the aforementioned tools and services: LeXkit, a software for creating generic XML-based dictionaries that is being used for simultaneous management of the digital and printed versions of the Cuban Scholar Dictionary. EUSKALTERM, a public service for terminological queries; Termkate, a service that assists in the elaboration and publication of specialised multilingual knowledge resources; Adizkitegia, an app that helps users search and conjugate verbs in Basque; BertsolariXa, which looks for words ending in a given rhyme, and DiaTech, a web tool for analyzing and visualising linguistic variation.

In sum, although most basic LT tools are available for Basque, a significant gap remains between Basque and other languages in terms of data. The mC4 Multilingual Dataset, for instance, offers 10.401 Gb for English, 1.613 Gb for Spanish (6 times smaller), and only 5 Gb for Basque (2,000 times smaller).[36] Similarly, the original BERT language model for English was trained using a Google Books corpus that contains 155 billion words in American English and 34 billion words in British English. This means that the English corpus was almost 500 times larger than its Basque equivalent (384 million words) in 2020.[37] This difference is also observed in speech resources. As a case in point, Common Voice provides 2015 validated hours of speech for English, 377 for Spanish and only 91 for Basque. Similarly, there is little domain-specific data in Basque. If we wish to fine-tune models to specific domains in order to perform better, domain-specific corpora are required and thus an effort should be made in this respect. This handful of examples underlines the endemic digital inequality that exists in LT, although one bright spot for languages with few resources, such as Basque, is that pretrained monolingual and multilingual language models have proven quite useful in NLP tasks, even when based on a far smaller corpus (Agerri et al., 2020).

Sign languages in the Basque Country are based on those of Spain and France. Yet, while there is no independent Basque Sign Language, there are many dialectical elements present in the Basque Country and sign language in the Spanish Basque Country can vary between 10 and 30%[38] with respect to Spanish Sign Language: mutual understanding is achieved, but some signs differ.

4.3 Projects, Initiatives, and Stakeholders

Spain has had a well-funded plan for LT in place since 2015[39] that exists alongside the Coordinated Plan on AI[40] and the Spanish strategy R+D+i for AI.[41] A few autonomous communities, such as Catalonia and Galicia, have developed plans for their respective languages, but as of now there is no equivalent plan in the Basque Country for the Basque language. In the

[36] https://github.com/allenai/allennlp/discussions/5265
[37] https://www.ehu.eus/ehusfera/ixa/2020/09/30/ixambert-good-news-for-languages-with-few-resources/
[38] https://www.berria.eus/paperekoa/1881/027/001/2020-06-12/keinu-telebista.htm
[39] 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
[40] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence
[41] http://www.ciencia.gob.es/portal/site/MICINN/menuitem.26172fcf4eb029fa6ec7da6901432ea0/?vgnextoid=70fcdb77ec929610VgnVCM1000001d04140aRCR

French area of the Basque Country, the Euskal Hirigune Elkargoa[42] institution promotes a digital agenda, as does the IKER[43] research centre located in Bayonne, the sole laboratory in France that specialises in Basque Studies.

This differs south of the border in the Spanish Basque region. While the Chartered Community of Navarre currently has no strategic line in this area, the Basque Autonomous Community has fostered LT since 2002 through the Etortek and ElkarTek Industry Programmes. The Hizking21 (2002-2004), Anhitz (2006-2008), Berbatek (2009-2011), Ber2Tek (2012-2014), ElkarOla (2015-2017), Modela (2018-2019), BerbaOla (2018-2019) and MintzAI (2019-2020) projects have all resulted from these initiatives and several resources for Basque were created within this context. Participants include universities, research centres and several companies. By way of example, the Berbaola project (2018-2019) involved the University of the Basque Country (UPV/EHU), the Elhuyar Foundation, and the companies Tecnalia and Vicomtech. Nevertheless, despite these efforts, support for LT has declined in local science and technology plans. Indeed, these technologies faded into the background in the 2015-2020 plan.

This perception is unfortunate, given that over the last five years several language applications have emerged across Basque's technological landscape that could prove to be a catalysts for encouraging the use of the language in the public sphere. These applications allow speakers of other languages to understand text and speech in Basque and Basque speakers to understand texts in other languages. Besides the already mentioned and well-known local translation technologies, the following represent some of the most noteworthy:

- Content Translation.[44] This tool allows Wikipedia editors to create translations next to an original article utilising an automated process that copies text across browser tabs and looks for corresponding wiki-links, wiki-categories, wiki-templates and programmed components, etc. The deep intrinsic multilingualism of Wikipedia and Wikidata allows for easy translation of all languages that appear in Wikipedia article infoboxes. Content Translation offers translation to and from Basque using elia.eus, Google Translate or Yandex.
- Aditu.[45] The Aditu web service recognises both Basque and Spanish speech. It can obtain high-quality instant transcriptions, automatic generation of subtitles, and direct transcription from a microphone. Transcriptions and subtitles may be edited through an online editing interface.
- Interprest.[46] The main goal of this system is to enable low-cost and portable interpretation services for different types of events. It is a wireless system based on a straightforward process of mobile phone communication: the interpreter's mobile phone sends audio through a small microphone and each attendant can use his/her own phone to listen to the simultaneous translation. Interprest was a technological platform powered by Donostia/San Sebastián 2016, the foundation put in place to implement the cultural program for San Sebastián's turn as European Capital of Culture.
- Bidaide.[47] This web service allows tour takers to read or listen to descriptions and general information about sites on their own mobile and in their own language (Cortes et al., 2018), a technology that simultaneously provides accessibility for the visually impaired. MT is employed to translate texts, while speech synthesis is utilised to produce audio materials.

[42] https://www.communaute-paysbasque.fr/la-communaute-pays-basque
[43] https://iker.cnrs.fr/?lang=en
[44] https://www.mediawiki.org/wiki/Content_translation
[45] https://aditu.eus
[46] https://talaios.coop/2016/09/interprest/
[47] http://bidaide.elhuyar.eus

In addition to these language applications, there are various local research groups and institutions that study or develop technology for the Basque language. The most common services offered by these organisations are machine translators, corpora and speech recognition systems. The main LT providers include:

- The Basque Center for Language Technology (HiTZ).[48] HiTZ is composed of Ixa[49] and Aholab,[50] research groups at the University of the Basque Country (UPV/EHU).
- UZEI[51] – Terminologia eta Lexikografia Zentroa, a non-profit organisation.
- Euskara Institutua,[52] Institute for the Basque Language at the University of the Basque Country (UPV/EHU).
- Elhuyar Fundazioa,[53] a private non-profit organisation.
- Euskaltzaindia,[54] the Royal Academy of the Basque Language.
- The Basque Government.
- VICOMTECH,[55] an applied research centre.
- The Basque Wikipedia.

As is normal for a language with a status akin to Basque, most LT providers are locally based in the Basque Country and, likewise, most of the resources for Basque have been produced by publicly funded research groups at the University of the Basque Country or other public entities. There are also a few companies involved in publicly funded projects. Regrettably, however, resources resulting from these projects have not always been open-sourced and greater care must be taken to ensure resources resulting from public funding are publicly available. That said, it should be noted that the Langune cluster brings together leading agents of Basque LT from the public and private sectors. Langune sets out to foster, consolidate and unite the Basque Country's Language Industry. Its main goal is to enhance the competitiveness and visibility of the sector, as well as the association's members, through management excellence, co-operation, innovation, technology development and internationalisation.

Finally, we should not fail to mention the European CLARIN infrastructure (Krauwer and Hinrichs, 2014). CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources. Although Basque has around 500 resources accessible from the infrastructure, the fact that Spain is not a member of CLARIN limits Basque's capacity to develop more resources and make them available to researchers. The opportunity to take an active role in the infrastructure would result in the creation of new resources as well as help in the proper maintenance of existing ones.

[48] http://www.hitz.eus
[49] http://www.ixa.eus
[50] https://aholab.ehu.eus/aholab/
[51] https://uzei.eus/en/
[52] https://www.ehu.eus/en/web/eins
[53] https://www.elhuyar.eus/en
[54] https://www.euskaltzaindia.eus/en/
[55] https://www.vicomtech.org/en

5 Cross-Language Comparison

The LT field[56] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results never seen before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[57] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

[56] This section has been provided by the editors.

[57] Tools tagged as "language independent" without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[58]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises of more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[59] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[60]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

[58] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

[59] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

[60] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

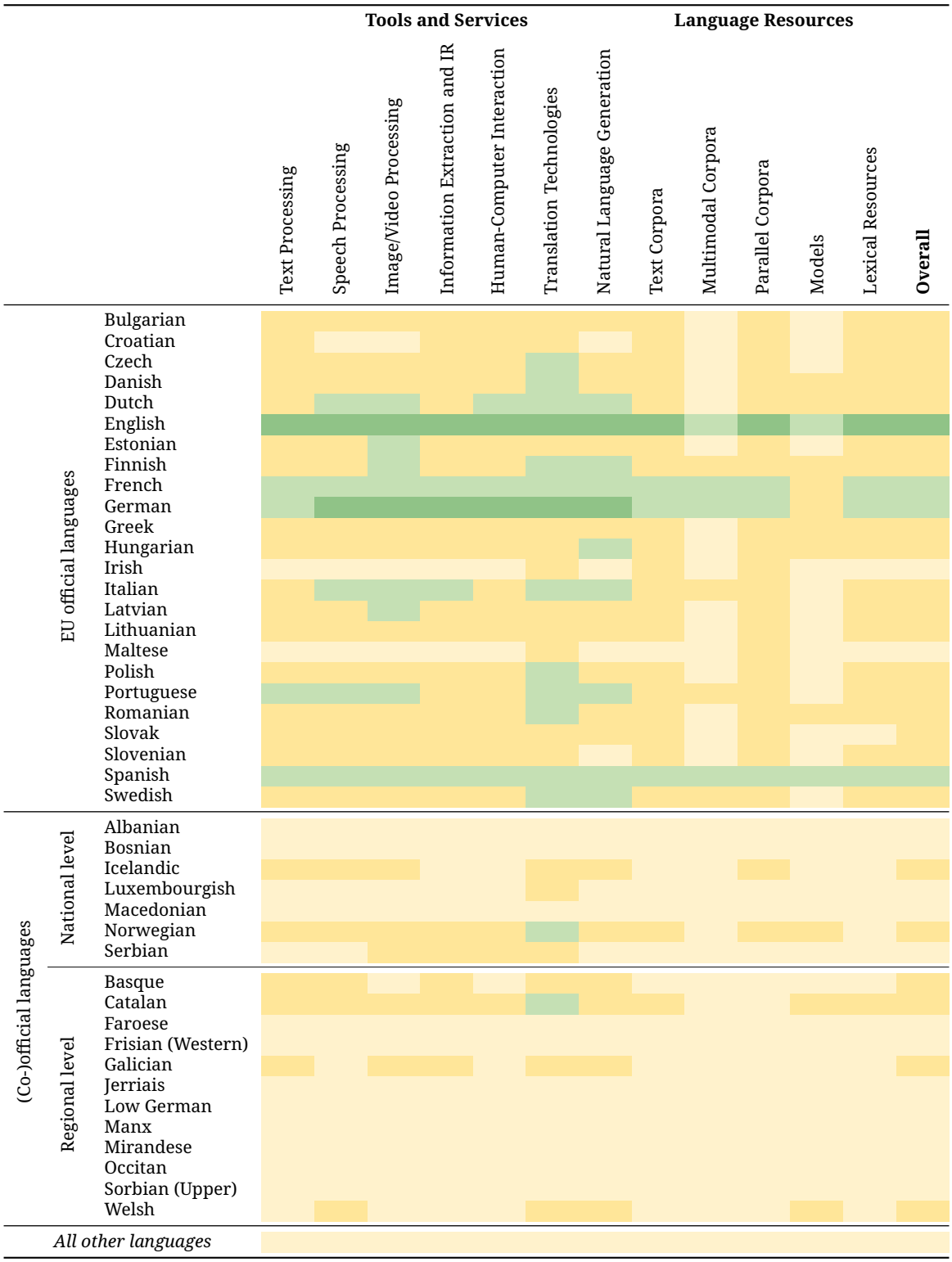

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[61] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 2 visualises our findings.

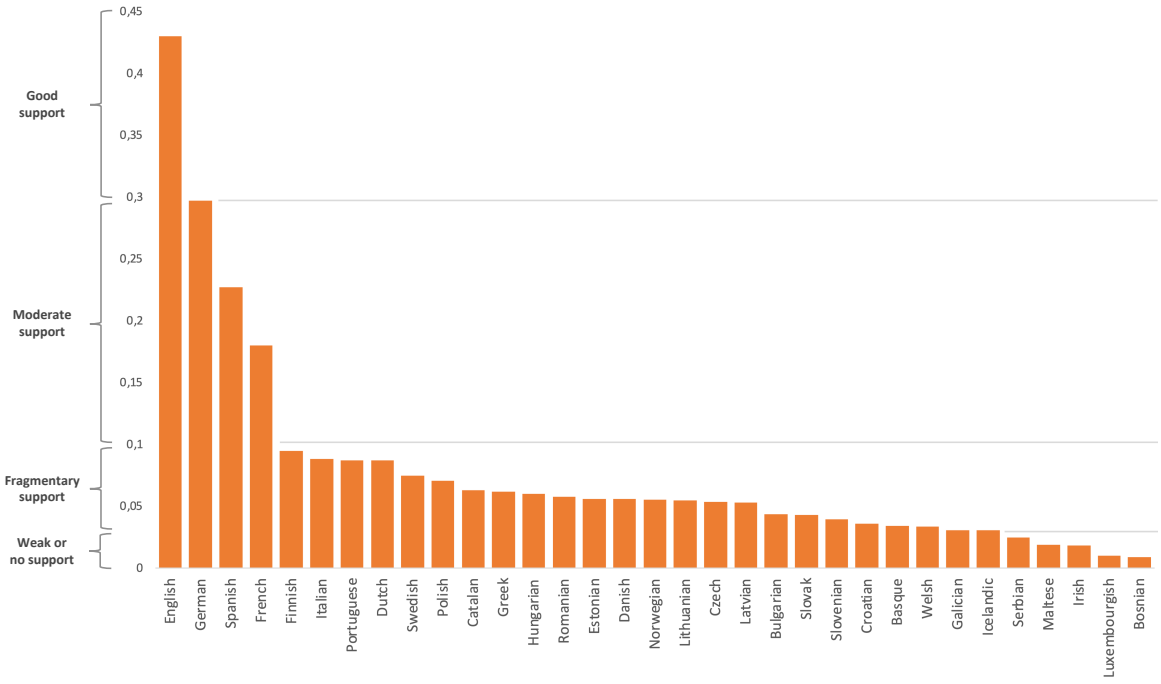

Figure 2: Overall state of technology support for selected European languages (2022)

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning,

[61] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evident in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

The advent of the Digital Age in the Basque Country has arrived at a moment when Basque's linguistic terrain is uneven. Its various dialects are spread across a region in which one's daily encounter with the language can differ sharply; it spans two European states and exists in multiple jurisdictions that afford it disparate levels of protection and status. The incremental and widely successful societal assimilation of Standard Basque over the past half century has helped to ameliorate Basque's disjointed linguistic body by providing a lingua franca for Basque speakers throughout the territory and abroad. The same also aided in nurturing a language community that is now able to develop, disseminate, and exploit Language Technology. This collaborative work in natural language processing has resulted in state-of-the-art technology for Basque and a solid foundation on which to innovate now and in the future. The significance of this potential should not be minimised given that LT is recognised as an additional means to support languages and propel their revitalisation in ways other approaches cannot. Be that as it may, Basque's ultimate digital fate will be predicated on the continuing capacity to cultivate the effective and high-quality LT solutions that are imperative for the everyday digital use of Basque.

Furthermore, Basque's sociolinguistic reality proves to be an interesting test case for NLP and LT. Because the digital space both reflects this reality and provides its own unique sphere within which to engage with Basque, distinctive relationships with the language and the people who utilise it may be forged that are unbounded by the constraints of locality. And it is evident that this supralocal sociolinguistic digital tapestry is being woven together at great speed. Basque enjoys a firm presence online and in social media, where it is utilised across virtually the entire spectrum of digital life. By the same token, it is clear that demand for digital resources and tools is significant and that the Basque community is taking full advantage of available language technologies in their everyday lives. This activity, coupled with the wide range of data resources and tools that exist for Basque, points to the current online vitality of the language and augurs well for its future digital survival.

Nevertheless, while Basque's digital condition may not be qualified as endangered, let

alone critically endangered, it remains vulnerable in certain ambits. More work must be done, e. g., to deepen Basque's integration into social network applications, expand its use in business and employment-oriented services, and extend its reach into entertainment-related products. Moreover, although the breadth of Language Technology for Basque is adequate when measured in isolation, there are significant gaps in the availability of language data and tools that must be addressed so that research may be improved and better applications developed for commercial use. Some of the more obvious lacunae include a lack of sufficient multimodal corpora, public datasets, and advanced language models for Basque. While it is true that pretrained monolingual and multilingual language models are employed to great effect in a variety of NLP tasks, a dearth of domain-specific data in Basque continues to hinder the ability to fine-tune models for those domains. This is an area that not only requires attention with respect to Basque, but also suitably underscores the prevailing chasm in LT between the most utilised online languages, such as English, and those with far fewer digital resources. In light of this, it is as understandable as it is troublesome that a high percentage of Basque speakers continue to meet with obstacles when going about their online lives, too often finding it easier or even necessary to rely on other, more widely available, languages for determined services and information. This prima facie case of linguistic inequality, not limited to Basque alone, does not bode well for the outlook of Europe's cultural heritage.

Fortunately, a remedy may yet be found if action is taken now. Basque's digital health would benefit from bolder and nimbler Language Technology strategies at the European, national and regional levels that can increase the scope and volume of applications for under-resourced languages. Although Spain and the Basque Autonomous Community have invested in this area in conjunction with AI, more must be done to establish plans that are both committed to sustained funding streams and responsive to the dynamic nature of digital technologies, which will continue to represent a critical sphere in European research and development in the coming years. A program of this kind that prioritises Basque and is compatible with its European and Spanish counterparts must be designed and put in place by the Basque LT community. For these endeavors to succeed, it is essential that future LT plans strive to guarantee data and resources will be made publicly accessible whenever possible because the amount of available data will determine the quality of prospective applications. Licences that provide fewer restrictions in content creation should be more widespread so that greater amounts of linguistic data may be collected. Infrastructures and trained personnel are required to manage the influx of data and curate it for research and development. At one level, taking these steps will help ensure Language Technology continues to adapt to Basque's digital needs and keep pace with advances at the global level. At another, such a strategy would impart greater visibility to Language Technology and reinforce its vital role in enabling Basque to thrive in today's rapidly evolving sociodigital space.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your Text Representation Models some Love: the Case for Basque, 2020.

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Lore Agirrezabal. *The Basque Experience: Some Keys to Language and Identity Recovery*. Garabide Elkartea, Eskoriatza, Gipuzkoa, 2010. ISBN 978-84-613-6835-8.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

Itziar Cortes, Igor Leturia, Iñaki Alegria, Aitzol Astigarraga, Kepa Sarasola, and Manex Garaio. Massively multilingual accessible audioguides via cell phones. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation 2018*, page 349, 2018.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete Ugarte, Aitor Alvarez, Ander González-Docasal, and Edson Benites Fernandez. mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation. In *Proceedings of IberSPEECH 2020*, pages 190–194, 2021.

Maite Goñi. Basque language on the web: Making an impact. http://basquetribune.com/lost-in-translation/, 2013. Accessed: 2022-01-31.

Antton Gurrutxaga and Klara Ceberio. Report: Basque – a digital language? , 2017. Reports on Digital Language Diversity in Europe | Editors: Claudia Soria, Irene Russo, Valeria Quochi. The Digital Language Diversity Project (www.dldp.eu), funded by the European Union under the Erasmus+ Programme (GrantAgreement no. 2015-1-IT02-KA204-015090).

Inmaculada Hernáez, Eva Navas, Igor Odriozola, Kepa Sarasola, Arantza Diaz de Ilarraza, Igor Leturia, Araceli Diaz de Lezana, Beñat Oihartzabal, and Jasone Salaberria. *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 2012. URL http://www.meta-net.eu/whitepapers/volumes/basque. Georg Rehm and Hans Uszkoreit (series editors).

Steven Krauwer and Erhard Hinrichs. The clarin research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, 2014.

Itziar Laka. *A Brief Grammar of Euskara, the Basque Language*. University of the Basque Country, 1996. ISBN 84-8373-850-3. URL http://www.ehu.es/grammar.

Luis Mitxelena. Lengua comun y dialectos vascos. *International Journal of Basque Linguistics and Philology*, 15:291–313, 1981.

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

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Iñigo Urrutia. Régimen jurídico de las lenguas y reconocimiento de la diversidad lingüística en el tratado por el que se establece una constitución para europa. *Revista de Llengua i Dret, 42*, 42:231–273, 12 2004.

Iñigo Urrutia and Iñaki Lasagabaster. Language rights as a general principle of community law. *German Law Journal, Vol. 8, Nº. 5, 2007, pags. 479-500*, 8, 05 2007. doi: 10.1017/S2071832200005733.

Koldo Zuazo. *Euskalkiak*. Elkar, 2014. ISBN 978-84-9027-238-1. URL http://euskalkiak.eus/en/sailkapenak.php.