IS CATALAN SYLLABLE-TIMED? AN ANSWER BASED ON CLUSTER ANALYSIS

Paolo Roseano[1,2], Patrícia Marsà Morales[3], Laura Alañá Vilas[1]

[1]Universitat de Barcelona, [2]University of South Africa, [3]Universitat Oberta de Catalunya
paolo.roseano@ub.edu

ABSTRACT

Studies about rhythm often try to determine which rhythm class a language belongs to. In this paper, we argue that cluster analysis is a method that can provide a statistically motivated clear-cut answer to this kind of question. In order to test this method, we first used *Correlatore* (Mairano & Romano, 2010) to compute nine rhythm metrics (namely V%, ΔC, ΔV, varcoΔC, varcoΔV, CrPVI, VrPVI, CnPVI, and VnPVI) for Catalan, Spanish, Italian, Southern British English and Mainstream American English. Such metrics were then used as variables in a cluster analysis that was carried out with Ward's method by means of *Gabmap* (Nerbonne, Colen, Gooskens, Kleiweg, & Leinonen, 2011). The outcome of the cluster analysis was plotted on a dendrogram. The results suggest that Catalan is syllable-timed, since it clusters with Spanish and Italian.

Keywords: rhythm, rhythm class, cluster analysis, Catalan

1. INTRODUCTION

The aim of this paper is twofold. On the one hand, it aims at contributing to the discussion about the rhythmic properties of Catalan, a language whose classification has been an object of discussion since the '90s (Nespor, 1990). On the other hand, it aims at bringing about a methodological contribution, insofar as it suggests that cluster analysis can be a useful tool for the classification of languages according to their rhythm.

The paper is organized as follows. Section 2 summarizes a few concepts about the rhythm that will be employed to motivate the use of cluster analysis. Section 3 presents the method, while Section 4 shows the most important results. Some conclusions are put forward in Section 5.

2. CLASSIFICATION METHODS IN RHYTHM STUDIES

In order to understand why cluster analysis can be an appropriate statistical tool in rhythm studies, one has to focus on some commonly accepted ideas in this field. Although studies about rhythm have undergone important paradigm shifts since the concept of linguistic rhythm appeared in the first half of the 20th Century (for an overview see Mairano, 2011), the majority of authors nowadays would agree on some basic concepts:

- Languages have different rhythmic patterns, and such differences can be measured.
- There is a linguistic continuum between (at least) two ideal poles of the continuum: stress-timed languages and syllable-timed languages.
- Some languages are commonly acknowledged to be stress-timed (e.g. Southern British English, German or Dutch); other languages are generally considered syllable-timed (e.g. Central Peninsular Spanish or Mandarin Chinese).

In spite of the fact that authors agree that there is a continuum, when researchers study the rhythmic properties of a specific language, they often ask themselves –in a more or less overt manner– which rhythmic *class* such language belongs to: is it syllable-timed or stress-timed? Trying to answer this question means trying to split the *continuum* in (at least) two *classes* (stress-timed and syllable-timed languages) or –which is the same– trying to *classify* rhythmic variation.

The methods used to classify languages rhythmically (i.e. to assign them to a rhythm class) are of two kinds: *graphic* methods and *statistical* methods. Both types of methods have advantages and disadvantages.

2.1. Graphic methods

Graphic methods have been used since the very first studies that employed rhythm metrics (Ramus, Nespor, & Mehler, 1999; Grabe & Low, 2002). These graphics are representations on a 2D Cartesian coordinate system where the axes show the values of the metrics. 2D representations are chosen for obvious practical reasons, but authors are aware that the best solution would be a representation in a multi-dimensional space (Ramus et al., 1999, p. 272). In the case of Catalan, graphic representations can be found in different studies, e.g. in Gavaldà-Ferré (2007) (Figure 1) or in Marsà and Roseano (2019) (Figure 2).

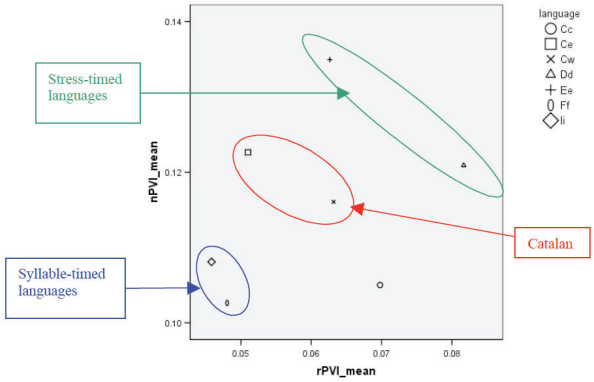

Figure 1: Distribution of languages in the nPVI/rPVI plane according to Gavaldà-Ferré (2007); Catalan marked in red.

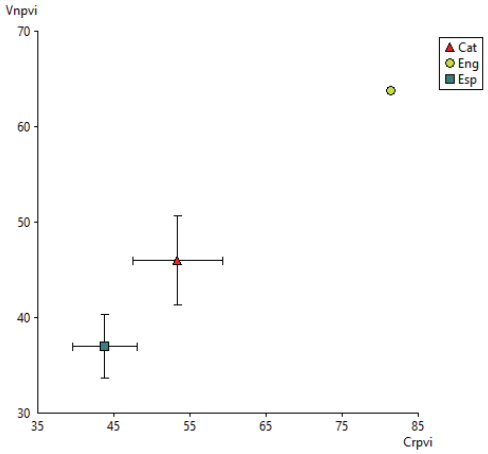

Figure 2: Distribution of languages in the nPVI/rPVI plane according to Marsà and Roseano (2019); Catalan marked in red.

Such representations usually include languages that represent the *poles* of the continuum (e.g. English or Spanish) and the language whose rhythmic properties are analysed. The representations are accompanied by non-quantified comments on the distance between the language studies and the poles of the continuum, e.g. "both dialects of Catalan are grouped together with the syllable-timed languages –Italian and French–, while the stress-timed languages are far away" (Gavaldà-Ferré, 2007, p. 28) or "our data allow us to conclude that Catalan is much closer to Spanish than to English" (Marsà & Roseano, 2019, p. 69).

The main advantage of graphic methods consists in the fact that 2D representations are persuasive and easy to interpret. Their possible disadvantages are two: 1) the clusters of languages are defined on the base of visual inspection of the graph and are not validated quantitatively; 2) graphs have only 2 dimensions corresponding to 2 metrics, even in the studies where three or more rhythm metrics are considered (e.g. Ramus et al., 1999).

2.2. Statistical methods

In order to overcome the limitations of the results provided by graphic analyses, different kinds of statistical methods have been used, though with different goals.

For example, Ramus et al. (1999, p. 272) use ANOVA to test to what extent the classification they had previously obtained using %V and ΔC was reliable. Tortel and Hirst (2010) use discriminant analysis to determine which rhythm metrics allow separating three predefined groups of native and non-native speakers of English. Loukina, Kochanski, Rosner, Keane, and Shih (2010) use machine classification to identify different languages from their rhythm measures. Crucially, in the studies we have mentioned so far, the objective of the statistical analysis was not exactly classifying languages in rhythm classes (at most, they aimed at validating an *existing* classification).

The studies that do use statistics to classify a language usually carry out metric-by-metric analyses. For example, Gavaldà-Ferré (2007) uses ANOVA and post hoc tests to check whether the language to be classified (Catalan) differs from the reference languages in 5 rhythm metrics taken individually. Similarly, Prieto, Vanrell, Astruc, Payne and Post (2012) use a Generalized Linear Mixed Model to test whether Catalan differs from English and Spanish in 7 rhythm metrics taken one by one.

To the best of our knowledge, there has been only one attempt to carry out a multi-parameter statistical analysis to classify a language rhythmically: Pukevičiūtė and Kazlauskienė (2013) use cluster analysis to determine which rhythm class Lithuanian and Latvian belong to. In order to do so,

they take five metrics (ΔV, ΔC, %V, varcoΔV, varcoΔC) as variables, and use data from other languages considered to be the prototypical examples for the two rhythmic classes.

The use of cluster analysis has several advantages in comparison to the previous approaches. The most noteworthy of them is that it provides a clear-cut answer to the research question 'Is language X stress-timed or syllable-timed?', and such answer is based on a statistical analysis.

3. METHODOLOGY

3.1. Speakers, recordings and acoustic analysis

We used recordings of the language to be classified (6 speakers of Central Catalan), in addition to the recordings of 2 syllable-timed varieties (6 speakers of Southern British English, 6 speakers of Mainstream American English) and 2 syllable-timed languages (6 speakers of Central Peninsular Spanish and 6 speakers of Italian). All speakers are native and were asked to read *The North Wind and the Sun* in their respective languages.

As usual in studies on rhythm, vocalic and consonantal intervals were annotated in a *Praat* textgrid (Boersma & Weenink, 2019) for each recording.

3.2. Statistical analysis

The statistical analysis was carried out in two stages. In the first stage, *Correlatore* (Mairano & Romano, 2010) was used to calculate nine rhythm metrics that had been put forward in classical studies about rhythm: V%, ΔC, ΔV, varcoΔC, varcoΔV, CrPVI, VrPVI, CnPVI, VnPVI (Ramus et al., 1999; Grabe & Low, 2002; Dellwo & Wagner, 2003).

In the second stage, *Gabmap* (Nerbonne, Colen, Gooskens, Kleiweg, & Leinonen, 2011) was used. *Gabmap* is a free web-based application which has been designed for dialectometric studies. In addition to being able to process alphabetic data (like phonetic transcriptions of words), it also allows the analysis of numeric data (like those of rhythm metrics provided by *Correlatore*). The distance calculated for numeric data is the Euclidean one, while the linkage method several options are offered: Ward's minimum variance method, group average, weighted average, single.

The two basic kinds of statistica analyses that *Gabmap* carries out are hierarchical cluster analysis and multidimensional scaling. The two techniques are complementary. Multidimensional scaling *per se* does not group the varieties, but represents them on a continuum, in a way which is very similar to the plots generally used in rhythmic studies (like the ones in Figures 1 and 2). On the other hand, cluster analysis does classify varieties by creating groups or classes. In addition, *Gabmap* also provides graphic representations of the results of the statistical analyses it carries out.

By means of *Gabmap*, we carried out two different analysis:

- Like Pukevičiūtė and Kazlauskienė (2013), we run a cluster analysis. Nevertheless, differently from the above-mentioned authors, we plotted the results of the cluster analysis on dendrograms, that are easy to interpret. From a methodological point of view, one has to point out that for our analysis we have chosen Ward's linkage method, which minimizes the total within-cluster variance and, therefore, favours the creation of homogeneous clusters.
- In addition to the cluster analysis used by Pukevičiūtė and Kazlauskienė (2013), we also used MDS plots, which are similar to 2D Cartesian representations typically used in rhythmic studies. The main difference between 2D plots (like the ones in Figure 1 and 2) and MDS plots is that the latter are based on the analysis of more than two metrics. To put it in intuitive terms, an MDS plot is the simplification in 2D of a N-dimensional space. In our case, since we have 9 metrics, our MDS plot will be the reduction on a 2D plot of a 9-dimensional space.

4. RESULTS

The values of the metrics obtained by means of *Correlatore* (Table 1) were uploaded in *Gabmap*, which generated the MDS plot in Figure 3 and the dendrogram in Figure 4.

Table 1: Values of the rhythm metrics for Central Catalan (CAT), Central Peninsular Spanish (CPS), Italian (ITA), Mainstream American English (MAE) and Southern British English (SBE).

	CAT	CPS	ITA	MAE	SBE
V%	44.51	45.36	51.15	39.67	44.48
ΔV	36.81	27.68	44.27	51.01	36.55
ΔC	46.80	38.61	41.09	68.67	68.85
varcoΔV	53.30	40.32	50.53	58.70	60.44
varcoΔC	55.54	47.29	47.58	53.23	53.74
VnPVI	45.99	36.22	44.57	62.63	65.99
CnPVI	63.54	54.76	56.28	60.63	60.80
VrPVI	34.42	26.06	41.35	55.89	71.26
CrPVI	53.39	44.06	48.87	76.48	76.33

This MDS allows us to get an approximate idea of the distribution of the five varieties in the 9-dimensionl space. The MDS plot gives a result that is very close to Gavalda-Ferré (2007) and Marsà and Roseano (2019) (see Figures 1 and 2): graphically speaking, Catalan ends up between stress-timed and syllable-timed languages, although it seems to be closer to stress-timed languages.

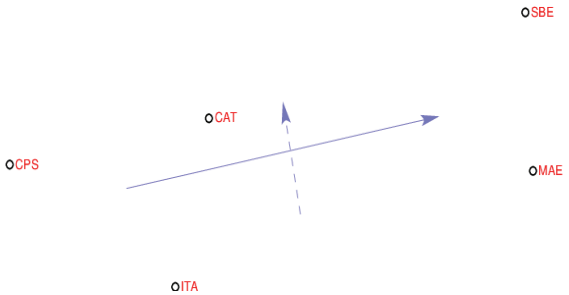

Figure 3: MDS plot based on the values of rhythm metrics for Central Catalan (CAT), Central Peninsular Spanish (CPS), Italian (ITA), Mainstream American English (MAE) and Southern British English (SBE) (*r* = 1.0).

Nevertheless, the MDS plot does not give a clear-cut answer to the research question "Is Catalan stress-timed or syllable-timed"?". On the contrary, the dendrogram corresponding to the cluster analysis (Figure 4) allows the researcher to easily find an unequivocal answer. In fact, Catalan appears grouped up with syllable-timed languages (Central Peninsular Spanish and Italian).

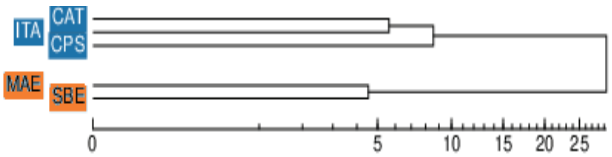

Figure 4: Dendrogram for Central Catalan (CAT), Central Peninsular Spanish (CPS), Italian (ITA), Mainstream American English (MAE) and Southern British English (SBE).

5. DISCUSSION

As we stated in the introduction, the aim of this paper was twofold. On the one hand, it aimed at contributing to characterize from a rhythmic point of view Catalan, a language whose classification has been an object of discussion since the '90s (Nespor, 1990). On the other hand, it intended to bring about a methodological contribution.

As far as the second objective is concerned, we argue that cluster analysis is a useful method in studies about rhythm because it can provide an answer to the research question "What rhythm class does language X belong to?" and such answer is 1) statistically motivated and 2) clear-cut. MDS plots, on the other hand, do not display these advantages, insofar as they are statistically-based, but do not offer a clear-cut answer to the research question.

As far as the first objective is concerned, our analysis shows that Catalan is clustered with the two syllable-timed varieties analysed (Central Peninsular Spanish and Italian), and not with the two stress-timed varieties (Southern British English and Mainstream American English). This suggests that our data confirm the claim by Ramus, Dupoux and Mehler (2003) and Marsà and Roseano (2019) that Catalan is a syllable-timed language, and not an intermediate language or a language with mixed syllable/stress-timed features like other studies mentioned above argue.

6. FINAL REMARKS

Although the results obtained by means of cluster analysis are encouraging, further work is needed in order to solve some issues raised by this technique.

The first important issue has to do with the amount of gathered data: cluster analysis makes sense and performs well only if the number of objects to be classified is high. This implies that the results of the cluster analysis will be more solid when we add several more languages to the database.

The second point that needs to be solved is to find out which metrics should better be used in the cluster analysis and which should be avoided (and why).

Finally, one will need to consider the possible advantages of running a cluster discriminant analysis to determine which metrics define better each of the two clusters. This could also help shed some light on the capacity of the different rhythm metrics to capture the rhythmic properties of languages.

6. REFERENCES

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer [Computer program]. Version 6.1.07*, retrieved from http://www.praat.org/

Dellwo, V., & Wagner, P. (2003): Relations between language rhythm and speech rate. In D. Recasens, M. J. Solé, & J. Romero (Eds.), *Proceedings of the 15th ICPhS* (pp. 471-474). Barcelona: Universitat Autónoma de Barcelona.

Gavaldà-Ferré, N. (2007). *Vowel reduction and Catalan speech rhythm* (MA dissertation, University College London, London, United Kingdom).

Grabe, E., & Low E. L. (2002): Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven, & N. Warner (Eds.), *Papers in Laboratory Phonology 7*, (pp. 515-546). Berlin: de Gruyer.

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2010). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America*, *129*, 3258-3270.

Mairano, P. (2010). *Rhythm Typology: Acoustic and perceptive studies* (Doctoral dissertation, Università degli Studi di Torino, Torino, Italy).

Mairano, P., & Romano, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. In Schmid, S., Schwarzenbach, M., & Studer, D. (Eds.), *La dimensione temporale del parlato, (Proc. of the V National AISV Congress, University of Zurich, Collegiengebaude, 4-6 February 2009)* (pp. 79-100). Torriana (RN): EDK.

Marsà, P., & Roseano, P. (2019). El ritme del català: anàlisi a partir de textos fonèticament equilibrats. *Estudios de Fonética Experimental*, *28*, 47-79.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., & Leinonen, T. (2011). Gabmap: A web application for dialectology. *Dialectologia, Special Issue II*, 65-89.

Nespor, M. (1990). On the rhythm parameter in phonology. In I. M. Roca (Ed.), *Logical issues in language acquisition* (pp. 157-175). Dordrecht: Foris.

Prieto, P., Vanrell, M.M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English, and Spanish. *Speech Communication 54*(6), pp. 681-702.

Pukevičiūtė, A., & Kazlauskienė, A. (2013). Klasterinė tipologinių kalbos ritmo grupių analizė. *Lietuvių Kalbotyros Klausimai, 69*, 168-184.

Ramus, F., Nespor, M., & Mehler, J. (1999): Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 263-292.

Ramus, F., Dupoux, E., & Mehler, J. (2003): The psychological reality of rhythm classes: Perceptual studies. Paper presented at the *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 3rd-9th August 2003.

Tortel, A., & Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. In Hasegawa Johnson, M. (Ed.), *Proceedings of Speech Prosody, May 10–14 2010*, Chicago, IL.