Physica A 484 (2017) 152–162

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# From language identification to language distance

Pablo Gamallo [a,*], José Ramom Pichel [b], Iñaki Alegria [c]

[a] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), University of Santiago de Compostela, Galiza, Spain
[b] Imaxin|Software, Galiza, Spain
[c] IXA Nlp Group, UPV/EHU, Basque Country, Spain

## HIGHLIGHTS

- A quantitative metric to measure language distance is proposed.
- Language distance is measured using the perplexity of corpus-based *n*-gram models.
- A current map of distances among almost all languages of Europe is presented.

## ARTICLE INFO

*Article history:*
Received 12 January 2017
Received in revised form 22 March 2017
Available online 11 May 2017

*Keywords:*
Language distance
*N*-gram models
Perplexity
Corpus-based linguistics
Natural language processing
Language identification

## ABSTRACT

In this paper, we define two quantitative distances to measure how far apart two languages are. The distance measure that we have identified as more accurate is based on the *perplexity* of *n*-gram models extracted from text corpora. An experiment to compare forty-four European languages has been performed. For this purpose, we computed the distances for all the possible language pairs and built a network whose nodes are languages and edges are distances. The network we have built on the basis of linguistic distances represents the current map of similarities and divergences among the main languages of Europe.



## 1. Introduction

In this article, we deal with the concept of *language distance*, which refers to how different one language or variety is from another. Even though there is no well-established measure to quantify the distance between two languages [1], some specific linguistic work relies heavily on the use of this concept, namely in phylogenetic studies within historical linguistics [2,3], in dialectology [4], or in studies about learning additional languages within the field of second language acquisition [5]. The prevailing view, however, is that linguistic distance cannot be measured since two languages may differ in many linguistic aspects, e.g. phonetics, written form, morphology, syntax, and so on. Quantifying all these aspects by reducing them to a single distance score is a difficult task which is far from being fulfilled or at least appropriately addressed, perhaps as it has not yet been a priority in the field of Natural language Processing (NLP).

The concept of language distance seems to be related to the process of language identification. In fact, language distance and language identification are two sides of the same coin. The more difficult the identification of differences between two languages is, the shorter the distance between them. Language identification was one of the first natural language processing problems for which a statistical approach was used and it is now considered as an (almost) solved problem

* Corresponding author.
*E-mail addresses:* pablo.gamallo@usc.es (P. Gamallo), jramompichel@imaxin.com (J.R. Pichel), i.alegria@ehu.eus (I. Alegria).

http://dx.doi.org/10.1016/j.physa.2017.05.011
0378-4371/

except for complex tasks such as similar variety discrimination or short text classification. The best language identification systems are based on *n*-gram models of characters extracted from text corpora.

The main objective of our work is to define a linguistic distance between two languages by considering character-based *n*-gram models, in a similar way to traditional language identification strategies. Character *n*-grams not only encode lexical and morphological information, but also phonological features since written systems are related to the way languages were pronounced in the past. In addition, long *n*-grams ($\geq$5-grams) also encode syntactic and syntagmatic relations as they may represent the end of a word and the beginning of the next one in a sequence. For instance, the 7-gram *ion#de#* (where '#' represents a blank space) is a frequent sequence of letters shared by several Romance languages (e.g. French, Spanish, or Galician).[1] This 7-gram might be considered as an instance of the generic pattern "*noun-prep-noun*" since *ion* is a noun suffix and *de* a very frequent preposition (translated as *of* or *from* in English) introducing prepositional phrases. So, models built from corpora and based on long character *n*-grams are complex linguist artifacts provided with linguistic features at different levels, including phonological, morphological, lexical, and even (very basic) syntactic information. We must point out that our study is aimed at comparing not a continuum of dialectal varieties, but well-defined written standards. These are *standardized varieties* including not only standards that are distinctly separate from any other language (*Abstand* languages or languages by distance), but also cultural and political constructs known as *Ausbau* (elaboration) languages. The latter are called *elaboration* languages because their distance to each other has been elaborated historically even though they are mutually intelligible [6].

In order to compute language distance, two specific metrics will be proposed. Firstly, we measure the *perplexity* of a *n*-gram model on a test text. Perplexity is defined as the inverse probability of the test text given the model. Most of the best systems for language identification use probability-based metrics with *n*-gram models. Secondly, we also use a *ranked-based* method that ranks *n*-grams according to frequency. *N*-grams with highest frequencies are retained and the rest are discarded. This gives us pruned character *n*-gram models which are used for defining the distance between languages. These two metrics were chosen because they represent two well-known families of language identification strategies: those that classify languages according to *n*-gram probabilities, and those relying on ranked lists of *n*-grams.

The two metrics will be tested in different experimental setups. We start by testing their performance in a language identification task, and then, we use them to measure the distance between European languages. The latter experiment will allow us to draw a diagram showing the linguistic distance among most European languages. The diagram will be derived from a 2D-matrix of languages and their relationship to each other.

The remainder of the article is organized as follows. Section 2 introduces the works using the notion of language distance in historical linguistics, as well as the main methods used in language identification. Following this, Section 3 defines two distance measures based on *n*-grams of characters. Two experiments are reported in Section 4: the first one uses our distance measures for the difficult task of identifying similar languages and varieties, and the second one applies them for building a network of the main languages of Europe. Finally, conclusions are drawn in Section 5.

## 2. Related work

Linguistic distance has been measured and defined from different perspectives using different methods. Many of them compare lists of words to find phylogenetic links, while there are few corpus-based approaches from a synchronic point of view.

### 2.1. Phylogenetics and lexicostatistics

The objective of linguistic phylogenetics, a sub-field of historical and comparative linguistics, is to build a rooted tree describing the evolutionary history of a set of related languages or varieties [3]. In order to automatically build such a phylogenetic tree, many researchers make use of what they call *lexicostatistics*, which is an approach of comparative linguistics that involves quantitative comparison of lexical cognates [7–9,2,3]. More precisely, these computational studies are based on cross-lingual word lists (e.g. Swadesh list [10] or ASJP database [11]) to measure distances from the percentage of shared cognates, which are words with a common historical origin. Given a standardized word list, the distance between a pair of languages is defined by considering the cognates they share. More precisely, as described by Wichmann [12], the basic lexicostatistical technique defined by Swadesh consists of the following steps: (1) a cross-lingual word list is created, (2) cognates are identified, (3) the percentage of shared cognates is computed for each pair of languages to produce a pairwise inter-language distance matrix, and (5) the lexical distances are transformed into separation times: the more distant two languages are, the more time is needed to find a common ancestor. This last step is one of the main objectives in *glottochronology*.

Other related work proposed an automated method which uses Levenshtein distances among words in a cross-lingual list [2]. Unlike lexicostatistical strategy, this method does not aim to distinguish cognates from non-cognates. The global

[1] The stress accent (e.g. *ión*) has been removed to simplify language encoding.

distance between two languages is computed by considering a normalized Levenshtein distance between words and then finding the average of all such distances contained in the list.

A slightly different strategy is based on more standard supervised machine learning approaches. The input to a phylogenetic analysis is generally a data matrix, where the rows represent the given languages, and the columns represent different linguistic features (also called *characters*) by which the languages are described [13]. Features need not be lexical; they can also be syntactic and phonological. Some of these approaches use Bayesian inference to classify new data on the basis of the language models coded in the data matrix [14].

Computational methods taken from computational phylogenetics have been applied not only on lists of lexical units but also on phonetic data [7]. And they have been used to explore the origins of Indo-European languages [15,16], Austronesian language groups [17,16], Bantu languages [18], as well as the subfamily of Dravidian languages [19].

In sum, computational phylogenetics use cross-lingual lists to compute string or/and phonological distances among words, which are in turn used to measure distances among languages. These distances are then submitted to tree-building or clustering algorithms for the purpose of generating phylogenetic trees or clusters showing historical relationships among languages [20]. An excellent survey explaining the different types of phylogenetic strategies is reported in Wichmann [12].

### 2.2. Distributional-based approaches

To compare different languages, very recent approaches construct complex language models not from word lists, but from large cross-lingual and parallel corpora [21–23]. In these works, models are mainly built with distributional information on words, i.e. they are based on co-occurrences of words, and therefore languages are compared by computing cross-lingual similarity on the basis of word co-occurrences. The works by Liu and Cong [21,22] were performed on a relatively small number of languages. More precisely, Liu and Cong [21] compared fourteen languages and Gao et al. [22] studied merely six languages. By contrast, Asgari and Mofrad [23] performed language comparison on fifty natural languages from different linguistic families, including Indo-European (Germanic, Romance, Slavic, Indo-Iranian), Austronesian, Sino-Tibetan, Altaic, Uralic, and Afro-Asiatic. The authors built the language models for each language from a collection of sentence-aligned parallel corpora. The corpora used is the Bible Translations Project described in Christodoulopoulos et al. [24]. The results of this large-scale language comparison are, however, not very promising, since the similarity measure gives rise to several counter-intuitive findings. For instance, Norwegian and Hebrew, belonging to two different language families (Indo-European and Semitic), are wrongly grouped together. The system also separates in different clusters the two main languages of the Finno-Permian family: Estonian is clustered with Arabic and Korean while Finish is grouped with Icelandic, an Indo-European language.

Another limitation of the distributional-based approaches is that they require parallel corpora to build the models to be compared, and this kind of data is not easily available for many pairs of languages.

### 2.3. Language identification

Two specific tasks of language identification have attracted a lot of research attention in recent years, namely discriminating among closely related languages [25] and language detection on noisy short texts such as tweets [26,27].

The Discriminating between Similar Languages (DSL) workshop [28,29,25] is a shared task where participants are asked to train systems to discriminate between similar languages, language varieties, and dialects. In the three editions organized so far, most of the best systems were based on models built with high-order character $n$-grams ($\geq$5) using traditional supervised learning methods such as SVM, logistic regression, or Bayesian classifiers. By contrast, deep learning approaches based on neural algorithms did not perform very well.

TweetLID [30,27] is another shared task aimed at comparing language detection systems tested on tweets written in the 5 most spoken languages from the Iberian Peninsula (Basque, Catalan, Galician/Portuguese, and Spanish), and English. Some of the target languages are closely related: e.g. Spanish and Galician or Spanish and Catalan, and there are even varieties of the same language in two different spelling rules, e.g. Portuguese and Galician. So the systems are tested, not only on noisy short texts (tweets), but also on a set of texts written in very similar languages/varieties. As in DSL Shared Task, the best systems were also based on character $n$-grams and traditional classifiers.

In addition to $n$-gram models, other traditional approach with satisfactory results in language identification is that relying on ranked $n$-grams [31,32]. This approach relies on the observation that the most frequent $n$-grams are almost always highly correlated with the language. The rank-based measure sums up the differences in rankings of the $n$-grams found in the test data as compared to the training data. Rank-based systems seem to be stable across different domains and perform reasonably well on out-of-domain tests [26,33]. Ranking-based methods have also been applied successfully in machine learning to order classification algorithms [34].

Given that corpus-based $n$-grams are still the best way of building language models for language identification and classification, we will use them for quantifying the distance between languages, which is a task very similar to language identification.

## 3. Measures for computing language distance

We propose defining language distances using $n$-grams extracted from text corpora, in a very similar way as linguistic identification systems learn their language models. More precisely, two different $n$-gram-based coefficients to measure language distance are proposed: *perplexity* and *ranking*.

### 3.1. Perplexity

The most widely used evaluation metric for language models is the perplexity of test data. In language modeling, perplexity is frequently used as a quality measure for language models built with $n$-grams extracted from text corpora [35,36]. It has also been used in very specific tasks, such as to classify between formal and colloquial tweets [37].

Perplexity is a measure of how well a model fit the test data. More formally, the perplexity (called *PP* for short) of a language model on a test set is the inverse probability of the test set. For a test set of sequences of characters $CH = ch_1, ch_2, \ldots, ch_n$ and a language model $LM$ with $n$-gram probabilities $P(\cdot)$ estimated on a training set, the perplexity $PP$ of $CH$ given a character-based $n$-gram model $LM$ is computed as follows:

$$PP(CH, LM) = \sqrt[n]{\prod_{i}^{n} \frac{1}{P(ch_i|ch_1^{i-1})}} \tag{1}$$

where $n$-gram probabilities $P(\cdot)$ are defined in this way:

$$P(ch_n|ch_1^{n-1}) = \frac{C(ch_1^{n-1}ch_n)}{C(ch_1^{n-1})}. \tag{2}$$

Eq. (2) estimates the $n$-gram probability by dividing the observed frequency ($C$) of a particular sequence of characters by the observed frequency of the prefix, where the prefix stands for the same sequence without the last character. To take into account unseen $n$-grams, we use a smoothing technique based on linear interpolation.

A perplexity-based distance between two languages is defined by comparing the $n$-grams of a text in one language with the $n$-gram model trained for the other language. Then, the perplexity of the test text $CH$ in language $L2$, given the language model $LM$ of language $L1$, can be used to define the distance, $Dist_{perp}$, between $L1$ and $L2$:

$$Dist_{perp}(L1, L2) = PP(CH_{L2}, LM_{L1}). \tag{3}$$

The lower the perplexity of $CH_{L2}$ given $LM_{L1}$, the lower the distance between languages $L1$ and $L2$. The distance $Dist_{perp}$ is an asymmetric measure.

### 3.2. Ranking

The ranking-based distance derives from the observation that, for each language, there is a set of sequence of characters that make up a large portion of any text and their presence is to be expected as word distribution follows Zipf's law. Like in Cavnar and Trenkle's method [31], we used pruned $n$-grams profiles of two languages to be compared. $N$-grams are ranked according to frequency in a training corpus, and those with highest frequencies are selected while the rest are discarded. This gives us the pruned character $n$-grams profile for each language. A *language profile* is thus the ranked list of the most frequent $n$-grams in the training corpus. Unlike $n$-gram language models, language profiles do not make use of prior probabilities but simply of ranked lists.

The ranking-based distance between two languages is obtained by comparing the ranked lists of the two languages. It takes two $n$-gram profiles and calculates a simple rank-order statistic based on an "out-of-place" measure. This measure determines how far out of place an $n$-gram in one profile is from its place in the other profile [31]. More precisely, given the ranked profiles $Rank_{L1}$ and $Rank_{L2}$ of languages $L1$ and $L2$, respectively, $Dist_{rank}$ is computed as follows:

$$Dist_{rank}(L1, L2) = \sum_{\substack{i=1 \\ gr_i \in Rank_{L1}}}^{K} abs(Rank_{L1}(gr_i) - Rank_{L2}(gr_i)) \tag{4}$$

where $Rank_{L1}(gr)$ is the rank of a specific $n$-gram, $gr_i$, in the profile of $L1$, and $Rank_{L2}(gr_i)$ is the rank of the same $n$-gram in the profile of $L2$. Notice that the measure only considers those $n$-grams appearing in the profile of $L1$, which might also appear in that of $L2$. For those cases where the $n$-gram is not in the profile of $L2$, subtraction of zero is not a good solution since it gives low values for very frequent $n$-grams appearing only in $L1$. In such a case, we apply a smoothing technique which consists of subtracting the rank of the $n$-gram in $L1$ from the total size of the profile: $K - Rank_{L1}(gr_i)$.

The range of this measure is from 0 (identical profiles) to $K^2$ (entirely different ones). Like $Dist_{perp}$, the distance $Dist_{rank}$ is an asymmetric measure.

**Table 1**
Results for test A in DSL Shared Task 2016.

| Systems | Accuracy |
|---|---|
| Best (1) | 0.8938 |
| **Perplexity (2)** | **0.8926** |
| Median (9) | 0.8779 |
| Lowest (18) | 0.8240 |
| **Rank (19)** | 0.7940 |

**Table 2**
Results for test B1 in DSL Shared Task 2016.

| Systems | Accuracy |
|---|---|
| Best (1) | 0.920 |
| **Perplexity (5)** | **0.884** |
| **Rank (7)** | 0.804 |
| Median (9) | 0.688 |
| Last (18) | 0.530 |

## 4. Experiments

Our main objective is to use the language distance metrics defined above to build a current map of the European languages (Section 4.2). However, first we will evaluate the two metrics by applying them on the standard language identification task (Section 4.1).

### 4.1. Discrimination between similar varieties

The two distance metrics, $Dist_{perp}$ and $Dist_{rank}$, were used to build two language detection systems which were evaluated against the gold standard provided by the Discriminating Similar Languages Shared Task 2016 [25,38]. The objective is to compare our methods with the participant systems at the Shared Task, and observe how they behave when they are applied on the difficult task of discriminating between very closely related languages or similar varieties.

The State-of-the-art language identification systems perform very well when discriminating between unrelated languages on standard datasets. Yet, this is not a solved problem, and there are a number of scenarios in which language identification has proven to be a very challenging task, especially in the case of very closely related languages or varieties [29]. This is the scenario in which we are evaluating the systems based on our two distance metrics.

Tables 1–3 show the accuracy obtained by our two strategies (in bold) on the three tests of DSL Shared Task: test A consists of journal news as the training data used to build the language models (in-domain dataset), while tests B1 and B2 are constituted by tweets (out-domain dataset). The tables also contain three representative scores for each test: the best, the median, and the lowest accuracies achieved by the participants to the shared task. We specify the position of each system between brackets. This allows us to compare our techniques with the systems that participated to the DSL Shared Task 2016.

In test A (Table 1), our perplexity-based strategy would reach the second position, very close to the best system [39]. By contrast, the rank-based method would be the last one on this task. However, this system is very stable across domains, since it reaches similar scores in out-domain tests (see Tables 2 and 3), where its accuracy is now above the median. The accuracy of the perplexity system slightly decreases in the out-domain tests but it is still clearly above the median, being in total one of the best three systems in the shared task. Most systems yield mixed results across domains. For instance, the best system on test A is the 5th on tests B1 and B2, whereas the second one on test A is the 12th on tests B1 and B2.[2]

The results of these experiments show that our distance-based strategies, even though they were not primarily conceived for the task of language detection, are able to reach very competitive scores. More precisely, the perplexity-based distance is very close to the state-of-the-art measures in the specific task of identifying similar varieties.

### 4.2. Distance among the languages of Europe

In the following experiment, we use our distance metrics to build up a network linking forty-four European languages according to their current linguistic distances. This is a more natural application for the two metrics defined above. In this case, instead of a quantitative evaluation, we will provide a visual diagram and a qualitative analysis of the results.

2 The best system [40] on test B1 is also the best system on B2.

**Table 3**
Results for test B2 in DSL Shared Task 2016.

| Systems | Accuracy |
|---|---|
| Best (1) | 0.878 |
| **Perplexity (6)** | **0.820** |
| **Rank (7)** | 0.762 |
| Median (9) | 0.698 |
| Last (18) | 0.554 |

#### 4.2.1. Comparable corpora

The goal of the current experiment is to compare forty-four language models. In order to make them comparable, the texts from which they are generated should belong to similar domains and genres. Thus, we trained the models from *comparable corpora*, that is, from collection of documents in several languages which are not translations of each other, but which share the same genre and/or domain [41,42].

Two different comparable corpora for the 44 targeted languages were built.

The first corpora was built using the BootCat strategy defined in Baroni and Bernardini [43] and the corresponding Web tool[3] described in Baroni et al. [44]. BootCat is a method to automatically generate a corpus. It starts from a set of seed words which are sent as queries to a search engine. The resulting pages which are at the top of the search engine's hits pages are then retrieved and used to build a corpus [44]. To generate our BootCat comparable corpus, we used the same seed words (translated by means of Google Translate[4]) for the forty-four languages. Given a query in a particular language, most of the documents returned by the system are in the target language even though some of the seed words of the query were not well translated. The final corpus was manually revised and odd pages returned by the search engine were removed.

Following this, we divided the texts generated for each language in two parts: training and test corpora. We followed the same procedure for all languages in order to have the same size: the training corpus consists of a selection of $\sim$120$k$ tokens while the test is three times smaller: $\sim$40$k$.

The second comparable corpus was derived from different versions of the Bible. Recently, a parallel corpus based on 100 translations of the Bible has been created in XML format [24]. As this corpus does not cover all the European languages, we used additional sources[5] to fill out the same forty-four languages of the BootCat corpus. The train and test parts were created in the same manner as previously, except for those languages (e.g. Gaelic) whose Bible version is just a partial translation with few chapters. In those cases, the language is kept in the list even though the size of the training and test corpora does not reach the number of tokens we have established.

All languages were transliterated to the Latin script and normalized using a generic orthography. The encoding of the final spelling normalization consists of 34 symbols, representing 10 vowels and 24 consonants, designed to cover most of the commonly occurring sounds, including several consonant palatalizations and a variety of vowel articulation. The encoding is thus close to a phonological one.

Finally, we generated 7-gram models for all languages, which are the input of the language distances.

#### 4.2.2. Building language-to-language matrices

By applying the two distances, $Dist_{perp}$ and $Dist_{rank}$, on the language models (created from both the Web and the Bible corpora), we obtained four 44 $\times$ 44 matrices, each one derived from a distance-corpus strategy: *perp-web*, *perp-bible*, *rank-web*, and *rank-bible*. Since the two distance metrics are asymmetric, each matrix consists of 1936 different values.

We measured the similarity between the four distance-corpus methods by computing the Spearman correlation of the values they generated. Given two strategies, we compare whether their distance values are ranked in a similar manner. Table 4 shows the Spearman coefficient between each pair of methods. We observe that there is strong correlation (75.481) between the two methods based on perplexity, *perp-web* and *perp-bible*, even though they are applied on two very different corpora. When the two distances are applied on the same corpus, the correlation is moderate (65.087, 57.386). Not surprisingly, the correlation is lower (46.056, 33,934) if the two compared strategies are completely different. However, the correlation between the two rank-based strategies is quite weak: 46.256. It follows that, in this experiment, perplexity seems to be more stable across domains than the ranking distance.

#### 4.2.3. Language interaction network

As previously mentioned, in most works on historical linguistics the distance values among languages are computed from lists of words with a great stability in terms of form/meaning change. The inter-language distances are then supplied to hierarchical clustering algorithms to infer a tree structure for the set of languages. Hierarchical clusters and trees are intended to represent language families and phylogenetic evolution from a diachronic perspective. However, in our

3 WebBootCat is available at https://the.sketchengine.co.uk.

4 https://translate.google.com.

5 https://www.bible.com/.

**Table 4**
Spearman coefficient between pairs of methods.

| | Perp-web | Perp-bible | Rank-web | Rank-bible |
|---|---|---|---|---|
| Perp-web | 1 | **75.481** | 65.087 | 46.056 |
| Perp-bible | | 1 | 33,934 | 57.386 |
| Rank-web | | | 1 | 46.256 |
| Rank-bible | | | | 1 |

**Table 5**
Accuracy reached by the four language networks with the best *max* and *min* thresholds for each one (column 3).

| Networks | Accuracy | Thresholds |
|---|---|---|
| Perp-web | **.85** | min = 30, max = 100 |
| Perp-bible | **.85** | min = 50, max = 200 |
| Rank-web | .825 | min = 5, max = 10 |
| Rank-bible | .825 | min = 5, max = 10 |

work, language distance is not computed from pre-defined lists of stable and universal vocabulary, but from text corpora containing a great variety of linguistic phenomena including loan and foreign words. So, the language distance we have defined intends to measure interactions among languages from a synchronic perspective. The most suitable representation for this type of data is not a hierarchical tree but rather a network showing language interactions.

To create a visual language network, we need to identify true interactions between languages. Given a language and a list of languages ranked by their distance to the first one, we are required to distinguish between those that are actually related (by an arch in the network) to the given language and those that are so far that can be considered as unrelated. For this purpose, we select languages (nodes) and interactions (arcs) from each language-to-language matrix according to a set of filters and requirements (i.e. conditions). More precisely, given a target language, we create an arc with another language if their distance fulfills at least one of the two following conditions:

- It is lower than a minimum score.
- It is lower than a maximum score and is one of the two closest distances to the target language.

To set the optimum values of the two thresholds (minimum and maximum), we built a *gold standard* dataset consisting of 45 well-known language interactions annotated by a linguist who took into account the classification reported in Ethnologue [45]. Only interactions between languages by elaboration (*Ausbau* languages) were considered since they are clearly related. For instance, two examples of manually annotated interactions are the following:

| | | |
|---|---|---|
| Portuguese | Galician | 1 |
| Galician | Spanish | 2 |

The first row means that Galician is the closest language, rank 1, to Portuguese. The second row means that Spanish should be among the 2 closest languages to Galician, since this language is between Portuguese and Spanish. The gold standard dataset only contains language relationships that are well established in comparative linguistics. It is used as a reference test to evaluate the accuracy of all possible networks built from the four language-to-language matrices by using different thresholds. The threshold values giving rise to the highest accuracy are considered to build the best networks. In the end, we select the best network for each one of the four matrices. Table 5 shows the highest accuracy reached by each network (they are called by the name of the method used to create their original matrix). The last column shows the minimum and maximum values that maximize the accuracy. Table 6 shows a sample of languages with their two most similar languages and their distance.[6] The sample was extracted from the *perp-web* network.

To visualize language networks, we use Cytoscape, an open-source software designed to simulate biochemical reactions and molecular interactions [46]. Languages are attracted and disassociated in a similar way as to how molecules interact with each other. Fig. 1 is a network graph, with languages represented as nodes and inter-language interactions represented as links, that is, edges or arcs, between nodes. The length of each arc is a complex function that considers both the distance score between the two linked languages and the number of common nodes to which they are also linked [46].

#### *4.2.4. Analysis*

Fig. 1 shows that groups of languages having short distances and several internal arcs (only shared by the nodes of the group) tend to form a language family or sub-family: e.g. Romance, Slavic, Germanic, Celtic, Finno-Permian, or Turkic languages. The two groups with strongest internal cohesion (i.e. those having more internal links and shortest distances)

[6] The best network configuration was obtained by removing Romance languages from the ranked list associated to non-Romance ones. Given the strong Latin influence over many European languages, the distance between many non-Romance languages and those derived from Latin tend to be short.

**Table 6**
Sample of some languages extracted from the *perp-web* network. Only their two closest languages are shown (second column), as well as the distance score between each pair (third column).

| Target language | Closest languages | Distance |
|---|---|---|
| Bosnian | Croatian | 5 |
| Bosnian | Slovene | 8 |
| Bulgarian | Macedonian | 15 |
| Bulgarian | Serbian | 20 |
| Catalan | Spanish | 8 |
| Catalan | Galician | 10 |
| Croatian | Bosnian | 7 |
| Croatian | Serbian | 11 |
| Czech | Slovak | 9 |
| Czech | Slovene | 21 |
| English | French | 16 |
| English | Dutch | 31 |
| French | Catalan | 14 |
| French | Spanish | 15 |
| Georgian | Basque | 37 |
| Georgian | Serbian | 47 |
| Irish | Gaelic | 9 |
| Irish | English | 33 |
| Maltese | Italian | 24 |
| Maltese | English | 25 |
| Portuguese | Galician | 6 |
| Portuguese | Spanish | 8 |
| Serbian | Croatian | 13 |
| Serbian | Bosnian | 13 |
| Spanish | Galician | 6 |
| Spanish | Portuguese | 8 |
| Swedish | Danish | 12 |
| Swedish | Norwegian | 13 |
| Turkish | Azeri | 20 |
| Turkish | English | 46 |

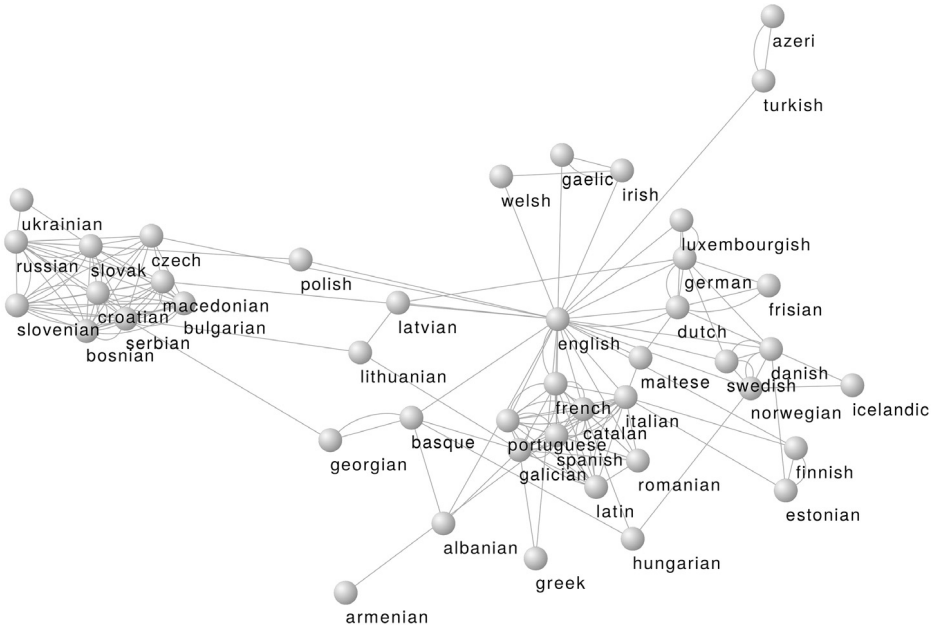


**Fig. 1.** Network of languages spoken in Europe. It has been built using the perplexity-based distance and the Web corpus (*perp-web* strategy).

are Romance and Slavic. However, Romance languages have a central position in the network since their elements are more connected to external nodes than the Slavic languages. The centrality of Romance language is explained by the fact that most languages have borrowed morphemes and lexical units from Latin in the past, and many neologisms from English

nowadays. Notice that a significant portion of English vocabulary (about 56%) comes from Romance languages, a portion of these borrowings come directly from Latin (15%) and another portion through French (41%) [47]. This makes English a special language between Romance and Germanic languages, as we can observe in Fig. 1. Moreover, it has many interactions with other languages from different families. English turns out to be the core of the map since it is the node with more connections to different sub-areas of the network.

The figure also shows us other interesting cases. Maltese, which is an Arabic language written in Latin alphabet, is interconnected with both English, the other national language in Malta, and Italian, probably because of its close geographical and cultural proximity.

Basque, a non-Indo-European language spoken between Spain and France, is identified by our distance measure as the closest language to Georgian (anyway the distance is quite high as can be observed in Table 6), belonging to the non-Indo-European Kartvelian family indigenous to the Caucasus. In fact, both languages are mutually related because Georgian is also identified as the closest non-Romance language to Basque, which is also strongly connected by our distance to Romance languages probably because of the great lexical influence of Latin and Spanish. Some controversial comparative-historical and typological approaches have tried to find a Basque–Caucasian connection [48]. However, according to other authors, the case for a link remains unproven, or even, they firmly rejected it [49].

It is also interesting to note that, in our network, Hungarian does not have any connections to Finnish and Estonian. Even if most historical linguists situate Hungarian as a member of the Uralic/Finno-Ugric family, it is also assumed that Hungarian is very detached from the Finno-Permic sub-family (Finnish, Estonian). Similarly, Fig. 1 also shows that Polish and the two Baltic languages (Lithuanian and Latvian), even though they belong to the Slavic family, are very far from the core of Slavic languages.

Finally, notice the network does not point at the presence of the Indo-European family. All languages, Indo-European or non-Indo-European, are somehow related either to the members of the family of Romance language or to English. As previously mentioned, our work does not intend to prove the existence of language families and historical relationships, but rather to show the existence of strong links and current interaction from a synchronic perspective.

## 5. Conclusions

To the best of our knowledge, this is the first time that models and methods from Language Identification have been applied to quantify the distance between languages. Basic *n*-gram models of characters extracted from text corpora can be used, not only for classifying languages or varieties as in the traditional task of language identification, but also for measuring the distance between language pairs in a global and quantitative way. We have shown that perplexity is an effective way of comparing models, but certainly not the only way. Other strategies, such as ranking-based methods can also be applied on the task of defining a distance measure working on *n*-grams.

We performed language comparison for forty-four European languages on the basis of two comparable corpora. We calculated the distances of $44 * *2$ language pairs and built a network that represents the current map of similarities and divergences among the main languages of Europe.

In many cases languages within the same family or sub-family have low distances as expected, but in some cases there are higher distances than one could expect for languages that are genetically related (e.g. Hungarian and Finish). The contrary is also true; we find low distances, as in the case of Maltese and Italian, for languages that are not related by phylogenetic links. This suggests that our quantitative measure can have applications not only on historical linguistics and the classification of language and language varieties, but also on NLP tasks such as machine translation, which requires knowing how close, or far apart, two languages are. This way, the choice of a specific machine translation strategy (e.g. rule-based, SMT, or neural-based) might rely on the distance between the source and target languages.

Finally, it is worth pointing out that our corpus-based strategy is just one more method to compute language distance, which should be seen as a complementary strategy to the existing ones. In particular, corpus-based *n*-grams might be seen as an additional linguistic source that complements the Swadesh list (and similar closed resources) used in phylogenetics and lexicostatistics. Unlike linguistic phylogenetics, which is focused on diachronic relationships, a *n*-gram method based on comparable corpora aims at relating languages from a synchronic perspective. The strategy defined in this article is an attempt to adapt the well-known and well-succeeded algorithms used in language identification to compute language distance. However, given that this is a complex and multidimensional task, further methods and strategies will be required to cover all the different aspects of languages. For instance, the use of delexicalized parsers trained and tested with different languages might be an interesting technique to compute the syntactic distance among them [50]. A more global strategy covering more linguistic aspects would be the use of the same technique in machine translation. Evaluating the translation quality of different target languages given the same source and the same models might provide us with a new quantitative metric for measuring the distance among languages.

Corpora and resulting datasets are freely available.[7]

[7] http://fegalaz.usc.es/~gamallo/resources/europe_languages.tar.gz.

## Acknowledgments

We would like to thank the linguist Marta Muñoz-González for her valuable help in building and cleaning the corpora as well as in setting the gold reference. This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), TADEEP (MINECO, ref:TIN2015-70214-P), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016–2019, ED431G/08) and the European Regional Development Fund (ERDF).

## References

[1] J. Nerbonne, E. Hinrichs, Linguistic distances, in: Proceedings of the Workshop on Linguistic Distances, LD'06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 1–6. URL http://dl.acm.org/citation.cfm?id=1641976.1641977.
[2] F. Petroni, M. Serva, Measures of lexical distance between languages, Physica A 389 (11) (2010) 2280–2283. URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:389:y:2010:i:11:p:2280-2283.
[3] F. Barbançon, S. Evans, L. Nakhleh, D. Ringe, T. Warnow, An experimental study comparing linguistic phylogenetic reconstruction methods, Diachronica 30 (2013) 143–170.
[4] J. Nerbonne, W. Heeringa, Measuring dialect distance phonetically, in: Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology, 1997, pp. 11–18.
[5] B. Chiswick, P. Miller, Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages, Discussion papers, IZA, 2004. URL https://books.google.es/books?id=nebHnQEACAAJ.
[6] H. Kloss, Abstand languages and ausbau languages, Anthropol. Linguist. 9 (7) (1967) 29–41.
[7] L. Nakhleh, D. Ringe, T. Warnow, Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages, Language 81 (2) (2005) 382–420.
[8] E. Holman, S. Wichmann, C. Brown, V. Velupillai, A. Muller, D. Bakker, Explorations in automated lexicostatistics, Folia Linguist. 42 (2) (2008) 331–354.
[9] D. Bakker, A. Muller, V. Velupillai, S. Wichmann, C.H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, E.W. Holman, Adding typology to lexicostatistics: A combined approach to language classification, Linguist. Typol. 13 (1) (2009) 169–181.
[10] M. Swadesh, Lexicostatistic dating of prehistoric ethnic contacts, Proc. Amer. Philos. Soc. 96 (1952) 452–463.
[11] C.H. Brown, E.W. Holman, S. Wichmann, V. Velupilla, Automated classification of the world's languages: a description of the method and preliminary results, Lang. Typol. Universals 61 (4) (2008).
[12] S. Wichmann, Genealogical classification in historical linguistics, in: M. Aronoff (Ed.), Oxford Research Encyclopedias of Linguistics, Oxford University Press, 2017.
[13] J. Nichols, T.J. Warnow, Tutorial on computational linguistic phylogeny, Lang. Linguist. Compass 2 (5) (2008) 760–820. URL http://dblp.uni-trier.de/db/journals/llc/llc2.html#NicholsW08.
[14] L.D. Michael, A Bayesian phylogenetic classification of tupí-guaraní, LIAMES 15 (2015).
[15] R. Gray, Q. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin, Nature (2011) URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&uid=14647380&cmd=showdetailview&indexed=google.
[16] F. Petroni, M. Serva, Language distance and tree reconstruction, J. Stat. Mech. Theory Exp. 2008 (08) (2008) P08012. URL http://stacks.iop.org/1742-5468/2008/i=08/a=P08012.
[17] R.D. Gray, F.M. Jordan, Language trees support the express-train sequence of austronesian expansion, Nature 405 (6790) (2000) 1052–1055. http://dx.doi.org/10.1038/35016575.
[18] C.J. Holden, R.D. Gray, Rapid radiation, borrowing and dialect continua in the bantu languages, in: P. Forster, C. Renfrew (Eds.), Phylogenetic Methods and the Prehistory of Languages, 2006, (Chapter 2). URL http://groups.lis.illinois.edu/amag/langev/paper/holden06phylogeneticMethods.html.
[19] T. Rama, S. Kolachina, B. Lakshmi Bai, Quantitative methods for phylogenetic inference in historical linguistics: An experimental case study of south central dravidian, Indian Linguist. 70 (2009).
[20] S. Wichmann, E.W. Holman, D. Bakker, C.H. Brown, Evaluating linguistic distance measures, Physica A 389 (17) (2010) 3632–3639. URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:389:y:2010:i:17:p:3632-3639.
[21] H. Liu, J. Cong, Language clustering with word co-occurrence networks based on parallel texts, Chinese Sci. Bull. 58 (10) (2013) 1139–1144.
[22] Y. Gao, W. Liang, Y. Shi, Q. Huang, Comparison of directed and weighted co-occurrence networks of six languages, Physica A 393 (C) (2014) 579–589. URL http://EconPapers.repec.org/RePEc:eee:phsmap:v:393:y:2014:i:c:p:579-589.
[23] E. Asgari, M.R.K. Mofrad, 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, San Diego, California, 2016, pp. 65–74. URL http://arxiv.org/abs/1604.08561.
[24] C. Christodoulopoulos, M. Steedman, A massively parallel corpus: the bible in 100 languages, Lang. Resour. Eval. 49 (2) (2015) 375–395. http://dx.doi.org/10.1007/s10579-014-9287-y.
[25] S. Malmasi, M. Zampieri, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task, in: Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, VarDial, Osaka, Japan, 2016.
[26] P. Gamallo, S. Sotelo, J.R. Pichel, Comparing ranking-based and naive bayes approaches to language detection on tweets, in: Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014, Girona, Spain, 2014.
[27] A. Zubiaga, I.S. Vicente, P. Gamallo, J.R. Pichel, I. Alegria, N. Aranberri, A. Ezeiza, V. Fresno, Tweetlid: a benchmark for tweet language identification, Lang. Resour. Eval. (2015) 1–38. http://dx.doi.org/10.1007/s10579-015-9317-4.
[28] M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, A report on the dsl shared task 2014, in: Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, VarDial, Dublin, Ireland, 2014, pp. 58–67.
[29] M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, P. Nakov, Overview of the dsl shared task 2015, in: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, LT4VarDial, Hissar, Bulgaria, 2015, pp. 1–9.
[30] A. Zubiaga, I.S. Vicente, P. Gamallo, J.R. Pichel, I. naki Alegria, N. Aranberri, A. Ezeiza, V. Fresno, Overview of tweetlid: Tweet language identification at sepln 2014, in: TweetLID - SEPLN 2014, Girona, Spain, 2014.
[31] W.B. Cavnar, J.M. Trenkle, N-gram-based text categorization, in: Proceedings of the Third Symposium on Document Analysis and Information Retrieval, Las Vegas, USA, 1994.
[32] R. Cordoba, L.F. D'Haro, F. Fernandez-Martinez, J. Macias-Guarasa, J. Ferreiros, Language identification based on *n*-gram frequency ranking, in: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 2007, pp. 27–31.
[33] P. Gamallo, I. Alegria, J.R. Pichel, M. Agirrezabal, Comparing two basic methods for discriminating between similar languages and varieties, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial3, 2016.
[34] P. Brazdil, C. Soares, A comparison of ranking methods for classification algorithm selection, in: Machine Learning: ECML 2000, 11th European Conference on Machine Learning, Barcelona, Catalonia, Spain, May 31–June 2, 2000, pp. 63–74. http://dx.doi.org/10.1007/3-540-45164-1_8.
[35] S.F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, in: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL'96, Association for Computational Linguistics, Stroudsburg, PA, USA, 1996, pp. 310–318. http://dx.doi.org/10.3115/981863.981904.

[36] R. Sennrich, Perplexity minimization for translation model domain adaptation in statistical machine translation, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL'12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 539–549. URL http://dl.acm.org/citation.cfm?id=2380816.2380881.
[37] M. González, An analysis of twitter corpora and the differences between formal and colloquial tweets, in: Proceedings of the Tweet Translation Workshop 2015, 2015, pp. 1–7.
[38] C. Goutte, S. Léger, S. Malmasi, M. Zampieri, Discriminating similar languages: Evaluations and explorations, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 2016.
[39] ÇağrıÇöltekin, T. Rama, Discriminating similar languages with linear SVMs and neural networks, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial3, 2016.
[40] B.D. Ayah Zirikly, M. Diab, The GW/LT3 VarDial 2016 shared task system for dialects and similar languages detection, in: COLING Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial3, 2016.
[41] X. Saralegi, I.S. Vicente, A. Gurrutxaga, Automatic generation of bilingual lexicons from comparable corpora in a popular science domain, in: LREC 2008 Workshop on Building and Using Comparable Corpora, 2008.
[42] P. Gamallo, J.R. Pichel, Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora, in: CICLING, in: LNCS, vol. 6008, Springer-Verlag, Iasi, Romania, 2010, pp. 473–483.
[43] M. Baroni, S. Bernardini, Bootcat: Bootstrapping corpora and terms from the web, in: Proceedings of LREC 2004, 2004, pp. 1313–1316.
[44] M. Baroni, A. Kilgarriff, J. Pomikálek, P. Rychlı, Webbootcat: a web tool for instant corpora, in: C. O. Elisa Corino, Carla Marello (Ed.), Proceedings of the 12th EURALEX International Congress, Edizioni dell'Orso, Torino, Italy, 2006, pp. 123–131.
[45] R.G. Gordon, J. Dallas, Ethnologue: Languages of the World (fifteenth ed.), IL International, 2005.
[46] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[47] J.M. Williams, In Origins of the English Language, The Free Press, 1975.
[48] N. Sturua, On the basque-caucasian hypothesis, Studia Linguist. 45 (1–2) (1991) 164–175.
[49] R.L. Trask, The History of Basque, Psychology Pres, 1997.
[50] J. Tiedemann, Cross-lingual dependency parsing for closely related languages, in: VarDial 2017, Valencia, Spain, 2017.