Measuring Language Distance Using Perplexity

Group 2
Giovanni Bortoletto, Leena Manninen,
Emma McKenzie, Oona Raatikainen

December 2018

In this research project, we apply a perplexity-based method to measure language distance between five closely related Romance languages: Catalan, Galician, European Portuguese, Brazilian Portuguese, and Spanish. Perplexity measures have previously been successfully used to discriminate between similar languages or varieties and to measure language distance both synchronically and diachronically. In previous work, the best results have been achieved with language models based on 7-grams of characters. This report describes how this method can be used to discriminate closely related languages.

In this project, we used Open Subtitles 2018 corpus from OPUS – an open source parallel corpus. For analyzing the data, we used Taito-Shell on CSC – IT Centre for Science. The analysis was conducted with Natural Language Toolkit (NLTK) and a language model package available for NLTK.

The results on perplexity-based language distance support our hypothesis in how those five languages are related to each other. The lower the perplexity, the lower the distance between languages. Languages and their perplexities form a continuum from Catalan to Spanish, from Spanish to Galician, from Galician to European Portuguese and finally from European Portuguese to Brazilian Portuguese. In this case, languages most distant geographically are also most distant linguistically, i.e. Brazilian Portuguese and Catalan. This method could be extended to other Romance languages.

1 Introduction

In this research project, we measure linguistic distance between five languages of the Romance language family using perplexity. Perplexity measures have previously been used to discriminate between similar languages (Gamallo et al., 2016, 2017a), to identify

languages, and to measure language distance synchronically (Gamallo et al., 2017b) and diachronically (Gamallo et al., 2018). In our project, we are applying the perplexity-based method for measuring language distance between Catalan, Spanish, Galician, Portuguese and Brazilian Portuguese.

In picking these languages, we've tried to create a spectrum, where Catalan should be on one end, perhaps closer to Spanish, then Galician closer to Portuguese, and then on the other end are the two varieties of Portuguese, most similar to each other. It is also possible that European Portuguese may not be most similar to Brazilian Portuguese due to the fact that Brazil and Portugal are separated by the Atlantic Ocean, whereas Spain and Portugal are located right next to each other.

The structure of this report is as follows: In Section 2 we present background research relevant for our project. The point of view is, on one hand, from the perspective of historical linguistics, which forms the basis of our expected results, and on the other hand, from the perspective of previous methods used to measure linguistic distance. In Section 3 we present our data and how it was processed before our analysis. Analysis and its results are described in Section 4. Finally, in Section 5, we offer some conclusions and propose further research possibilities.

2 Background

2.1 Language History

In order to understand the divergences between Catalan, Spanish, Galician, European Portuguese, and Brazilian Portuguese, we must look at the history of the languages there to see what has influenced them and how that could relate to their similarities and differences.

Catalan evolved from Latin in the 700s, making it much older than Portuguese and Galician as independent languages (Marinzel, 2014). After a difficult history of decline, re-emergence, and then repression from the Spanish government, Catalan now has official status with Spanish in Catalonia (Drew, 2017). Along with the repression of Catalan, which caused the language to be influenced by Spanish (even today everyone is bilingual in Spanish and Catalan), Catalan has also been influenced by French and other Romance languages (Drew, 2017). According to Ethnologue, Catalan has 87% lexical similarity to Italian and 85% with Portuguese and Spanish (Lewis and Simons, 2018). Because of this equal lexical similarity, it will be interesting to see where Catalan falls in linguistic distance. Since Catalan has had so much contact with Spanish, we think Catalan will have a smaller distance with Spanish than with Portuguese.

For Spanish, the year 711 is an essential date; in that year, Moslem invasion broke the hegemony of the Germanic people settled in the peninsula, occupying the South and the Centre. Galicia and Franks/Catalans were two main points of resistance against the new owners of the lands. After the Great Battle of Rio Salado in 1340, the dialects of the conquerors (Portuguese, Castilian and Catalan) occupied the lands in sharply defined spheres of influences. However, many borrowings from Mozarabs language still survive in Spanish, such as names, liturgical terms, and technical words (e.g. names of weapons,

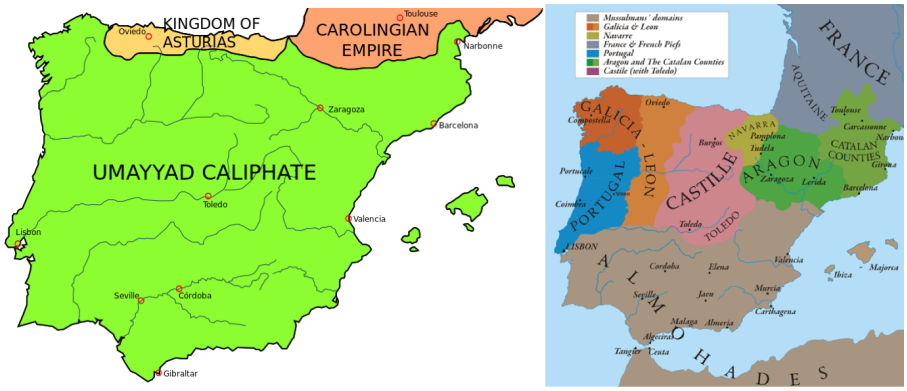

Figure 1: Iberian Peninsula territory comparison between VIII and XIII centuries, before and after the Christian Reconquista. (wikipedia.org/wiki/History_of_Spain)

war tactics, places, measurements, agriculture, etc.). We expect to find more Arabisms in Spanish than in other languages in the country (Entwistle, 1962).

Modern Spanish flourished in the north: their language was a heterogeneous compromise between conservative Galician, innovative Castilian, the Latin origins and some borrowings from Arabic Mozarabs. This is why we expect to find many similarities between Spanish and Galician (Entwistle, 1962).

Though Portuguese and Galician were once the same descendant of Latin, we don't expect them to have a small perplexity difference. Until the early 14th century, Galician and Portuguese were still the same language, but once they began to diverge, they diverged quickly due to a want to make Portuguese a distinct language (Azevedo, 2005). Any features in Portuguese that were too similar to Galician features were thrown out of use (Azevedo, 2005). While Galician is closely related to Portuguese, Spanish has also had a lot of impact on modern Galician (Bostrom 2006). This is not only due to the language contact, but also due to the fact that Galician had a lower social status, so more Spanish loanwords slipped into Galician language because Spanish was the high-status language (Bostrom, 2006). Because of this Spanish influence, we expect Galician to be more in the middle between Spanish and Portuguese, while slightly favoring Portuguese due to them once being the same language.

When Portugal first colonized Brazil, there were many different groups of indigenous people living there (Holm, 1989). Lingua Geral, the Lingua Franca of Brazil was based off of one of the indigenous languages and was spoken by most of the population for the first 250 years of colonization (Holm, 1989). Brazil was also a main layover for slaves, with nearly forty percent of the Atlantic slave trading using Brazil as a stopping point (Guy, 1981). From all this contact with indigenous and African languages, Brazilian Portuguese adopted a lot of lexical terms (Guy 1981). For example, while Catalan, Galician, and Spanish all have words for pineapple that are derived from Latin *pineum*, the Brazilian Portuguese word for pineapple is *abacaxi*, which originates from the extinct classical Tupi, an indigenous language in Brazil. In addition, European Portuguese also opts for a word that originates from a language native to Southern America: *ananás* has

its origin ultimately in the Guaraní word *naná*.

Table 1: Some Romance cognate sets.
Words mentioned in parentheses are infrequent or obsolete. **ninus* is Vulgar Latin, *puer* Classical Latin.

English	**boy**	**bread**	**dog**	**mother**	**pineapple**
Catalan	noi	pa	gos (ca)	mama	pinya
Galician	neno	pan	can	mamá	piña
European Portuguese	menino	pão	cão	mãe, mamãe	ananás
Brazilian Portuguese	menino	pão	cão	mãe, mamãe	abacaxi
Spanish	niño	pan	perro (can)	mamá	piña (ananás)
Latin	puer, *ninus	panem	canis	mater	ponum, pineum

It's also very possible that the other differences between Brazilian and European Portuguese also came from this language contact. Other than lexicon, the main differences between Brazilian and European Portuguese are: use of pronouns, use of verbal clitics, and patterns of number agreement (Guy, 1981). Lower class Brazilian Portuguese has more use the unique Brazilian features that come from the African and indigenous languages (Guy, 1981). We expect our data to be standard Brazilian Portuguese, which is upper class Brazilian Portuguese, which is more similar to European Portuguese (Guy, 1981). Because of this, we expect Brazilian Portuguese and European Portuguese to have a small perplexity measure between them.

Table 2 summarizes the modern situation in terms of the number of speakers around the globe.

Table 2: Speakers (million), (Ethnologue, 2018)

Language	Speakers (m)	
	L1	Total
Catalan	4	9
Galician	2	2
Portuguese	10	10
Brazilian Portuguese	204	237
Spanish	46	513

2.2 Measuring Linguistic Distance

Linguistic distance has been analyzed and measured traditionally by various methods, including phylogenetics, also known as the historical-comparative method, and dialectology. The historical-comparative method aims to reconstruct ancestor languages and form language classifications, such as in Figure 2. The core of this method is the comparison of cognate sets, such as in Table 1 above, identify changes and on the basis of these changes reconstruct the parent language. Dialectology, on the other hand, challenges the basic assumption of sound change being regular and exceptionless, and offers a 'wave theory' as opposed to the 'family-tree model'. Unlike the historical-comparative

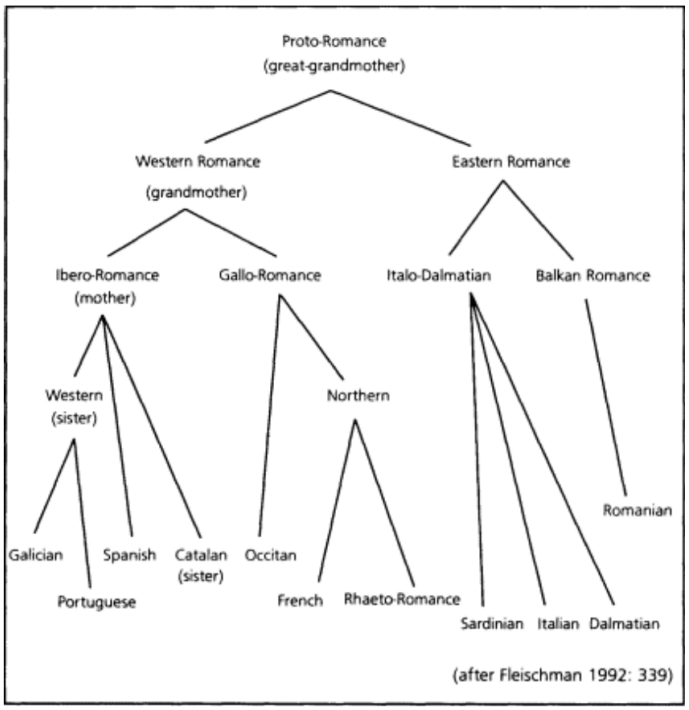

Figure 2: Proto-Romance family tree (Campbell, 2013, pg. 103)

method, dialectology also takes into account language external factors, such as prestige. (Campbell, 2013)

Lexicostatistics is a quantitative method for measuring linguistic distance in the vain of the historical-comparative method. In lexicostatistics, the linguistic distance is calculated on the basis of the cognates that languages share. One basic assumption behind this method is that the basic vocabulary resists replacement by borrowings. Consequently, linguistic distance can be calculated on the basis of cross-linguistic word lists, such as the Swadesh 200-list, which contains what is presumed to be universal "basic vocabulary". The more cognates a pair of languages share, based on this word list, the closer they are related. (Campbell, 2013)

Glottochronology is further application of lexicostatistics. This method has been described as the "carbon-dating" of linguistics. Glottochronology applies the same method as lexicostatistics, but the goal is to assign the date of the split-up of a given language based on the lexical items that they share. Both lexicostatistics and glottochronology have been criticized, for instance, for assuming that a basic vocabulary would be universal and not subject to cultural differences. (Campbell, 2013)

2.3 Previous perplexity-based work

According to Gamallo, Pichel & Alegria, concept of language distance is closely related to the process of language identification (Gamallo et al., 2017b). In general, two types

of methods have been used in language detection: those based on n-grams of characters and those based on word unigrams or dictionaries. Language detection systems work well when tested on long and well-written texts. Closely related languages, language varieties or dialects are more difficult to identify and separate than languages belonging to different linguistic families. (Gamallo et al., 2016) Language identification is difficult also when tested on noisy short texts such as those written in social media. (Gamallo et al., 2018)

In their study, Gamallo, Pichel & Alegria (2016) compared two different methods for language detection. First method based on classification with ranked dictionaries, and second naïve Bayes system based on word unigrams. They observed how these methods behave when used in the difficult task of discriminating between similar languages or varieties. They made experiments on similar languages with both character n-grams and word unigrams. They achieved best results using word unigrams. (Gamallo et al., 2016)

Gamallo, Pichel & Alegria (2017a) used a perplexity-based method for similar language discrimination. They used perplexity to compare different language models. Perplexity of test data is the most widely used evaluation metric for language models. In their study, they used different n-gram models of words and characters, and they observed that short n-grams for words and long n-grams for characters performed best. Therefore, they used in their runs 1-grams of words and 7-grams of characters. (Gamallo et al., 2017a)

Gamallo, Pichel & Alegria (2017b) defined two quantitative distances to measure linguistic distance, in other words how different one language or variety is from another: perplexity and ranking. The more accurate one based on the perplexity on n-grams models extracted from text corpora. (Gamallo et al., 2017b)

Measuring distance between two languages is not easy. Languages can differ in many linguistic aspects such as phonetics, morphology, syntax, semantics and pragmatics. Reducing these aspects to a single distance score between languages is difficult. "There is not any standard methodology to define a metric for language distance." The shorter the distance between two varieties or languages is, the more difficult the identification. "The best language identification systems are based on n-gram models of characters extracted from textual corpora." Long n-grams encode different linguistic aspects. (Gamallo et al., 2018)

Perplexity (PP) is a robust metric to calculate distance between languages. Perplexity measures how well a language model fit the test data. A perplexity-based distance (PLD) between languages or varieties is determined by comparing n-gram based language model (LM) of language L1 and test text (CH) of language L2. The comparison must be made in the two directions.

$$\text{PLD (L1, L2)} = (\text{PP(CHL1, LML2)} + \text{PP(CHL2, LML1)})$$

The perplexity-based distance between two languages is the mean derived from two perplexity values. The lower the perplexity the lower the distance between languages. (Gamallo et al., 2018)

Pichel, Gamallo and Alegria Gamallo et al. (2018) have used perplexity-based distance (PLD) to measure distance between historical varieties of European Portuguese.

They used text corpora from different historical periods. Their implementation of PLD contained 7-gram models and a smoothing technique. They expect that this metric can be applied to other languages. (Gamallo et al., 2018)

3 Data

As our corpus, we are using the OpenSubtitles 2018 corpus – an open source parallel corpus. OPUS (Tiedemann, 2012) is a collection of translated texts from the web. Data is aligned and available for free[1]. OpenSubtitles corpus[2] (Lison and Tiedemann, 2016) a collection of translated movie subtitles from OpenSubtitles.org[3]. We chose to use the OpenSubtitles 2018 corpus since it is available online for free and provides enough data for our project. The available corpus sizes for our set of languages are given in Table 3.

Table 3: Corpus sizes in tokens (m)

Parallel Corpus	Tokens (m)
Catalan - Spanish	7,6
Catalan - Portuguese	6,3
Catalan - Galician	0,2
Catalan - Brazilian Portuguese	4,7
Galician - Catalan	0,2
Galician - Spanish	3,5
Galician - Portuguese	1,4
Galician - Brazilian Portuguese	2,8
Portuguese - Catalan	6,3
Portuguese - Spanish	469,6
Portuguese - Galician	1,4
Portuguese - Brazilian Portuguese	469,5
Brazilian Portuguese - Catalan	7,4
Brazilian Portuguese - Spanish	823,9
Brazilian Portuguese - Galician	2,8
Brazilian Portuguese - Portuguese	469,9
Spanish - Catalan	7,6
Spanish - Portuguese	469,6
Spanish - Galician	3,5
Spanish - Brazilian Portuguese	823,9

[1] http://opus.nlpl.eu/
[2] http://opus.nlpl.eu/OpenSubtitles2018.php
[3] http://www.opensubtitles.org/

There is considerable variation in the sizes of the corpora. The parallel corpus for the language pair Spanish – Brazilian Portuguese is by far the biggest with over 800 million tokens. In addition, corpora for the language pairs Spanish – Portuguese and Portuguese – Brazilian Portuguese are among the largest with over 400 million tokens. Corpora including Catalan and Galician are the smallest, Galician – Catalan corpus being only 200 000 tokens large. However, taking into account the number of speakers of these languages, presented in Table 2, the small size of both Catalan and Galician corpora is not surprising.

To access this data, we used Taito-shell on CSC – IT Centre for Science. OPUS files are already stored there. First, we loaded the module:

```
module use -a /proj/nlpl/software/modulefiles/ module load nlpl-opus
```

Then, we specified the languages we are interested in. For our project, we needed datasets of comparable sizes for each language. Therefore, the language with the smallest corpus, i.e. Galician, was used as a pivot-language when extracting our datasets, and all other languages aligned with it.

```
opus2multi /proj/OPUS/corpus/OpenSubtitles/xml gl es pt pt_br ca
```

After this we have XML files with sentence ID numbers, with Galician as the pivot language. Files are gl-es.xml, gl-pt.xml, gl-pt_br.xml and gl-ca.xml. Then we read the files and write sentences to monolingual files. Next command reads gl-es.xml file and writes the Galician sentences into the gl.txt file and the Spanish sentences into the es.txt file.

```
opus2moses -d /proj/OPUS/corpus/OpenSubtitles/xml -e gl.txt -f es.txt
gl-es.xml
```

We did this to every language, getting sentences in monolingual text files (es.txt, pt.txt, pt_br.txt, ca.txt and gl.txt). The datasets were pre-processed as follows. First all characters were converted into lower case with the function, and all punctuation, including punctuation marks <¿> and <¡>, specific to Spanish, were removed. Second, all characters with diacritics were converted into their non-accented counterparts, so that characters such as <ã>, <ç>, <ñ>, <ó>, <ï> were converted to <a>, <c>, <n>, <o>, <i>, respectfully. Finally, all multiple white spaces, which were present in some of the corpora, were converted into singe white spaces.

The problem with this approach is that it doesn't take into account phonetic features, but is based on standardized orthography which differs more or less from language to language, even with closely related languages. For example, the same palatal nasal [ɲ] is written <ñ> in Spanish, but <ny> in Catalan, and the palatal fricative [ʃ] is marked with <x> in Catalan, Galician, European Portuguese and Brazilian Portuguese, but usually with <ch> in Spanish. Thus, words piña and pinya, presented in Table 1 above, are phonetically more or less identical, [piɲa]. To accurately model the phonological similarity of our language set, a unified, pseudo-phonetic transcription would be needed.

Also, the differences in intonation patterns as well as other spoken language phenomena are not attained with our model. However, within limits of the time frame for our project, taking these considerations into account is not feasible.

After pre-processing the data were divided into training and test sets, the training set comprising ca. 80% of the corpus, and the test set ca. 20%. The division was based on characters rather than lines, since lines vary considerably in their length. The training and test set sizes are given in Table 4. It is worth noticing that the Brazilian Portuguese corpus is approximately one seventh smaller than all the other corpora. The difference in size must have something to do with the process of data extraction rather than pre-processing, since the size of the raw data for Brazilian Portuguese is smaller (394,71 kB) compared to the raw data of all the other languages (ranging from 448,29 kB to 463,5 kB).

Table 4: Training and Test Set sizes in characters

Language		**Set sizes**		
		Train	Test	Total
Catalan	*Lines*	8200	2208	10408
	Characters	336902	83344	420246
Galician	*Lines*	8600	2305	10905
	Characters	336450	83545	419995
European Portuguese	*Lines*	8600	2362	10962
	Characters	345254	85970	431224
Brazilian Portuguese	*Lines*	7690	2321	10011
	Characters	291768	72287	364055
Spanish	*Lines*	8090	2273	10363
	Characters	335335	83632	418967

4 Experiments

4.1 Analysis

In our analysis, we used character based 7-grams, as was done in Gamallo, Pichel & Alegria's study of historical varieties of Portuguese (2018). We chose long character-based ngrams first of all since they are able take into account syntax, as 7-grams may represent the end of a word and the beginning of the next in one sequence, for example a preposition + article sequence such as #de#la#, or noun + preposition sequence such as ion#de#, which are both frequent sequences in many Romance languages. The second reason for choosing character based 7-grams was to make our results comparable with Gamallo et al.'s results for purposes of evaluation.

The analysis was conducted with Natural Language Toolkit (Loper and Bird, 2002) version 3.4, henceforth NLTK, and a language model package available for NLTK, called *nltk.lm*[4]. First step was to "pad" the data with symbols that indicate the start and

[4]http://www.nltk.org/api/nltk.lm.htmlmodule-nltk.lmk

end of a sentence. Then the data was split into 7-grams. The *nltk.lm* package offers one single function for this, called `padded_everygram_pipeline()`, which results in the pre-processed input for the language model. Same function was used for all the test sets as well. The training of the language model was done with the function `lm.fit()`, where lm is a model for a specific language with a specified order of 7.

On the basis of our trained language models, we can generate text with function `lm.generate()`. This can also be used to evaluate if our language model works at all. Our Spanish language model generates something that looks strikingly like Spanish, if not a bit foul-mouthed: *"... rosamente por tu vientre sus putas maquille o algunos ladrones y sus superstriciosa sobre sus putas m..."*. This roughly translates as "... [...] for your belly your whores [...] or some thieves and their [...] on their whores". Generated words *rosamente*, *maquille* and *superstriciosa* are not Spanish, but quite close: *maquilla* is similar to the verb *maquillar*, 'to put on make-up', and *superstriciosa*, without the second <r>, would be a feminine singular form of *supersticioso*, 'superstitious'. *Rosamente* follows the rules of adverb formation, i.e. combining an adjective with the ending –mente – freely translating *rosamente* would mean *rosely* or *pinkly*. The Catalan language model generates strings that also resemble Catalan: *"hi ha prou homes e spasa de portare el poder i un home mort apropos del rei es morts els nostres i protegireu en ella forca distribuir..."*, which would translate as "...There are enough men and [...] of I will take the power and a dead man approached of the king is dead ours and will protect in her it forces to distribute..." Here the non-word *spasa* resembles *espasa*, 'sword'. Of course, neither of the generated sequences makes much sense or has a coherent logic, but on phrase-level the language models are performing quite well.

The final step in our analysis was to calculate perplexities between every language pair, resulting in 25 perplexity measures. The *nltk.lm* package provides a function, `lm.perplexity()` for this. Perplexities were first measured line by line, and then the overall mean was calculated.

At first, we used a Maximum Likelihood Estimator (MLE) as our language model. This model has no smoothing method, which means that when the trained language model encounters a character or a character sequence it hasn't encountered in the training data, the perplexity of this line pair becomes infinite. Thus, in our first perplexity test runs resulted in mostly infinite perplexity measures, and calculating the mean would in turn yield an infinite perplexity. Since infinite values cannot be compared (there is are no values that are more or less infinite than others), our language model needed to be changed. We experimented with other language models which apply various smoothing methods provided by the *nltk.lm* package. Out of WittenBellInterpolated, KneserNeyInterpolated, Lidstone and Laplace models, Laplace was the only one that actually gave us any output, so our final experiments were conducted with it.

4.2 Results

The results of our analysis are presented in Table 5 below. Here the rows indicate language models, and columns the test set language. The perplexity measures range from 6,961 to 14,344, being higher than we expected. For instance, in Gamallo et al.'s previous

work on historical varieties of Portuguese (Gamallo et al., 2018), perplexity measures range from 2,849 to 7,692. In addition, these is more dispersion in the perplexity measures than we expected.

Table 5: Perplexity measures

Language Model	Test set				
	Catalan	Galician	European Portuguese	Brazilian Portuguese	Spanish
Catalan	6,961	11,941	12,585	12,423	11,339
Galician	13,339	7,082	10,147	10,026	10,282
European Portuguese	13,639	9,757	6,970	7,068	11,572
Brazilian Portuguese	14,344	10,395	7,706	7,113	12,177
Spanish	12,736	10,475	12,205	12,077	6,808

However, the overall pattern is exactly what was expected. A heatmap illustrates this better than a table. A simple heatmap including a dendrogram was made with the statistical computing environment R (R Development Core Team, 2008), with the basic `heatmap()` function and the *RColorBewer* package (Neuwirth, 2014), presented in Figure 3 below. In this heatmap, the darker colours indicate higher perplexities and lighter colours lower perplexities.

As expected, the lowest perplexity measures are obtained when a language model is tested with the same language it is trained with. However, perplexities between Brazilian and European Portuguese are also quite low. Galician seems to be an intermediate language between the two Portugueses on one hand, and Spanish and Catalan on the other hand. Spanish does better with Galician and Catalan than with Brazilian or European Portuguese.

The dendrogram resembles the proto-Romance family tree presented in Figure 2 above. The first major split is into Spanish/Catalan and Galician/Portuguese. The Galician–Portuguese node further splits into Galician and a node comprising the two Portuguese variants that split on the lowest level.

To illustrate the spectrum-like nature of our results, we also calculated the mean perplexities of each language pair. Here we calculated the mean of a language model^{A} with test set^{B}, and the language model^{B} with the test set^{A}, for example, the Catalan language model tested with Spanish, and the Spanish language model tested with Catalan. The results are presented in Table 6 below.

These values are also in line with our hypothesis. The pair European Portuguese/Brazilian Portuguese has the lowest value, Brazilian Portuguese/Catalan the highest. Here we can also see that European Portuguese and Galician are much closer to each other in comparison with Brazilian Portuguese and Galician. In fact, Brazilian Portuguese/Galician and Brazilian Portuguese/Spanish have the same value.

To present this a hierarchical cluster analysis was conducted in R with the function `hclust()`, presented in Figure 4 below. For this method, the distances between the mean perplexity values are calculated with the function `dist()`, resulting in a distance matrix, which is the input for the hierarchical cluster analysis.

This analysis confirms our original hypothesis. Catalan and Brazilian Portuguese are

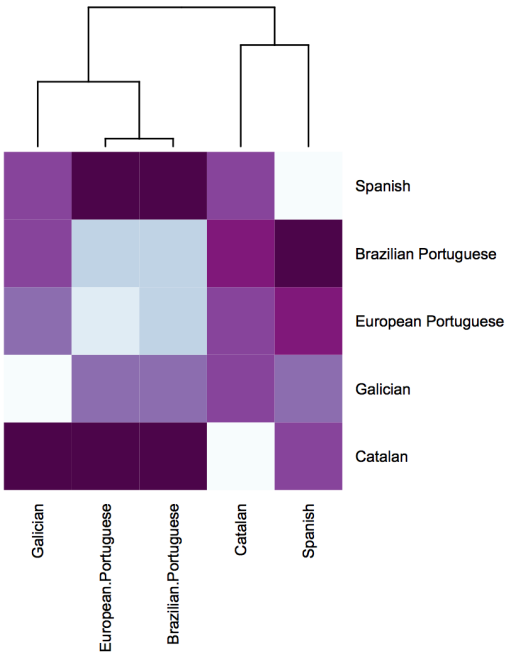

Figure 3: Heatmap of perplexity measures. Darker colors indicate higher perplexity.

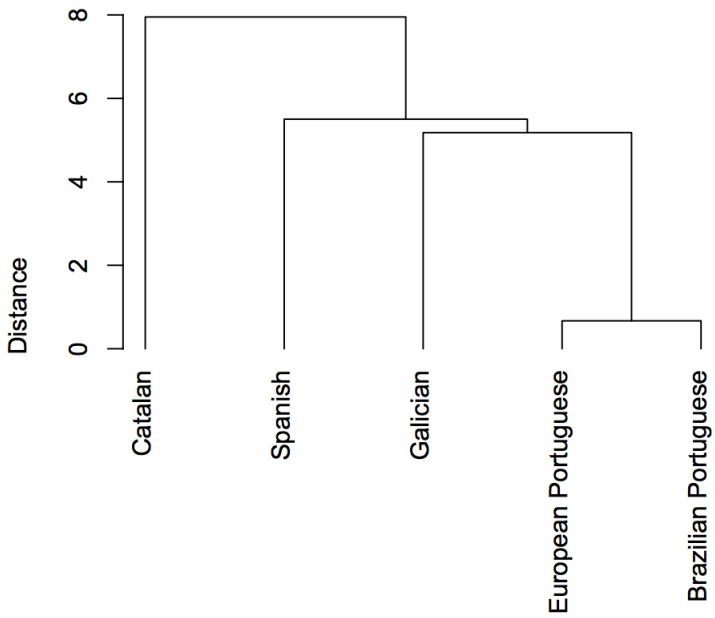

Figure 4: Hierarchical clustering based on mean perplexity values

Table 6: Mean perplexity values ranked from highest to lowest

Language Pair	Mean Perplexity
Catalan / Brazilian Portuguese	13,38
Catalan / European Portuguese	13,11
Catalan / Galician	12,64
Galician / Brazilian Portuguese	12,13
Spanish / Brazilian Portugurse	12,13
Catalan / Spanish	12,04
Spanish / European Portuguese	12,04
Galician / Spanish	10,38
Galician / European Portuguese	9,95
European Portuguese / Brazilian Portuguese	7,39

at the opposite ends of the spectrum, Spanish and Galician in the middle. However, the distance between the split of Spanish and Galician is much smaller than the distance between the split of Catalan and the other languages suggests that Catalan is the most divergent in this group of languages. In general terms, the distances between the splits converge with the history of these languages presented above in Section 2.1.

The major split between Catalan and all other languages in also fits the dispute of Catalan's classification within the Romance language family. Some linguists maintain that Catalan is an Ibero-Romance language, like the rest of the languages in our language set. This was the approach adopted in the Romance language tree presented above in Figure 1. On the contrary, some linguists argue that Catalan should be classified in the Gallo-Romance group, or in the smaller Occitano-Romance subgroup with Occitan, a language spoken in Southern France. (Bossong, 2016)

It is also interesting that European Portuguese and Brazilian Portuguese are so close to each other. It is possible that the lexical differences between the languages are not that significant, or it might be that they were not present in our data. One further possible explanation for the similarity between European and Brazilian Portuguese is the power of standardized orthography which doesn't evolve as rapidly as the spoken language.

5 Conclusion

Other than matching the geographical map, our results also match the history of the languages. Catalan was the furthest away from all the languages, while slightly closer to Spanish, which reflects its long history of being an independent language (Marinzel, 2014), as well as the influence of Spanish since it was the dominant language (Drew, 2017). Spanish has had a lot of contact with the other languages in the Iberian Peninsula, but it also has a lot more Arabic influence (Entwistle, 1962), which put it closer to the other languages than Catalan was, but still with distance, with quite a lot of distance from the Portuguese varieties. Galician was in the middle, between the Portugueses and

Spanish, which we guessed based on the fact that Galician and Portuguese were the same language until the 14th century (Azevedo, 2005) and the contact with Spanish, which was viewed as prestigious (Bostrom, 2006). Finally, European and Brazilian Portuguese were very similar, which we expected since standard written Brazilian Portuguese is very similar to European Portuguese (Guy, 1981) and we were using written text.

There are multiple possibilities for further research based on this experiment. First of all, it would be interesting to recreate this experiment with a pseudo-phonetic transcription to avoid the problems of our preprocessing choices that were discussed above in Section 3. This would, of course, ask for a close examination of the orthography and the phonology of each language and the creation of a transcription system should be careful. It would, for instance, be necessary to devise a way to distinguish between nasal and oral vowels. In addition, to be consistent with the transcription, multiple values of a given grapheme would have to be taken into account – for example, in Spanish alone the letter <c> may represent a voiceless stop [k], a sibilant [s] or a voiceless dental fricative [θ].

It would also be interesting to recreate this study with more languages, perhaps including the whole Romance spectrum from Latin American Spanish varieties to multiple Italian varieties and finally to Romanian at the easternmost end of the spectrum. This could also shed more light on the classification of Catalan. We could also try to confirm Ethnologue's claim of Catalan and Italian having 87% lexical similarities.

Our original idea was actually to have more languages and/or variants, for example Latin American variants of Spanish. When searching for data sets, we noticed that OPUS provides data for Latin American Spanish variants, mostly in the GNOME[5] and Ubuntu[6] corpora, which contain localization files. However, the parallel corpora are quite small, some as small as 300 000 tokens. At the beginning of this project, we were unsure about the amount of data we needed, to opting for the biggest data available seemed the best choice. Then again, our parallel corpus for Catalan – Galician was mere 200 000 tokens, and Galician was used as the pivot language for data extraction. Our results would suggest that ruling out smaller data sets is not by default necessary.

References

Milton M. Azevedo. *Portuguese: A Linguistic Introduction*. Cambridge University Press, Cambridge, 2005.

Georg Bossong. Classifications. In Adam Ledgeway and Martin Maiden, editors, *The Oxford Guide to the Romance Languages*, pages 63–72. Oxford University Press, Oxford, 2016.

Jay G. Bostrom. *Which Way for Catalan and Galician? Master's Thesis.* The University of Montana, Missoula, Montana, 2006. URL `https://scholarworks.umt.edu/cgi/viewcontent.cgi?article=2201context=etd`.

[5] http://opus.nlpl.eu/GNOME.php
[6] http://opus.nlpl.eu/Ubuntu.php

Lyle Campbell. *Historical Linguistics. An Introduction.* Cambridge University Press, Cambridge, third edition, 2013.

Lisel Drew. *'I'm from Barcelone': Boundaries and Transformations Between Catalan and Spanish Identities. Master's Thesis.* Uppsala Universitet, Uppsala, 2017. URL `https://www.diva-portal.org/smash/get/diva2:1112709/FULLTEXT01.pdf`.

William J. Entwistle. *The Spanish Language, Together with Portuguese, Catalan and Basque.* Faber & Faber, London, 1962.

Pablo Gamallo, José Ramon Pichel, Alegria Iñaki, and Agirrezabal Manex. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, 2016.

Pablo Gamallo, José Ramon Pichel, and Alegria Iñaki. A perplexity-based method for similar languages discrimination. In *Proceedings of Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, 2017a.

Pablo Gamallo, José Ramon Pichel, and Alegria Iñaki. From language identification to language distance. *Physica A*, 483:162–172, 2017b.

Pablo Gamallo, Pichel José Ramon, and Alegria Iñaki. Measuring language distance among historical varieties using perplexity. application to european portuguese. In *Proceedings of Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 2018.

Gregory R. Guy. *Linguistic Variation in Brazilian Portuguese: Aspects of the Phonology, Syntax, and Language History. Doctoral Dissertation.* University of Pennsylvania, Philadephia, 1981.

John A. Holm. *Pidgins and Creoles: Volume II: Reference Survey.* Cambridge University Press, Cambridge, 1989.

M. Paul Lewis and Gary F. Simons, editors. *Ethnologue: Languages of the World.* SIL International, Dallas, TX, USA, twenty-first edition, 2018. URL `https://www.ethnologue.com`.

Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).)*, 2016.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

Anastazia Marinzel. *Catalonia: The Quest for Independence from Spain.* John Carroll University, Missoula MT, 2014. URL `http://collected.jcu.edu/honorspapers/39`.

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. URL `https://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-2.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.