Levenshtein Distance as a measure of lexical proximity between languages

No Institute Given

Abstract. This paper intends to show the work in progress of a computational approach to the matter of studying the historical development of Portuguese in terms of its lexical distance from other languages. This distance is measured using the Levenshtein Distance algorithm between the target language and other languages that are related with its historical development, such as Spanish, French, Italian (Latin family of languages) and Arabic (historical contact from the earlier stages of the language) and at the same time, a comparison with historically unrelated languages, such as Dutch and Greek, in order to see if the method is a valid way to measure historical relations between languages. In order to obtain a cross-linguistic corpus with comparable lexical items we used Swadesh Lists in the chosen languages, and to eliminate the inherent problems of written form, such as comparing different writing systems, the work was based on IPA transcriptions of the lexical items in those languages. After obtaining the results for a language with solid historical documentation such as Portuguese and comparing what the computational method brought with the historical analysis de facto, the idea is to propose the automated analysis to languages that lack historical documentation, such as indigenous languages, in order to see if it is possible to assert something about their historical development and the relations between that language and its neighboring languages.

Keywords: Computational Linguistics, Comparative Linguistics, Phonology.

1 Introduction

This paper exposes the preliminary phase of the development of an automatic tool to measure the distance between different languages based on phonological differences when considering possible cognates.

The idea of defining a sort of distance between different languages is recurring throughout quantitative analysis in hitorical linguistics[1].

This notion of distance is empirically justified, since there are languages which are easily understandable for a native speaker of a given language and other languages that ar absolutely incomprehensible to the same speaker. Based on that, is it not reasonable to think that some languages are, in a way, closer than others?

Rather than throwing away this line of questioning, the linguist should consider how to find the formal aspects of the language that could verify or falsify this way of thinking.

When we analyse languages from a historical perspective, it becomes quite clear that there is such a thing as proximity and distance between languages. At some point, all the regions that nowadays speak languages such as Spanish and Italian in Europe were regions where Latin was the language spoken. At the same time, since we have access to a quite continous and regular written records from those regions, it becomes clear that at some point the Latin spoken in those regions became slightly different from what it once was, and this process is continous and hence we saw at some point different languages spoken in defferent regions. So we can clearly see a historical process of growing distance between what once was a single language community. This is a more than canonical view of language development, having a vast body of production such as viewed in [4].

It is important to note that one thing is the observable fenomenom of mutual comprehension, the fact that this can be related to the historical development of languages is a different process altogether: it involves different mechanisms and a more holistic analysis than a mere observation of facts. A third line of thought is if the empirical notion of distance is quantifiable and which method is the most accurate to quantify the concept.

We aim here to focus on the third notion, and later the goal is to apply the concept using different languages to enrich the discussion regarding the historical development of languages.

2 Development

2.1 Levenshtein Distance

Levenshtein's Distance (henceforth LD), is a metric of difference between two strings of characters, presented by Vladimir Levenshetein in[3].

The notion behind LD is rather simple: from a source string of characters, we calculate the minimum amount of basic steps (insertion of a character, deletion or replacement) that are needed to form the target string of characters.

The algorithm implemented for this paper was adapted from an implementation available on the internet, which guaranteed a more efficient way to processs the LD for any given two strings.[1].

[1] For a more in-depth discussion about implementing the LD algorithm, we recommend reading [2] or [5]

LD is widely used in natural language processing, it has a considerable efficiency in listing the best candidates to replace a mispelled word in a typed text. Also, because the algorithm is quite simple, it is an efficient way to test simple linguistic processing.

It is important to point out that there are already some works that have applied LD as a metric to attribute distance between languages, in specific, we point the works of [7] and [6], nonetheless, both articles does not cover the methodological traps of the resource or even discuss in depth the results obtained.

That being said, we will now cover the importance of choosing an appropriate set of words to attribute the distance between languages and also to recognize the shortcomings of any given standard set of words.

2.2 Corpus: Swadesh List

The Swadesh list is a lexical resource commonly used in lexicostatistics due to the fact that it is a well defined parallel corpus for several languages with a fixed number of words. The original proposal was made by the linguist Morris Swadesh throughout the 1950's [8] with lexical items considered somehow common to the majority of the human languages. For this work we used a version of the list with 207 items.

It is vital to bring up several aspects of the list that have been target to criticism: the notion that there is some sort of basic vocabulary that is shared by all languages is a very complicated notion, specially when it comes to words like 'freeze' (n. 145) and snow (n. 164). It is quite easy to imagine that a indigenous language from the Xingu river would not have such words, which would generate some sort of bias regarding the comparison amongst different languages.

Apart from the problem with basic vocabulary, the languages chosen for this first comparison are not too apart from each other and several of them actually hold some sort of relatedness which was helpful to assert the quality of the distante measured. The languages chosen were Portuguese (PT), which would be the base for the distance measurements, Spanish (ES), Italian (IT), French (FR), Dutch (DT) and Egiptian Arabic (EA). All words were transcribed using th International Phonetic Alphabet (IPA).

3 Partial results

Once that the set of languages was defined, the LD algorithm was executed on each word of a given pair of languages and then those results where normalized

by the medium lenght of the words in the target language.

Prior to applying the methods, we developed a rank of the languages regarding their proximity to Portuguese in terms of their historical relatedness (represented on the Expected column). After we measured the normalized results, we present as well the rank of the languages taking their distance to Portuguese (smaller avarage distance from swadesh list's words pairs) into account.

Expected	**Obtained**
Spanish	Spanish
Italian	Italian
French	French
Dutch	Greek
Greek	Dutch
Arabic	Arabic

Table 1: Comparison between the expected and obtained results for the distance between Portuguese and other languages using LD.

This shows an interesting precision for the method to locate related languages. It was capable of showing that Latin languages are closer, and that French is an outlier of the Latin group due to its historical phonological development. At the same time, the only difference between the expected list of languages and the obtained one is related to Greek, which is a language that is indeed more historically - in a phylogenetical way of thinking - distant from Portuguese than Dutch, but at the same time, Portuguese has a lot of Greek loans, which would justify the proximity between those languages.

4 Prospects

Finally, this in development research aims to provide a computational method to measure the distance between languages, which is a valid endeavour by itself. The final goal, though, is to develop a probabilistc classifier which would take into account several kinds of metrics and informations about two given languages and return the probability of they being related. Of course, this is a rather useless development to Indo-European languages, which have a rich written history that is used to provide historical informations. But for indigenous language this tool can actually provide new insights.

References

[1] Crowley, T., Bowern, C.: An Introduction to Historical Linguistics. OUP USA (2010), https://books.google.com.br/books?id=_N8v-s6fyt0C

[2] Jurafsky, D., Martin, J.H.: Speech & language processing. Pearson Education India (2000)

[3] Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)

[4] McMahon, A., McMahon, R.: Language classification by numbers. Oxford University Press on Demand (2005)

[5] Navarro, G.: A guided tour to approximate string matching. ACM computing surveys (CSUR) **33**(1), 31–88 (2001)

[6] Petroni, F., Serva, M.: Language distance and tree reconstruction. Journal of Statistical Mechanics: Theory and Experiment **2008**(08), P08012 (2008)

[7] Serva, M., Petroni, F.: Indo-european languages tree by levenshtein distance. EPL (Europhysics Letters) **81**(6), 68005 (2008)

[8] Swadesh, M.: Towards greater accuracy in lexicostatistic dating. International journal of American linguistics **21**(2), 121–137 (1955)