Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 53-61 recibido 12-12-2022 revisado 02-02-2023 aceptado13-02-2023

Measuring language distance for historical texts in Basque

Cálculo de distancia lingüística para textos históricos en euskera

Ainara Estarrona[1], Izaskun Etxeberria[1], Manuel Padilla-Moyano[2] and Ander Soraluze[1]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]University of the Basque Country UPV/EHU
{*ainara.estarrona, izaskun.etxeberria, manuel.padilla, ander.soraluze*}*@ehu.eus*

Abstract: Measuring distance between languages, dialects and language varieties, both synchronically and diachronically, is a topic of growing interest in NLP. Based on our Syntactically Annotated Historical COrpus in BAsque (SAHCOBA) and previous work in perplexity-based language distance proposed by Gamallo, Pichel and Alegria (2017, 2020), we have compared historical corpora with current texts in the standard variety and calculated the language distances between them. As the standard Basque is based on the central dialects, the starting hypothesis is that the oldest texts and the dialects on the extremes will be the most distant. The results obtained have largely confirmed the thesis of traditional dialectology: peripheral dialects show a strong idiosyncrasy and are more distant from the rest.
Keywords: Language distance, dialectology, historical texts, perplexity.

Resumen: Medir la distancia entre diferentes lenguas, dialectos o variantes de lengua, tanto sincrónica como diacrónicamente, es un área de interés creciente dentro del PLN. Basándonos en el corpus histórico sintácticamente anotado del euskera (SAHCOBA), y en el trabajo previo realizado por Gamallo, Pichel y Alegría (2017, 2020) en relación con la distancia entre lenguas basada en perplejidad, hemos comparado textos históricos en euskera con textos actuales y hemos calculado la distancia entre ellos. Dado que el euskera estándar se basa en los dialectos centrales, la hipótesis inicial es que los textos más antiguos, así como los textos de los dialectos periféricos serán los más distantes. Los resultados obtenidos confirman de forma contundente las tesis propuestas por la dialectología tradicional: los dialectos periféricos muestran una fuerte idiosincrasia y su distancia respecto al estándar es mayor que la del resto de dialectos.
Palabras clave: Distancia lingüística, dialectología, textos históricos, perplexity.

1 Introduction

Measuring distance between languages, dialects and language varieties, both synchronically and diachronically, is a topic of growing interest in NLP. Under the BIM and SAHCOBA projects[1] we have collected the most relevant historical texts in Basque written in different dialects. As a next step, we hoped to quantify how different these texts are as a means to confirm and modulate theories about the historical and dialectal development of the language. For this purpose, we have compared the historical corpora with current texts in the standard variety and calculated the distances between them. As the standard is based on the central dialects, the starting hypothesis is that the oldest texts and the dialects on the extremes will be the most distant.

For measurements we have used information theory based on *perplexity*. Perplexity-based measures have been employed successfully for language identification (Gamallo et al., 2016), to calculate distance between

[1]Basque in the Making (BIM): A Historical Look at a European Language Isolate project (ANR-17-CE27-0011 - BIM, Agence Nationale de la Recherche, France) and the Syntactically Annotated Historical COrpus in BAsque (SAHCOBA, RTI2018-098082-J-I00) project (Ministry of Science and Innovation (MICINN), Spain).

ISSN 1135-5948 DOI 10.26342/2023-70-4 ©2023 Sociedad Española para el Procesamiento del Lenguaje Natural

languages (Gamallo, Pichel, and Alegria, 2017b), and to quantify the diachronic distance in a language (Pichel, Gamallo, and Alegria, 2018). The software is open and, being an unsupervised method, only raw historical corpora are required.

The remainder of this paper is organised into several sections. Specific features of the Basque language and its dialects are introduced in Section 2. Section 3 is devoted to describing the corpus, while Section 4 covers why and how perplexity is applied. In Section 5 we detail the design of the experiments and briefly discuss the results in Section 6. Finally, Section 7 outlines our conclusions and possible future work.

2 *Basque language and dialects*

As a non-Indo-European language, indeed an isolate, Basque grammar differs considerably from that of the neighbouring languages. Basque is agglutinative, head-final, pro-drop, and usually assumed to be a Subject-Object-Verb (SOV) type language (de Rijk, 1969), but it is also described as having 'free word order', meaning that the order of phrases in the sentence can vary (Laka, 1996). Moreover, the Basque language exhibits a high level of dialectal fragmentation over an area of 10,000 km^2. The dialectal split began in the early Middle Ages (Mitxelena, 1981), and over the past 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.

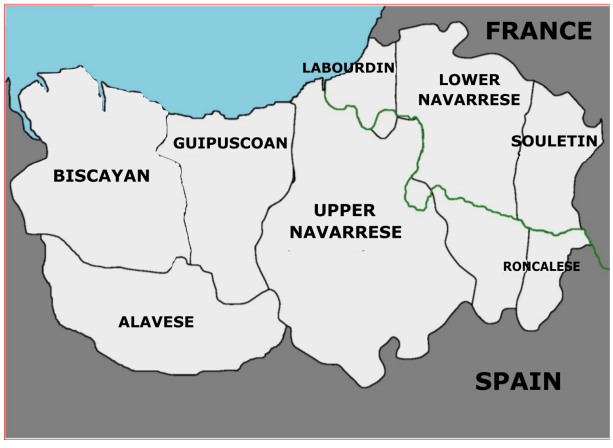

Figure 1: Historical Basque dialects. Alavese and Roncalese are extinct varieties. The green line represents the French-Spanish border.

At present, Zuazo (2014) distinguishes between five main Basque dialects: the Western dialect, traditionally called Biscayan, the Central dialect, traditionally known as Guipuscoan, the Navarrese dialect, the Navarrese-Labourdin dialect, and the Souletin dialect[2].

Figure 2: The five main dialects of the Basque Language (after Zuazo (2014)).

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 and historical dialects differ from standard Basque to varying degrees.

The distance between Basque dialects has often been a matter of discussion among linguists, but a scientific consideration of this problem requires some operational procedure for quantifying linguistic distance (Mitxelena, 1981). To our knowledge, the only attempts to quantify the differences between dialects have been based on dialectometry (Séguy, 1973) and have been carried out by linguists from the Eudia[4] group at the University of the Basque Country, including Aurrekoetxea, Gaminde, and Videgain, among others (Aurrekoetxea, 1992; Aurrekoetxea and Videgain, 2009; Aurrekoetxea et al., 2019). Their research does not deal with historical dialectology, but with dialects spoken

[2]http://euskalkiak.eus/en/ezaugarriak.php
[3]https://www.euskaltzaindia.eus/en/
[4]http://eudia.ehu.es/en/home/

today, from a synchronic perspective. However, language history and dialectology must go hand in hand (Camino, 2008) since every historical text is by definition a dialectal one.

Biscayan and Souletin are the two dialects at the corners. As traditional dialectological studies attest, they display the greatest differences from the others and have the most marked idiosyncrasy. The case of Biscayan is particularly relevant, as in the past certain scholars claimed that there were only two dialects: Biscayan on the one hand, and the central-eastern dialect, which would include the rest of the dialects, on the other (Lacombe, 1924). Although Biscayan has indeed noticeable characteristics that are lacking in the other dialects, it is no less true that many of these idiosyncrasies are innovations due to its peripheral character. Bear in mind that the lateral areas, unlike the central ones, are not only repositories of archaisms but also a breeding ground for innovations fostered by the heat of languages from the surrounding area (Mitxelena, 1981).

As we have already said the standard Basque (*Batua*) is based on the central dialects (mainly Guipuscoan and Labourdin), from which we can deduce that the peripheral dialects are the most distant from the standard. The main goal of this paper is to quantify the distance of the different historical dialects from standard Basque in order to confirm (or reject) existing dialectological theses in a quantitative way, thus contributing to historical dialectology from computational linguistics. To carry out this work we are going to use perplexity-based measures (see Section 4).

3 Corpus

Basque's historical corpus is quite scarce compared to those of neighbouring languages. Moreover, the corpus is asymmetrical geographically and historically: most varieties have significant gaps in their textual history, and at certain periods we do not have written records for all dialect. Along with scarcity and asymmetry, we must also mention homogeneity since, until the nineteenth century, works of a religious nature constituted more than 95% of the corpus (Lakarra, 1997). Most of these are also simple texts (doctrines, catechisms, etc.), which conceal many characteristics of the language, as lexicon, morphology and syntax are constrained by the type of discourse (Ulibarri, 2013).

As mentioned above, two projects have been involved in the creation of the Basque annotated historical corpus: BIM and SAHCOBA. The BIM-SAHCOBA corpus needed be representative of all dialects with a written tradition. Therefore, in these two projects we decided to establish a philologically reliable corpus covering most of the textual production between the fifteenth and mid-eighteenth centuries (Estarrona et al., 2021). This, on the one hand, is the minimal span that includes regular attestations for all Basque dialects and, on the other, is representative of the divide between Archaic and Old Basque from early modern Basque (Gorrochategui, Igartua, and Lakarra, 2018).

For the time being, we are creating a corpus of around one million words. Considering the issues associated with the written past of languages, especially in cases like Basque, this size is considered acceptable for a historical corpus (Claridge, 2009).

We have picked out nine works from the sixteenth and seventeenth centuries for our experiments. The oldest texts have been chosen because, although it is true that over time the dialects have become ever distant from each other (Mitxelena, 1981), it is no less true that the more recent the text, the closer it is to the standard variety. The main criterion for the choice of works was diversity of dialect. Thus, the texts selected are relevant to the history of Basque and that, in addition, reflect the main characteristics of each historical dialect. They are as follows [5]:

- Lazarraga's manuscript (1565)[6]
- *Iesus Krist Gure Iaunaren Testamentu Berria* (New Testament), Leizarraga (1571)
- *Dotrina Christiana. Bigarren impressionean debocionozco othoitz eta Oracino batçuez berreturic*, Materra (1617)
- *Gvero bi partetan partitua eta berecia*, Axular (1643)
- *Iesusen imitacionea*, Pouvreau (1669)

[5] The works are arranged by dialect and chronologically within each dialect.

[6] We will use the following abbreviations in the tables: Lazarraga=Laz; Leizarraga=Lç; Materra=Mat; Axular=Ax; Pouvreau=SP; Beriain=Ber; Kapanaga=Cap; Tartas=Tt: and Belapeire=Bp.

- *Tratado de como se ha de Oyr Missa*, Beriain (1621)
- *Exposición breve de la doctrina christiana*, Kapanaga (1656)
- *Onsa hilceco bidia*, Tartas (1666)
- *Catechima laburra eta Jesus-Christ Goure ginco jaunaren ecagutcia*, Belapeire(1696)

Table 1 shows the description of the selected works:

Author	Century	Dialect	Size
Laz	XVI	Alavese	12,072
Lç	XVI	Labourdin	73,906
Mat	XVII	Labourdin	16,323
Ax	XVII	Labourdin	90,029
SP	XVII	Labourdin	46,363
Ber	XVII	Upper Navarrese	14,995
Cap	XVII	Biscayan	11,408
Tt	XVII	Lower Navarrese	34,505
Bp	XVII	Souletin	23,735

Table 1: Works chosen for the experiments, century and year in which they were written, dialect and size.

Note that we consider Lazarraga's work as written in the Alavese dialect, which is an extinct variety. From today's perspective, Lazarraga's text is commonly considered a western dialect (Pagola, 2006), a classification created by Zuazo (2014).

As can be seen in Table 1, and due to the aforementioned asymmetry of the corpus, we do not have works in all the dialects for each century. As a case in point, there are no sources for the Guipuscoan dialect until the middle of the eighteenth century[7]. An interesting avenue for future work would be to quantify the distance of this central or Guipuscoan dialect with respect to the standard since in principle it should be the closest to it, both because the standard is based on the central dialects and because the texts in Guipuscoan are much more recent than those analysed in this paper.

Finally, we should mention the philological work that we carried out to begin from the best possible transcription of the works that we treated. We compared the transcriptions with their facsimiles (and/or with reliable critical editions) and, depending on the quality of each one, opted for one of the following: i) to correct the transcript, or ii) to create a new one. This task is highly time-consuming, but necessary to ensure our corpus is based on reliable versions of historical texts. The main criterion behind this philological effort is modernising the spelling — not to be confused with the adoption of present-day standard Basque orthography. In our corpus, the updating of spelling preserves the phonological shape of each text. For instance, Eastern Basque dialects have a set of aspirated plosive phonemes *ph, th, kh* that is not represented in the spelling system of standard Basque. However, we decided to maintain this phonological feature in the transcription of the texts (Estarrona et al., 2021).

4 Language distance. Perplexity

The main approaches to measuring language distance for historical or dialectal texts compare phonetic forms (Kondrak, 2005), "but some researchers have argued against the possibility of obtaining meaningful results from crosslingual comparison of phonetic forms" (Singh and Surana, 2007).

In computational linguistics, language models have been utilised for this purpose. The models and the calculation of crosslingual similarity are often based on word co-occurrences (Liu and Cong, 2013; Gao et al., 2014; Asgari and Mofrad, 2016). Recently, Degaetano-Ortlieb and Teich (2018) have used relative entropy for the detection and analysis of periods of diachronic linguistic change.

Perplexity-based measures are related to entropy and have been employed successfully for language identification (Gamallo et al., 2016), to measure the distance between languages (Gamallo, Pichel, and Alegria, 2017b), and to quantify the diachronic distance in a given language (Pichel, Gamallo, and Alegria, 2018). Basque appears in two of the experiments carried out by these authors, one comparing forty-four European languages (Gamallo, Pichel, and Alegria, 2017a) and the other selecting a handful of isolated languages to measure the distance between them (Gamallo, Pichel, and Alegria, 2020).

The method has been quite successful and,

[7]There are, however, small texts of few words collected in Michelena (1964), Sarasola (1983) and Satrustegi (1987).

in addition to language identification and historical linguistics (Scherrer, Samardžić, and Glaser, 2019; Zugarini, Tiezzi, and Maggini, 2020), has been used in other fields, including machine translation (Barrault et al., 2019), sociolinguistics (Chavula and Suleman, 2020) and sociology (Sant'Anna and Weller, 2020).

We use perplexity according to the methodology proposed by Pichel, Gamallo, and Alegria (2020) and the software they offer[8]. A language model's perplexity is defined as the inverse probability of the test text given the model. It is calculated comparing the n-grams (characters) of a text in one language/dialect with the n-gram model trained for another language/dialect (or between two historical periods of the same language). Lower perplexity would indicate lower distance between languages (or language periods). The comparison can be made in both directions because perplexity is a divergence with asymmetric values.

Due to the size of the historical texts, we have calculated the distance only in one direction, building the model for the standard language (larger corpus) and using historical texts as test corpus. In order to have comparable results, we configured the distance and the corpora with the same hyper-parameters as those used by the authors: 7-grams.

5 Design of the experiments

As discussed in the previous section, at least one corpus of standard Basque is required to carry out the experiments that measure distance between today's standard and the various historical dialects of the language. The corpus of historical Basque has been described in Section 3.

Regarding standard Basque, we believed it best to ensure the subject of the standard Basque text (or texts) was 'similar' to that of the historical texts. As most of the latter are religious texts, we elected to use a digital version of the Bible written in standard Basque[9]. However, to determine whether the subject of the text is important when measuring distance between historical and standard Basque, we also relied on a non-religious second corpus written in standard Basque (EPEC, a reference corpus for the processing of Basque (Aduriz et al., 2006)).

In order to obtain the n-gram model of standard Basque (Bible or EPEC) we used the software previously mentioned in section 4. This software carries out a preprocessing step before obtaining the final 7-gram model that consists on: (1) text cleaning: figures and punctuation marks removed, uppercase letters converted to lowercase, extraneous characters eliminated, and so on; (2) tokenization. This preprocess reduced the size of the Bible and EPEC corpora, leaving them at 514,443 and 291,228 words, respectively. After preprocessing step, two n-gram models of standard Basque are trained, one utilising the Bible and the other using EPEC corpora.

The same cleaning process applied to the Bible and EPEC was repeated for each historical text (the resulting sizes of the corpora appear in Table 3). Given that the texts vary significantly in length, from 11,408 to 90,029 words, and because we wished to compare the distances between different dialects, we decided to conduct two experiments for each: one utilised a randomly selected predetermined portion of content similar in size to the shortest text (11,500), while the other used the full text. The results appear in Tables 2 and 3. In addition to cleaning process, n-grams of each historical text were also calculated. These n-grams were then used to compare with the previously obtained n-gram models of standard Basque in order to obtain perplexity-based distance.

6 Results and discussion

In this section we will present and analyse the results. Tables 2 and 3 contain the findings obtained in the two experiments previously described.

6.1 Results

Table 2[10] displays the results obtained for samples of similar size extracted from every source, while in Table 3 we see results for the complete text. As may be appreciated, the findings for the sample experiment differ little from those obtained from that done on the complete work, nor do they vary when we compare the historical texts with the Bible or with a corpus of a different subject matter, such as EPEC.

[8] https://github.com/gamallo/Perplexity

[9] https://www.biblija.net/biblija.cgi?l=eu

[10] In the following tables we will use abbreviations for dialects: Alavese=Al; Labourdin=L; Upper Navarrese=UN; Biscayan=B; Lower Navarrese=LN, and Souletin=S.

Cent.	Auth.	Dial.	Size	Dist. Bible	Dist. EPEC
XVI	Laz	Al	11,501	7.84	7.58
XVI	Lç	L	11,501	6.09	6.32
XVII	Mat	L	11,503	4.93	4.91
XVII	Ax	L	11,507	4.69	4.53
XVII	SP	L	11,510	4.77	4.79
XVII	Ber	UN	11,520	5.52	5.29
XVII	Cap	B	11,408	7.18	6.68
XVII	Tt	LN	11,503	6.07	5.68
XVII	Bp	S	11,500	11.33	9.48

Table 2: The two perplexity values for each historical text based on a portion of similar size. The first value was obtained using a contemporary version of the Bible written in standard Basque. The second was attained using the EPEC corpus.

Cent.	Auth.	Dial.	Size	Dist. Bible	Dist. EPEC
XVI	Laz	Al	12,072	7.83	7.58
XVI	Lç	L	73,906	6.13	6.28
XVII	Mat	L	16,323	4.94	4.90
XVII	Ax	L	90,029	4.72	4.57
XVII	SP	L	46,363	4.74	4.78
XVII	Ber	UN	14,995	5.58	5.31
XVII	Cap	B	11,408	7.18	6.68
XVII	Tt	LN	34,505	6.06	5.66
XVII	Bp	S	23,735	11.43	9.54

Table 3: The two perplexity values for each historical text. In this case, the entirety of each historical text was utilised in the experiment. The two corpora are the same as in the previous experiment.

Unsurprisingly, the works closest to the standard are those belonging to the central dialects: Labourdin, Upper Navarrese and Lower Navarrese. Moreover, the Labourdin texts are somewhat closer than their Navarrese counterparts, which was to be expected since the standard was essentially built on Guipuscoan and Labourdin, as mentioned above. Interestingly, Leizarraga's work, despite being essentialy written in Labourdin, departs somewhat from the standard. Leizarraga was presumably a native speaker of Lower Navarrese and he noted that he translated the Bible so that it would be understood by most readers, i.e. in a sort of northern koiné. We should also mention that Leizarraga's work is one of the oldest and it is therefore logical that it differs most from the standard. These factors may help explain why the results demonstrate a greater distance between this text and the standard. The case of Tartas's contribution may be similar, given that despite being a Souletin writer, he attempted to move away from the Souletin dialect in order to address a wider public.

The next most distant works from the standard are those by Lazarraga and Kapanaga, written in Alavese and Biscayan, respectively (both two western varieties). Once again, this result was expected. Because these varieties move away from the centre, they are more peripheral and, therefore, more distant from that standard variety.

Finally, the gap between the Souletin dialect and the standard should be highlighted. In view of the findings, we can clearly state that the work written in Souletin is the most distant from the standard. Yet again, this is an expected result since Souletin, like Biscayan, is a highly idiosyncratic peripheral variety. One of the most conspicuous features of this Souletin idiosyncrasy is the so-called sixth vowel "ü" (/y/), non-existing in the rest of dialects. We believe that it is this characteristic that makes the distance so quantitatively great. Tartas, however, opted against using a specific spelling for the sixth vowel in his works, or used it in a very defective way and perhaps this also helps explain why the distance is not as great as in other Souletin authors.

In short, we emphasize once more that the results obtained confirm our expectations and that they validate quantitatively what traditional dialectology affirms.

6.2 Comparing with other languages

Although the quantitative values of the distances are not directly comparable to similar experiments with corpora in other languages, we can consider whether the range of values that we have found (4.53 minimum and 11.43 maximum) coheres with those obtained in the diachronic study of other languages or in crosslingual comparisons.

Gamallo, Pichel, and Alegria (2017a) demonstrate that the distance between older English (sixteenth-eighteenth centuries) and today is 5.80, but that for previous centuries

(twelfth-fifteenth) it is up to 15.85 (original spelling in both cases). In the case of Portuguese, the measured values for the same periods in original spelling are 7.40 and 7.73, while for Spanish they are 5.97 and 8.02.

Thus, for the period between the sixteenth and eighteenth centuries, the values for the three languages are 5.80, 7.40 and 5.97. In our case, with the exception of Belapeire, the distances for most of the Basque historical texts are close to these figures.

With respect to distance between languages, Gamallo, Pichel, and Alegria (2017a) compute distances among 44 European languages using perplexity, yielding interesting figures that are close to the values we obtained:

- The smallest distances obtained are the Bosnian-Croatian distance (5) and the Portuguese-Galician distance (6).
- Within the 7-9 range are Bosnian-Slovene, Catalan-Spanish, Czech-Slovak and Portuguese-Spanish.
- The value for the Swedish-Danish distance is 12 and for Swedish-Norwegian 13.

Hence, the smallest diachronic distance between the Basque dialects and standard Basque is similar to the distance between the closest languages, just as the greater distances within Basque are similar to those between languages that are somewhat more differentiated.

7 Conclusions and Future Work

7.1 Conclusions

Utilising our Syntactically Annotated Historical Corpus in Basque and the previous work in language distance by Gamallo, Pichel, and Alegria (2017a), we have compared Basque historical corpora with current texts in the standard variety and calculated the language distances between them. Since the standard is based on the central dialects, the starting hypothesis is that the texts of the dialects of the extremes on the one hand, and the oldest on the other, will be the furthest away.

The results obtained have largely confirmed the thesis of traditional dialectology: peripheral dialects have a strong idiosyncrasy and are more distant from the rest. We have verified this by measuring the distance of all historical dialects from the standard Basque (built upon the central dialects).

We must not forget that these are initial experiments and that the findings, while significant, also raise further questions. For example, we hope to study in depth the effect of the spelling "ü" in Souletin texts, as it is involved in a series of morphophonological changes. In addition, another interesting avenue to pursue is the fact that Lazarraga's work is at the same distance from the standard as Kapanaga's written in the Biscayan dialect. Traditional dialectology tells us that Lazarraga's text is written in what is today known as the western dialect (as is Kapanaga's). But within that dialect, Lazarraga would correspond to a more eastern variety (closer to the central dialects) (Pagola, 2006). Therefore, one would expect that the distance with respect to the standard is not as great as in the case of the westernmost Biscayan.

We believe that this work opens up a new line of research in Basque dialectology and that the next step is to measure the distances of the historical dialects from each other to establish whether the results confirm this study's findings.

7.2 Future work

These first experiments and the results obtained encourage us to continue working along these lines. The next step will be to include all the works in the corpus in the experiments to see what occurs with varieties of the language that are not attested to until later, such as the case of the Guipuscoan dialect.

We would also plan to measure the distance between the different historical dialects, although we foresee that the scarcity of records will be a major roadblock. Once the distances between the historical dialects are calculated, it will be worthwhile to compile a contemporary corpus of the different dialects in order to measure the distances between them and compare the results with those obtained for the historical dialects. In this way, we will be able to test the thesis of traditional dialectology that the distance between Basque dialects increased over time. It nevertheless remains that, in some aspects, the spreading of standard Basque during the last decades favours dynamics of convergence between dialects.

Acknowledgments

We are very grateful to Pablo Gamallo of the University of Santiago de Compostela for his contributions to the development of the experiments. Special acknowledgment are due to José Ramón Pichel and Iñaki Alegría for their expertise in perplexity measure, to Ricardo Etxepare of the IKER UMR 5478-CNRS for his leadership in the BIM project and for always being committed to interdisciplinarity, and finally to Aritz Farwell for his help in revising the text.

This research has been partially supported by the Agence nationale de la recherche of France (ANR-17-CE27-572 0011-BIM); the Ministry of Science, Innovation, and Universities of Spain (RTI2018-573 098082-J-I00); and the Basque Government (IT1570-22).

References

Aduriz, I., M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. D. de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In *Corpus linguistics around the world*. Brill, pages 1–15.

Asgari, E. and M. R. K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 65–74, San Diego, California.

Aurrekoetxea, G. 1992. Nafarroako euskara: azterketa dialektometrikoa. *Uztaro*, 5:59–109.

Aurrekoetxea, G., I. Gaminde, J. L. Ormaetxea, and C. Videgain. 2019. *Euskalkien sailkapen berria*. UPV/EHU, Bilbao.

Aurrekoetxea, G. and C. Videgain. 2009. Le projet Bourciez: traitement géolinguistique d'un corpus dialectal de 1895. *Dialectologia*, 2:81–111.

Barrault, L., O. Bojar, M. R. Costa-Jussa, C. Federmann, M. Fishel, and Y. Graham. 2019. Findings of the 2019 conference on machine translation (WMT19). Association for Computational Linguistics (ACL).

Camino, I. 2008. Dialektologiaren alderdi kronologikoaz. *Fontes Linguae Vasconum (FLV)*, 108:209–247.

Chavula, C. and H. Suleman. 2020. Intercomprehension in retrieval: User perspectives on six related scarce resource languages. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 263–272.

Claridge, C. 2009. Historical corpora. In *Corpus linguistics. An International Handbook*, pages 242–259, Berlin, Germany.

de Rijk, R. 1969. Is Basque an SOV language? *Fontes Linguae Vasconum (FLV)*, 1:319–351.

Degaetano-Ortlieb, S. and E. Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33.

Estarrona, A., I. Etxeberria, R. Etxepare, M. Padilla-Moyano, and A. Soraluze. 2021. The first annotated corpus of historical basque. *Digital Scholarship in the Humanities*, 37(2):391–404.

Gamallo, P., I. Alegria, J. R. Pichel, and M. Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177.

Gamallo, P., J. R. Pichel, and I. Alegria. 2017a. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

Gamallo, P., J. R. Pichel, and I. Alegria. 2017b. A perplexity-based method for similar languages discrimination. *VarDial 2017*, page 109.

Gamallo, P., J. R. Pichel, and I. Alegria. 2020. Measuring language distance of isolated European Languages. *Information*, 11(4):181.

Gao, Y., W. Liang, Y. Shi, and Q. Huang. 2014. Comparison of directed and

weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393(C):579–589.

Gorrochategui, J., I. Igartua, and J. A. Lakarra. 2018. *Historia de la lengua vasca*.

Kondrak, G. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Lacombe, G. 1924. La langue basque. In *Les langues du monde*, pages 255–270, Paris.

Laka, I. 1996. *A brief grammar of Euskara, the Basque language*. UPV/EHU, Bilbao.

Lakarra, J. A. 1997. Euskararen historia eta filologia: arazo zahar, bide berri. *ASJU*, 31(2):447–535.

Liu, H. and J. Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Michelena, L. 1964. *Textos Arcaicos Vascos*.

Mitxelena, K. 1981. Lengua común y dialectos vascos. *International Journal of Basque Linguistics and Philology*, 15:289–313.

Pagola, R. M. 2006. Lazarragaren eskuizkribua: grafiak, hotsak eta hitzak. In *Lingüística Vasco-Románica. I Jornadas = Euskal-Erromantze Linguistika. I. Jardunaldiak*, pages 539–561, Donostia.

Pichel, J. R., P. Gamallo, and I. Alegria. 2018. Measuring language distance among historical varieties using perplexity. Application to European Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155.

Pichel, J. R., P. Gamallo, and I. Alegria. 2020. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, 26(4):433–454.

Sant'Anna, A. A. and L. Weller. 2020. The threat of communism during the cold war: A constraint to income inequality? *Comparative Politics*, 52(3):359–393.

Sarasola, I. 1983. Contribución al estudio y edición de textos antiguos vascos. *ASJU*, pages 69–212.

Satrustegi, J. M. 1987. *Euskal Testu Zaharrak*.

Scherrer, Y., T. Samardžić, and E. Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Singh, A. K. and H. Surana. 2007. Can corpus based measures be used for comparative study of languages? In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, pages 40–47. Association for Computational Linguistics.

Séguy, J. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane (RLiR)*, 37:1–24.

Ulibarri, K. 2013. Testuak kokatuz dialektologia historikoan: egiteetatik metodologiara. In *Koldo Mitxelena Katedraren III. Biltzarra / III Congreso de la Cátedra Luis Michelena / 3rd Conference of the Luis Michelena Chair*, pages 511–532, Vitoria-Gasteiz.

Zuazo, K. 2014. *Euskalkiak*. Elkar.

Zugarini, A., M. Tiezzi, and M. Maggini. 2020. Vulgaris: Analysis of a corpus for middle-age varieties of italian language. *arXiv preprint arXiv:2010.05993*.