On the Syllabic Similarities of Romance Languages

Anca Dinu[1] and Liviu P. Dinu[2]

[1] University of Bucharest, Faculty of Foreign Languages,
5-7 Edgar Quinet, 70106, Bucharest, Romania, anca_radulescu@yahoo.com
[2] University of Bucharest, Faculty of Mathematics and Computer Science,
14 Academiei, 70109, Bucharest, Romania, ldinu@funinf.cs.unibuc.ro

Abstract. In this paper we study the syllabic similarity between Romance languages via rank distance. The results confirm the linguistical theories, bringing a plus of quantification and rigor.

1 The Syllabic Similarity of Romance Languages

The problem of classifying Romance languages is an intensely studied issue. Unfortunately, in many studies of this kind, the data referring to Romanian are incomplete or even missing (as it happens, for example, in Ziegler, 2000). Here we study the "syllabic" similarity of Romance languages. The work corpus is formed by the representative vocabularies of Romance languages (Latin, Romanian, Italian, Spanish, Catalan, French and Portuguese languages) (Sala, 1988). We syllabified the vocabularies. For each vocabulary we constructed a classification of syllables: on the first position we put the most frequent syllable of the vocabulary, on the second position the next frequent syllable, and so on.

The method we applied in investigating the syllabic similarity of Romance languages is the following: each of the seven Romance languages is compared to the other six (using rank distance (Dinu, 2003)), for each comparison having a graphic as a result. We apply the normalized rank distance between all pairs of such classifications and we obtain a series of results which express the "syllabic" similarity between Romance languages. We also investigate the distances between partial classifications. Each graphic represents the behavior of the function $f_{\Delta}(i)$ with i varying between 1 and 561 (the minimum number of syllables correspondent to the Latin language). The function f_{Δ} expresses the variation of the normalized rank distance between two classifications (see Appendix).

We chose this method for the following reasons: when a listener hears for the first time a language, it is difficult to believe that he is able to distinguish syntactic constructions or even words. In fact, it is more plausible that he can distinguish and individualize syllables; due to this fact, he is able to say to which language (or family of languages) the language he hears is similar.
In Table 1, we present the number of distinct syllables (*types*) and the number of all the syllables (*tokens*) from every language analyzed. The frequency of the syllables from every language is not uniformly distributed. Table 1 shows