Olga Redko

Nancy Ide

CMPU 336

3 May 2020

Comparing the Lexicons of Romance Languages

Abstract

In this paper I propose a computational method for extracting translations in Romance languages that are highly orthographically similar to each other for a word. I disregard differences in diacritical marks and capitalization when calculating similarity scores, and I color-code diacritics to visually outline differences. The final product is a spreadsheet that orders translations based on orthographic similarity, and the purpose of this product is to help learners efficiently study the vocabulary of Romance languages for multilingual acquisition.

1. Problem and Motivations

According to the contrastive analysis hypothesis, where similarities between a learner's first language and second language occur, the acquisition of the second language would be easier compared with the situation in which there were differences between both languages (Benati & VanPatten, 2011). Inspired by this hypothesis, I confronted the problem of compiling a comparative list of multiple languages that could be utilized for language learning. I created a spreadsheet that indicates similarities as well as differences between the lexicons of seven Romance languages—Spanish, Portuguese, Italian, French, Romanian, Catalan, and Galician. By sorting the translations based on orthographic similarity in ascending order, I hoped to outline

similarities between different languages to aid learners with the acquisition of multiple Romance languages.

2. Overview of Existing Work

I was not able to find existing work on computationally gathering translations that are similar to each other in order to create a practical comparative list, but there has been an abundance of work in measuring similarities between foreign languages, often to determine the extent of their relationships. According to Campbell (2003), the most popular approaches employed for establishing relationships between languages are using methods based on the comparisons of sound correspondences and cognate lists. Ciobanu and Dinu (2014a) carried out a study that determined the orthographic similarity between Romanian and other languages as well as language families. They also carried out a study where they computed the orthographic similarity of Romance languages (Romanian, Italian, French, Spanish, and Portuguese) and clustered these languages based on their degrees of similarity (Ciobanu & Dinu, 2014). The recognition of similarities between different languages has also been utilized to develop machine translations between closely related minority languages such as Irish and Scottish Gaelic despite the lack of pre-existing bilingual lexical resources (Scannell, 2006).

3. Data, Algorithms, and Methodology

In this section I introduce a technique for determining the orthographic similarity of languages using the help of WordNet and MED (Minimum Edit Distance), also known as Levenshtein distance, calculations.

3.1 Data

I obtained a list of 3,000 common English words found on a webpage of Education First Australia ("3,000 Most Common Words in English"). I then utilized Princeton WordNet of English to obtain data of WordNets for my 3,000 collected English words. WordNets group words into synsets (cognitive synonyms) that each express a different concept, and synsets are interlinked by means of conceptual-semantic and lexical relations ("What is WordNet?"). I then used the WordNet corpus reader to gain access to the Open Multilingual WordNet, which allowed me to use foreign WordNets that were all linked to the Princeton WordNet of English. By linked, it is meant that the synsets of the foreign WordNets were created in correspondence with the Princeton WordNet synsets whenever possible, and semantic relations were imported from the corresponding English synsets.

The Romance language WordNets that I used were created by different projects and varied in size and accuracy. Multilingual Central Repository created the Spanish, Catalan, and Galician WordNets; OpenWN-PT created the Portuguese WordNet; MultiWordNet created the Italian WordNet; WOLF created the French WordNet; and finally, Romanian WordNet created the Romanian WordNet (Bond).

3.2 Methodology and Algorithm

I would like to clarify that when I say "word," I refer to one of the 3,000 analyzed English words, and when I say "lemma," I refer to one of the elements found in the translated synsets of a word. First, I created eight empty lists, where each list was reserved for collecting the contents of synsets for each language. Let us call each of these lists "X_list," where "X" denotes one of the eight languages like "English" and "Spanish." Then, for each word, I extracted the lemmas of its non-English translated synsets and added them as elements to a list that was later appended to

the corresponding X_list; English_list was an exception that simply had a list of lists where each list within the bigger list had a single English word from the list of 3,000 English words. After the X_lists were filled, I filled in each empty list within these lists with a string of a hyphen to prevent indexing errors, and I replaced all cases of "_" with " " so that the elements would be more convenient for humans to read (e.g. "de_alguna_manera" would become "de alguna manera"). Afterwards, I preprocessed the X_lists, forming preprocessed_X_lists, by stripping off diacritics with Unidecode and by converting uppercase letters to lowercase letters. I stripped off diacritics because from an orthographic perspective, the resemblance of words is higher between words without diacritics than between words with diacritics.

Initially, for each lemma of language X, I calculated MED scores between this lemma and the lemmas of the same corresponding word from other languages except for English, and I kept track of the index of the lemma for X of each word that had the lowest calculated MED score, and I kept track of what the lowest, or minimum, calculated MED score was. The lemmas I extracted from this led to noticeably poor results (i.e., the individual lemmas I selected tended to be egregiously different from each other)—this is because the selected lemmas were not solely compared to lemmas that had already been selected. So, this often led to cases where a lemma was extracted because it was very similar to another foreign lemma—a foreign lemma that was ultimately not extracted because a very different lemma within the same language found greater similarity with the lemma of a third language.

To alleviate this problem, I ensured that, aside from the first extracted lemma, the lemmas of each language would only be compared to lemmas that had already been extracted. The order of languages I chose to do this was

Spanish→Portuguese→Italian→French→Romanian→Catalan→Galician. This might have not been the most ideal order for obtaining lemmas that were most similar to each other, but it led to markedly better results, and finding an ideal order by going through every permutation of the languages would have been excessively demanding within the scope of my project and likely would not have led to significantly better results.

Through my refined method, for the first language, Spanish, I calculated MED scores between its lemmas and the lemmas of the same corresponding word from other languages except for English, and I kept track of the index of the Spanish lemma of each word that had the lowest calculated MED score, and I kept track of what the lowest, or minimum, calculated MED score was. I did this instead of comparing Spanish lemmas solely to lemmas of the second language, Portuguese, to ensure that Spanish and Portuguese would not have set an overpowering precedent in the selection of other lemmas. Afterwards, I calculated MED scores between Portuguese lemmas and the extracted Spanish lemma of the same word with the lowest MED score, and I kept track of the index of the Portuguese lemma of each word with the lowest calculated MED score, and I kept track of what the lowest, or minimum, calculated MED score was. A generalized description of the method is that for the *n*th language, where n is an integer $1 < n < 8$, I calculated MED scores between the lemmas of the *n*th language and the extracted lemmas of all the *m*th languages for each word, where m is all the integers between 1 and n-1, and I kept track of the index of the lemma of the nth language for each word with the lowest calculated MED score, and I kept track of what the lowest, or minimum, calculated MED score was.

After creating a list of lists for each language, where each list within a list contained the index of the lemma of a word with the lowest MED score followed by the lowest, or minimum, MED score, I pickled each of these lists to save time. Each pickled list was saved in the format "X_MED_list.txt." Afterwards, using the information of indices, I extracted lemmas from X_lists (i.e., I gathered lemmas that were not preprocessed), and, for each word, I calculated and recorded the sum of the lowest, or minimum, MED scores of lemmas that were associated with the word. A sum of the minimum MED scores for a word, in other words, provides information about the sum of the smallest MED scores between the extracted lemma of the *n*th language and the extracted lemmas of the *m*th languages for each language for a word. I skipped over words that did not have lemmas for all seven Romance languages due to the incompleteness of WordNets; this lowered the number of considered words from 3,000 to 1,818. I created a list of lists from this information, where each list within the list contained an English word followed by its corresponding extracted lemmas from each of the Romance languages and ending in the corresponding summed minimum MED score. I sorted the list based on the summed minimum MED scores.

Then, using XlsxWriter, I created a .xlsx spreadsheet called "comparison_sheet_colored.xlsx" with English words along with corresponding lemmas for Romance languages. Each row had the following format, where column types are separated by commas: "English word, corresponding Spanish lemma, corresponding Portuguese lemma, corresponding Italian lemma, corresponding French lemma, corresponding Romanian lemma, corresponding Catalan lemma, corresponding Galician lemma, summed minimum MED score." For example, "peace, paz, paz, pace, paix, pace, pau, paz, 5." Rows were ordered by ascending

summed minimum MED scores. I also color-coded diacritics to emphasize differences. I color-coded letters with the acute with red; grave with blue; circumflex with green; cedilla with yellow; diaeresis with brown; tilde with pink; breve with orange; and comma with purple. Finally, my product was complete.

4. Results

In this section I present some statistics that reflect the characteristics of my collected data.

Summed minimum MED score	Number of cases	Percentage of cases
0-4	290	15.95%
5-9	581	31.96%
10-14	506	27.83%
15-19	273	15.02%
20-24	126	6.93%
25-29	30	1.65%
30+	12	0.66%

Table 1: Number and percentage of cases of certain summed minimum MED scores out of 1,818 cases.

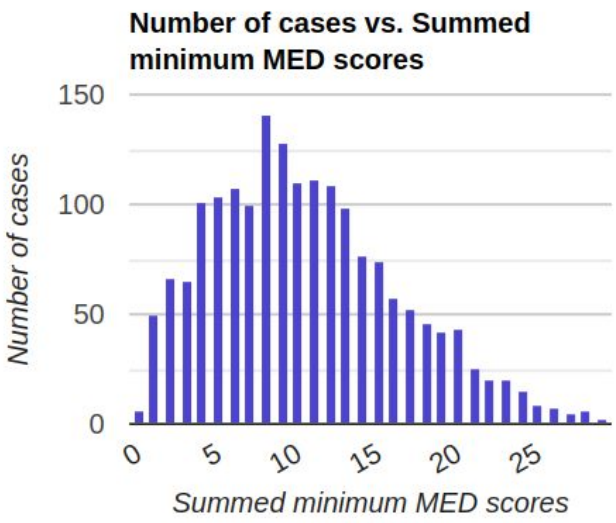

Figure 1: Bar graph of number of cases of certain summed minimum MED scores from 0-29.

The range of summed minimum MED scores was 0 to 43, the mean summed minimum MED score was 10.67, the median summed minimum MED score was 10, and the mode was 8.

	Summed minimum MED scores within a language	Average minimum MED score for a lemma within a language
Spanish	241	0.13
Portuguese	2502	1.38
Italian	4332	2.38
French	4056	2.23
Romanian	3945	2.17
Catalan	1356	0.75
Galician	2962	1.63

Table 2: Sums and average scores per extracted lemma of minimum MED scores for each language. Each language contained 1,818 lemmas and minimum MED scores.

Considering the methodology used, it is not surprising that extracted Spanish lemmas have a relatively low average minimum MED score—Spanish lemmas, unlike the lemmas of other Romance languages, were selected through calculations and comparisons between Spanish lemmas and all other non-English lemmas for a word, including lemmas that were not ultimately extracted for the spreadsheet. Since Spanish lemmas were compared with a larger group of lemmas than the lemmas of other languages, it makes sense that extracted Spanish lemmas were more often compared with lemmas with low MED scores.

4. Analysis of Shortcomings and Ideas for Future Research

There are several shortcomings in my project that can be improved upon through additional research and time. Some of my shortcomings came from the quality of WordNets I used—the WordNets I used were incomplete and sometimes inaccurate. Due to the incompleteness of the WordNets I used, my final results only displayed 1,818 words instead of the originally expected 3,000 words.

Some of the extracted lemmas were also highly impractical to use and were misleading. For example, some extracted lemmas that corresponded to the English word “bond” included “James Bond.” This is a proper noun that virtually has nothing to do with the typical meanings of “bond,” such as, for example, “something that binds or restrains” (Bond, 2020). To reduce this problem, I could filter out proper nouns or implement a method to only consider lemmas that pass a certain frequency threshold (this would filter out lemmas that are quite infrequent and likely impractical to learn).

Another shortcoming of my project is the methodology I used. Although I am fairly satisfied with the generated spreadsheet considering the limitations I had and the relatively little time it took to carry out computations, it is highly unlikely that I chose the best lemmas for minimizing edit distances between all lemmas for a corresponding word. I likely would have been closer to reaching this goal if I tried going through every permutation of the Romance languages when extracting lemmas and, in the end, searching for whichever permutation had the lowest total summed minimum MED scores. A more useful way to calculate minimum MED scores also would have been to calculate MED scores between all of the extracted lemmas for a word instead of only calculating MED scores between lemmas that had been extracted before the lemma I am in the process of extracting.

References

Benati, A. G., & VanPatten, B. (2011). Key Terms in Second Language Acquisition. *International Journal of Applied Linguistics*, 270–273. doi: 10.5040/9781474227544.ch-004

Bond. (2020). In *Merriam-Webster.com.* Retrieved May 8, 2020, from https://www.merriam-webster.com/dictionary/bond

Bond, F. (n.d.). Open Multilingual Wordnet. Retrieved from http://compling.hss.ntu.edu.sg/omw/

Campbell, L. (2003). How to Show Languages are Related: Methods for Distant Genetic Relationship. *The Handbook of Historical Linguistics*, 262–282. doi: 10.1002/9780470756393.ch4

Ciobanu, A. M., & Dinu, L. P. (2014a). An Etymological Approach to Cross-Language Orthographic Similarity: Application on Romanian. Retrieved from https://www.aclweb.org/anthology/D14-1112/

Ciobanu, A. M., & Dinu, L. P. (2014b). On the Romance Languages Mutual Intelligibility. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1183_Paper.pdf

Scannell, K. (2006). Machine Translation for Closely Related Language Pairs. *Proceedings of the Workshop on Strategies for Developing Machine Translation for Minority Languages,* 103–107.

3,000 Most Common Words in English. (n.d.). Retrieved from https://www.ef-australia.com.au/english-resources/english-vocabulary/top-3000-words/

What is WordNet? (n.d.). Retrieved from https://wordnet.princeton.edu/