The phonology of rhythm from Classical to Modern Portuguese

Sónia Frota*, Charlotte Galves+, Marina Vigário*,

Verónica Gonzalez-Lopez+ & Bernadete Abaurre+

*Universidade de Lisboa, +Universidade de Campinas

Sónia Frota

Laboratório de Fonética (CLUL)

Faculdade de Letras

Universidade de Lisboa

Alameda da Universidade

1600-214 Lisboa

Portugal

sonia.frota@mail.telepac.pt

Tel: +351-21-7920052 / +351-91-8245419

Short title: Rhythm from Classical to Modern Portuguese

Revised version, May, 2012

The phonology of rhythm from Classical to Modern Portuguese

Abstract

The prosodic change that has been reported to have occurred from Classical to Modern Portuguese is investigated by means of a new approach to the study of rhythm in language change. Assuming that rhythm is a by-product of the presence/absence of a set of properties in a given linguistic system, we computed frequency information on rhythm-related properties from written texts of the 16th to the 19th centuries, by means of the electronic tool FreP. Results show a change in the distributions of properties related to word stress and prosodic word shape after the 16th century, indicating that the prosodic change occurred between the 16th and 17th centuries. A predictive analysis based on Bayesian statistics provided strong support for the timing of the change, and successfully modelled our data showing a time line consistent with the direction of the prosodic shift towards the integration of stress-timing properties into Romance syllable-timed rhythm.

1. Introduction

Reading the phonology of a language from written text is a challenge, but also a need in order to go beyond the traditional sources that are usually confined to a few and sketchy comments by grammarians, the study of rhythm in poetry, or the description of the speech of literary characters. Historical phonology analyses of large sets of data require tools and methodologies which allow studies to be replicable and results to be independent of the circumstantial observer. The study of prosodic changes, which is even more challenging due to the subtle nature of the phenomena involved, is an area of research that would benefit from the broadening of the sources and data available, especially if some prosodic properties can be shown to be successfully computed from written text. In this paper, we will focus on the prosodic property of rhythm, and present a new approach to the study of rhythm in language change, using texts written in prose and computational tools together with a statistical model to quantify and validate the results obtained. Our main goal is to investigate whether rhythm-related properties can be computed from written text in any meaningful away, and thus contribute with new tools and methods for the study of prosodic changes over large empirical datasets.

Following a well-established line of thought in the framework of Generative Grammar (Lightfoot, 2006; Roberts, 2007, a.o.), we view language change as the effect of the interaction between External-Language (E-Language) and Internal-Language (I-Language). E-language can be defined as the totality of utterances produced in a speech community, whereas I-language is a mental object, the knowledge that every speaker develops as the result of language acquisition, his linguistic competence, or in other words, his grammar (Chomsky, 1985). In this picture, the competence attained by children of a given generation (their I-language) depends on the linguistic data they are exposed to (the E-language of their linguistic community), which in turn is produced by the grammar of the previous generations. It is therefore natural to expect that, from generation to generation, children of the same speech community acquire the same grammar. However, it is a well-known fact that grammatical changes occur in languages. The reason is that E-languages can undergo changes in such a way that the Corpus of utterances available to a generation G is different from the Corpus of utterances available to G-1, for reasons independent from the grammar of G-1. Changes can occur in the use of the language. According to Roberts (2007:129), phonetic and phonological changes can play a crucial role in changing the linguistic data available to the next generation.

To track phonological changes, and especially prosodic changes, in written texts is a challenge. This has been mainly performed on poetic texts, since poetry has an essential prosodic component. However, it is known that prosody is a property of language, independent of its spoken or written form (e.g., Chaffe, 1988; Fodor, 2002). We thus take the view that rhythm, as a prosodic feature, is a characteristic of any piece of language, and is expressed in oral texts as well as in texts written in prose. To establish whether a prosodic change has occurred in the corpus of utterances under observation, like in corpus linguistics (e.g. McEnery and Wilson, 2001) we see language change as that change which is identifiable and measurable within a given corpus of text. It can manifest itself in different ways, including a change in patterns of frequency of linguistic units or properties across time.

In the history of Portuguese, a prosodic change has been reported to have occurred between the beginning of the 17th and the end of the 18th century. There are many reasons for the study of the prosodic change in Portuguese using texts and the modern resources offered by computational methods and statistical models. Firstly, phonological changes are a very pervasive phenomenon in language and have been at the heart of the foundation of modern historical linguistics. Among such changes, the role played by stress patterns has been acknowledged since the work of 19th century Neo-grammarians. In the case of Portuguese, the phenomenon that most attracted the attention of grammarians was the raising of pre-tonic vowels, which was dated from the second half of the 18^{th} century in Teyssier (1980), and from one century earlier in Castro (2006) because it was already mentioned in one grammar written in 1682 and reflected in the writing of documents dated from the beginning of the 17^{th} century, such as those studied by Marquilhas (2003). However, references to this change are scarce and mostly limited to short comments made by earlier grammarians. The prosodic relevance of this phenomenon is described by Gonçalves Vianna (1892) in the following way: "A distância entre tónicas e as átonas, isto é, a diferença de intensidade entre elas, pode ser maior ou menor. Deste modo, a diferença entre as tónicas e as átonas das línguas germânicas é máxima; menor a que se dá em português; menor ainda a do castelhano; e mínima a francesa." [The distance between stressed and unstressed vowels, that is the difference of intensity between them, may be big or small. Thus, the difference between stressed and unstressed in Germanic languages is maximal; smaller is the difference in Portuguese; even smaller is the difference in Spanish; and minimal is the difference in French] (Gonçalves Vianna, 1892:16). The observations of this scholar suggest a relation between the changes in the vowel system and the changes in the realization of stress, that are reflected in a rhythmic change that places the rhythm of Modern European Portuguese between the rhythm of Germanic Languages and that of most other Romance Languages. In present-day accounts of linguistic rhythm, other properties besides stress-related vowel reduction patterns have been shown to play a role in rhythmic distinctions, namely the degree of variation in syllable types and word shapes (Dauer, 1983; Mehler & Nespor, 2002). As rhythm is a global long-term phenomenon, these properties need to be examined within large domains that allow the computation of the most recurrent and frequent patterns. In the present paper, we aim to study the prosodic change reported to have occurred in Portuguese between the beginning of the 17th and the end of the 18th centuries by computing several rhythm-related properties from written texts, including syllable, stress and word-based properties.

Secondly, it is known that prosodic factors are likely to have syntactic effects. This has been argued in the case of the loss of the Verb Second phenomenon in French, which was the result of a chain of changes initiated by the loss of the Germanic initial stress, creating the conditions for a reanalysis of Subject-Verb as no longer produced by a Verb Second grammar, but by a Subject-Verb-Object grammar (Adams, 1987; Dufresne, 1993). Another example is the decrease in the frequency of object topicalization in English, due to a trochaic rhythmic requirement that disfavors sentences with contiguous accented words, as is the case when a subject immediately follows a fronted topic (Speyer, 2008). In Portuguese, there was a change of clitic placement, as well as of the position of the subject in the sentence, that is observed in texts written by authors born from the beginning of the 18th century onwards: specifically, clitic location in the contexts that previously allowed for variation evolved to an enclitic pattern (Galves, Britto & Paixão de Sousa, 2005; see also Martins, 1994); and the position of the subject was established as pre-verbal, as seen by the strong decrease of the use of post-verbal subjects (Paixão de Sousa, 2004; Galves & Paixão de Sousa, 2010). In the absence of any external factor arguably responsible for such syntactic change, it is very tempting to think that it was triggered by the emergence of a new prosodic pattern, as Galves & Galves (1995) have proposed. However, if the syntactic change has a robust description supported by a large empirical basis and an established location in time, phonetic and phonological evidence for the prosodic change is still scarce and controversial, dependent on scant individual observations and thus extremely difficult to time. This situation makes claims about a possible connection between the two linguistic changes virtually untestable.

Thirdly, the Portuguese case can be seen as a testing ground for large-scale studies of prosodic changes in other languages, using written texts already made available for different purposes. The use of large-scale written corpora in historical linguistics is becoming increasingly common, and requires processing software and statistical measures (e.g., Hilbert and Gries, 2009). The present work follows this recent trend, extending the areas of change that can be observed to also cover prosodic aspects of language. The methodology proposed in this paper may be applied to other languages, besides Portuguese, provided that computational resources exist for obtaining frequency information of phonological units from written texts in those languages. Thus, the present study may contribute to the broadening of the sources and data available to the study of diachronic changes in prosody.

The novelty of our approach resides in the combination of the following theoretical and methodological aspects: (1) an approach to language change that assigns a crucial role to the corpus of utterances used by generations of speakers (Lightfoot, 2006; Roberts, 2007), and measures change within a given corpus; (2) the view that prosody is inherent to language, whether spoken, gestural or written (Wagner and Watson, 2010; Sandler, 1999; Chaffe, 1988; Fodor, 2002); (3) a theory of linguistic rhythm where rhythmic distinctions among languages are seen as resulting from sets of phonological and phonetic properties in a given linguistic system (Dasher and Bolinger, 1982; Dauer, 1983, 1987; Nespor, 1990, inter alia); (4) the use of computational methods to extract such properties from written texts; and (5) the use of Bayesian statistics for the validation of results. To investigate the prosodic change, we draw on a written texts database similar to that used for the study of the syntactic change (and drawn from the same corpus, the Tycho Brahe Historical Corpus), and compute quantitative frequency information on relevant rhythm-related properties by means of the electronic tool FreP (Frequency in Portuguese). Our findings clearly point to the presence of a prosodic change between the 16th and 17th centuries. A statistical analysis by means of ANOVA provides a first confirmation of this result. Finally, we develop a statistical predictive approach, based on Bayesian statistics, to model our frequency data. The predictive analysis successfully confirms that the prosodic change is in place by the 17th century, and nicely models the direction of the prosodic shift along the time line.

In the following section, we present our approach to rhythm in language, and characterize Modern European Portuguese rhythm. Section 3 of the paper is devoted to the methods and tools used. In section 4, the frequency data results are presented. The Bayesian predictive model is described in section 5. The paper concludes with a discussion of our major findings and their implications to studies of prosodic change in language.

2. On linguistic rhythm: The derived approach

The traditional view of linguistic rhythm classifies languages into one of three rhythmic groupings: syllable-timed languages, in which the timing regularity is based on syllables, like in most Romance languages; stress-timed languages, in which the timing regularity is based on interstress intervals, like in Germanic languages; and mora-timed languages, in which the timing regularity is based on morae, such as Japanese (Lloyd James, 1940; Pike, 1945; Abercrombie, 1967, inter alia). The traditional view of rhythmic classes thus hinges on the idea of isochrony, either of syllables, interstress intervals, or morae.

It is by now well-known that the physical basis of rhythmic distinctions as differences in the isochrony of the various linguistic units has been difficult to ascertain. Indeed, numerous phonetic studies have failed to confirm the hypothesis of isochronous units (e.g. Wenk & Wioland, 1982; Manrique & Signorini, 1983; Dauer 1983). This finding led to a different approach to linguistic rhythm, whereby rhythmic distinctions among languages are seen as resulting from the presence/absence of particular phonological and phonetic properties in a given linguistic system (Dasher and Bolinger, 1982; Dauer, 1983, 1987; Nespor, 1990). In this derived approach, rhythm is thus the result of the co-occurrence of a set of phonological and phonetic properties (p) in a language (L). An illustration of the derived approach is given in (1). (1a) and (1b) show the clear cases of stress-timed languages and syllable-timed languages, which have distinct sets of characteristic p properties. English and Italian could be instantiations of (1a) and (1b), respectively. In contrast, (1c) shows the less clear case where a language combines properties typical of both (1a) and (1b). (1c) is an important predicted outcome of the derived approach, as p properties are taken to be independent of one another, even if they usually cluster together in specific subsets. The literature provides us with several examples of the (1c)-type, like Catalan, Bulgarian, Polish, or Portuguese (Nespor, 1990; Dimitrova, 1997).

(1)	a. Lx	b. Ly	c. Lz
	{A, B, C}	{D, E, F}	{A, B, E}

Unlike in the traditional view, in the derived approach languages may either cluster into distinct rhythmic classes or scatter along a rhythmic continuum. The possible groupings into rhythmic classes are now crucially an empirical question. Evidence from both perception studies and language acquisition strongly suggest that the rhythmic class notion is empirically adequate. Perception studies have reported rhythm-based language discrimination, both by adults and infants, providing evidence that languages are perceptually grouped according to their rhythmic properties (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Nazzi, Bertoncini & Mehler, 1998; Ramus & Mehler, 1999). Research in language acquisition has also shown that languages of different rhythmic classes may use rhythm differentially as a cue to speech segmentation and lexical identification (Morgan, 1986, 1996; Mehler & Nespor 2002; Milotte, Morgan, Margules, Bernal, Dutat & Christophe, 2010; Mersad, Goyet & Nazzi, 2010).

A central notion to the derived approach is the description of the set of rhythm-related properties. Along the lines of Dauer's proposals, there are three main kinds of p properties: syllable structure variety and complexity, which is reflected in the syllable types a language exhibits; the presence/absence of a reduced vowel system; and the properties of stress, namely its importance in the phonology of the language, its distribution and correlates. In Mehler & Nespor (2002), mean word size is also considered as a possible relevant p property, given the fact that it seems to vary in correlation with rhythmic type: mora-timed languages tend to have frequent long words (like Japanese), whereas stress-timed languages are characterized by frequent monosyllabic words (like English), and syllable-timed languages fall in between (like Italian). Furthermore, the stronger or weaker role of word boundaries in the phonology of the language seems also to be correlated with rhythmic type: typically, stress-timed languages show a constellation of phenomena bound by the word, such as phonotactic restrictions, segmental rules and resyllabification; by contrast, in syllable-timed languages word-based phonological facts are less common and resyllabification spans a phrasal domain (Kleinhenz, 1997; Vigário, 2003). Thus, besides syllable structure, stress and word shape properties have also been claimed to co-vary with rhythmic distinctions. Concentrating on the stress-timing/syllable-timing distinction, which is the relevant one for the purposes of the present investigation, we list in (2) the types of rhythmic indicators that may reflect the three main kinds of p properties mentioned above. Table 1 presents the characteristic set of indicators respectively associated with stress-timed and syllable-timed rhythm. C and V stand for consonant and vowel, respectively. The properties relative to words are computed on prosodic words (PW), that is on the word as a phonological entity that bears one stress. PW1 indicates a PW consisting of just one syllable (a monosyllable).

(2) Proportion of CV syllables

Proportion and regularity of occurrence of C and V

Mean word size

Word shape variation

Distribution of stress (as a property of syllables and prosodic words)

PLEASE INSERT TABLE 1 HERE

Following the derived approach to rhythm, it is our goal to inspect the distributions of segments (consonants and vowels), syllables, prosodic word shapes and stress in written texts from Classical to Modern European Portuguese. The general hypothesis is that the presence of a rhythmic change will be manifested as a change in the distribution of some of these units. These properties, that have been used in the literature on rhythm since Dauer's proposals, do not have all the exact same phonological nature. At least some of them clearly imply generalizations over the lexicon, such as mean word size and word shape variation. Importantly, many scholars today share the view that both children's and adults' phonology is crucially grounded on generalizations over the lexicon (Bybee, 2001; Beckman & Edwards, 2000; Martin, 2007; Edwards, Munson and Beckman, 2011). The interaction between the frequency of use of lexical items and phonology can also be seen in sound change, as shown in the work of Bybee and colleagues (e.g., Bybee and Hooper 2001). Our use of the set of properties described above is thus in line with recent research on speech production and perception, as well as on language acquisition and change, by extending the role played by the lexicon (and lexical choice) to the analysis of rhythm.

To develop a specific hypothesis about the substance and direction of the rhythmic change under investigation, a description of present day European Portuguese rhythm is required. Modern European Portuguese (hereafter MP) rhythm is characterized by syllable-timed Romance-like properties and by stress-timed Germanic-like properties (Frota & Vigário, 2001; Vigário, Frota & Freitas, 2003). Like syllable-timed Romance, MP is CV-dominant, shows few syllable types and few closed syllables, and the prosodic word shapes exhibit less monosyllables than longer words (with more than two syllables) (cf. Vigário, Frota & Freitas, 2003; Vigário, Martins & Frota, 2005; Vigário, Freitas & Frota, 2006). Unlike syllable-timed Romance, MP also includes in its set of p properties the presence of vowel reduction – with a seven vowel-system in stressed position and a reduced four-vowel system in unstressed position –, a strong contrast between stressed and unstressed syllables, and stronger cues for the PW than in a typical Romance language (Vigário 2003). The analysis of physical correlates of p properties in MP, along the lines of Ramus, Nespor & Mehler (1999), supports its mixed rhythmic nature: in MP the duration of the space occupied by vowels is high, like in Romance Languages, but the variability in the duration of consonantal intervals is also high, similarly to Germanic languages (Frota & Vigário, 2001; Duarte et al, 2001). Importantly, the latter property reflects a low level phonetic feature, namely the optional but frequent deletion of unstressed high back vowels leading to acoustic consonantal intervals of varying length (Frota & Vigário, 2001). On the perception side, MP has been shown to be discriminated from Dutch, a stress-timed language (Frota, Vigário & Martins, 2002).

Given the rhythmic properties of most Romance languages, which are typical syllable-timed languages (e.g., Italian, Spanish, French), we assume that stress-timing is an innovation of MP. Thus, we entertain the hypothesis that the prosodic change that may have occurred in the language between the beginning of the 17th and the end of the 18th century consisted of the integration of stress-timing properties into syllable-timed rhythm.

PLEASE INSERT TABLE 2 HERE

This hypothesis leads to the prediction that the p properties of the stress-timing profile (given in Table 1 above) that were integrated into syllable-timing were those related to stress and the prosodic word, as shown in Table 2. Crucial to our prediction is the fact that like in syllable-timed Romance languages, phonological syllable structure in MP is dominated by the CV type, shows few syllable types and few closed syllables (Vigário, Frota & Freitas, 2003; Vigário, Martins & Frota, 2005). Due to the low level phonetic nature of optional vowel deletion, we do not expect to find indicators of this property retrievable from a phonological coding of written texts. We thus expect that the rhythmic change will be signalled by a change in the distribution of prosodic word shapes and stress patterns measurable within a corpus of written texts.

3. Methods

3.1. Materials: The Tycho Brahe Historical corpus

The Tycho Brahe corpus is an annotated historical corpus freely accessible at http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html (Galves & Faria, 2010). It includes 53 texts written in Portuguese by authors born between 1380 and 1845, comprising a total of 2,464,191 orthographic words. The Tycho Brahe corpus uses the chronological criterion of the author's birthdate. Text dating is thus based on the year of birth of the authors rather than on the moment the text was produced. This choice is motivated in great part by the Generative Linguistics framework. Either from the point of view of grammar or from the point of view of use, language is a property of individuals. Moreover, given the strong relationship between language change and language acquisition in this framework, we rely on generations of speakers rather than generations of texts. It must be emphasized, however, that this methodological choice is permitted by the fact that our corpus is a corpus of authors whose dates of birth are a well-known fact.

Given that the prosodic change under investigation was previously reported to have occurred between the beginning of the 17th century and the end of the 18th century, we selected a subset of the Tycho Brahe corpus formed by texts by different authors born between 1502 and 1836 in order to examine this change. These texts, all in prose, are representative of several genres – sermons, theatre plays, historical narratives, memories, philosophical dissertations –, which can be located on a continuum with respect to their relationship with orality. We expect some variation to be produced by the different genres, as well as by the different authors, but our hypothesis is that this variation is independent from the diachronic change, which is the object of this paper.

Furthermore, the use of computational tools usually requires some degree of normalisation of the texts analysed (e.g. Heidelberg & Marquilhas, 2012), and thus we selected the texts that were edited following modern European Portuguese orthography (according to the 1945/1973 norm), since this is the orthographic norm read by the FreP computational tool (see section 3.2).

The final subset of texts consists of 17 historical texts written by 15 different authors, and distributed by 5 genres – narrative (narr), letter (lett), philosophical (phil), sermon (serm), and theatre (theat) –, as shown in Table 3. The 17 historical texts make up a total of 564,554 orthographic words. In the subset under observation, there are 4 texts by authors born in the 16th century, 4 texts by authors born in the 17th century, 4 texts by authors born in the 18th century, and 5 texts by authors born in the 19th century.[1]

PLEASE INSERT TABLE 3 HERE

In addition to the historical texts, written texts of mixed types, dated from 1990 to 2010 were also included in our materials. This corpus, with a total of 133,379 orthographic words, is a subset of the corpora collected for the FrePOP (Frequency of Phonological Objects in Portuguese Database), available at http://frepop.fl.ul.pt/ (Frota, Vigário, Martins & Cruz, 2010). The results from the 1990-2010 materials will provide us with a reference point for the present status of computed rhythmic-related properties in MP.

3.2. The FreP tool

The electronic tool FreP - Frequency in Portuguese (Martins, Vigário & Frota, 2009, http://www.fl.ulpt/ LaboratorioFonetica/frep) extracts the frequency of occurrence of phonological units at the word level and below from Portuguese written text. Specifically, FreP identifies and counts major classes of segments (consonants, vowels, glides), syllables and syllable types, phonological clitics, clitic type and size, prosodic words and their shape, word stress location, among other units and phonological patterns (Frota, Vigário & Martins, 2006). Reliability measures for the different functionalities of the tool are above 99.5% (http://www.fl.ul.pt/LaboratorioFonetica/FreP/Year2_report.pdf). FreP thus generates a phonological transcription from written text, by extracting from written language some critical properties of spoken language. The tool was used to provide a phonological coding of our historical corpus and compute the frequency of the several phonological units and patterns. On the basis of this coding and of the frequency computation, we were able to get the relevant p properties for the study of the prosodic change from the written materials which constitute our empirical basis.

A detailed description of the FreP tool is beyond the scope of this paper. We will thus concentrate on the aspects of the tool's input and output that are crucial to our present purposes. The input of FreP is a simple orthographic form (according to the 1945/1973 norm). As shown in Table 4, first the written text is divided into orthographic words (first column); in various intermediate steps the orthographic words are phonologically analysed (e.g., application of letter to sound rules, stress assignment, syllabification – see the second column); finally, an output is provided where orthographic words are transformed into a phonological template showing the segmental composition in major sound classes, the syllabic composition, the presence/absence of stress (and thus the prosodic word or clitic status), and the location of stress (third column). This output automatically generated by FreP was the basis of the data for the analysis of the behaviour of rhythmic-related properties between the 16th and the 19th centuries: namely, the distributions of consonants and vowels, of syllable types, of word stress and word shapes. It is important to note that all the properties under analysis were derived from FreP's phonological template, which has been shown to be reliably generated from written text. Low-level phonetic aspects, such as the actual pronunciation of a given segment and not its phonological status as a consonant or vowel, were not considered in the analysis as we do not expect to find reliable indicators of this property retrievable from a phonological coding of written text. After the FreP phonological coding of written texts, our empirical basis consisted of 311,718 prosodic words (PW), 928,908 syllables, and 1,977,115 segments.

PLEASE INSERT TABLE 4 HERE

The FreP tool assumes Modern Portuguese lexical phonology for the building of the phonological coding from written text. By using FreP with historical texts, we are thus assuming MP lexical phonology under the hypothesis that it did not change with time. This claim is strongly supported by the fact that the crucial aspects of MP lexical phonology at stake in our work are basically Romance-like (Mateus & Andrade, 2000; Frota & Vigário, 2001; Vigário, 2003): lexical Vs and Cs have kept their status as such; syllable types and their frequencies are comparable to other Romance Languages; stress location bears on morphological information, like in Romance Languages. For lexical Vs and Cs, notice that phonetic changes in vowel or consonantal quality have no impact on the prosodic properties we are measuring. Phonetic changes that would impact on syllable types (such as vowel loss or merger, intervocalic consonantal loss or insertion, and diphthongization) were mostly established by 1500 (Teyssier, 1980: 40-66). Finally, word stress has remained essentially in the same location since Vulgar Latin, and phonetic processes that may have altered the stress pattern of particular words (such as vowel loss or merger) were reported to have occurred in an earlier period, before the 16th century (Williams, 1975; Quednau 2000).

In short, FreP offers a methodology that allows us to compute rhythmic-related properties from written text over large sets of data. Rhythm, as prosody in general, is inherent to language, whether spoken, gestural or written (Wagner and Watson, 2010; Sandler, 1999; Chaffe, 1988; Fodor, 2002). Research on language processing, both spoken and written, has shown that prosody plays an active role in how we speak, read and write (Shattuck-Hufnagel & Turk, 1996; Fodor, 2002; Chaffe, 1988). By using FreP, we extend the analysis of rhythm to include written texts from large historical corpora, broadening the sources and data available for the study of prosodic changes.

The methodology proposed in this paper for the study of prosodic change may be applied to other languages, provided that resources exist for computing frequency information of phonological units from written texts in those languages. The basic architecture of FreP (its modular organization and its set of functions – Frota, Vigário & Martins, 2006) can be used to build equivalent tools for other languages. Given that other Romance languages, besides Portuguese, also show a fairly predictable relation between orthography and (lexical) phonology (Serrano, Genard, Sucena, Defior, Alegria, Mousty, Leybaert, Castro & Seymour, 2010), and that the lexical phonologies of MP and of other Romance languages show many similarities, the FreP developers are currently engaged in the construction of comparable tools for other Romance languages, namely Catalan and Spanish.

In the specific case under investigation in this paper, a prosodic change is predicted to have occurred by means of the integration of stress-timing properties into syllable-timed rhythm. This prosodic change would be manifested by a change in the frequency distribution of one or more p properties related to prosodic word shapes and word stress patterns. Such a change is expected to occur towards the present status of rhythmic-related properties in Modern European Portuguese.

3.3. Data analysis

To test our central hypothesis that the prosodic change that has occurred from Classical to Modern Portuguese consisted in the integration of stress-timing properties into syllable-timed rhythm, we inspected the frequency distributions of the phonological units and patterns involved in the stress- and syllable-timing profiles described in Table 2 above. We are assessing the emergence of the Modern Portuguese mixed rhythmic nature (that is, a change from a (1b)-like language into a (1c)-like language), and not a change from syllable-timed into stress-timed rhythm (that is, a change from a (1b)-like language into a (1a)-like one). For that reason, we applied the following rationale in the data analysis: we take the most frequent syllable type, word shape, or stress pattern, which are expected to be kept the same since the 16th century to the present day (as properties related to syllable-timing), and inspect the variation in the proportion of the other patterns relative to these (the other patterns are properties related to stress-timing). Therefore, p properties are expressed as ratios and each data point corresponds to the value of a given p property for a given text. For example, prosodic word shape variation is computed as the variation in the proportion of monosyllables relative to the most frequent word shape, the disyllable.

As our materials consist of texts written by different authors and of various types, we may expect an impact of these two factors on our data.[2] However, by grouping the data into successive periods of time of the same size we can inspect whether variation can be found across authors and genres as a function of time. This can be achieved by clustering the data by century (using the author's date of birth). Grouping is the usual technique of data clustering for analysis, and the chosen time interval is among the most commonly used in historical corpus linguistics (Hilpert & Gries, 2008). Taking the century as the time window for analysis allows for the computation of mean values that accommodate the variability due to factors such as author and genre. Another consequence of using this time window is the boosting of sample size and thus of the statistical power of the analysis: the grouping by century leads to a sample size approximately four times bigger than each text individually considered.

In section 4 the results obtained are first presented descriptively, for each p property examined. In order to test for significant differences between the behaviour of rhythmic indicators over time, an analysis of variance was calculated. The ANOVA took the normalised values for the rhythmic indicators for each text (the value computed by FreP divided by the overall median value of each indicator) as the raw data and time (defined as the century of birth of the text author) as a factor. Given a significant effect of time, pairwise comparisons of the different centuries were run to inspect which groups were significantly different (Post Hoc Bonferroni).

Finally we develop a predictive approach (based on Bayesian statistics) to model our frequency data (section 5). The predictive analysis based on the data from each of the centuries by rhythmic indicator shows whether each century is different or not from the others, locates the data points in the time line, and allows a selection of the best rhythmic indicators for the prosodic change.

4. Results

In section 2 above, a specific hypothesis about the substance and direction of the prosodic change under investigation was developed: the prosodic change consisted of the integration of stress-timing properties into syllable-timed rhythm. Stress-timing would thus be an innovation of MP, resulting in the mixed rhythm pattern of the language that combines both syllable-timing Romance-like properties and stress-timing properties. According to this hypothesis, the integration of stress-timing properties into syllable-timed rhythm should be seen in p properties related to prosodic word shapes and stress patterns, whereas properties related to segments and syllables should show a constant pattern reflecting the syllable-timed basis of Portuguese (Table 2). We start by examining the latter p properties.

PLEASE INSERT TABLE 2 HERE

The ratio between simple and complex syllable types was computed: simple syllables are those with the CV or V formats; complex syllables are those with a complex onset and/or a complex rhyme. Figure 1 shows the ratio between simple and complex syllables from 1502 to present. Despite the higher variation found at the beginning, no clear signs of change appear pointing to a constant pattern. The ratio between the two major classes of segments – consonants and vowels – was also computed. In Figure 2, the consonant/vowel ratio is plotted by century. There is important variation among texts within the same century, and the general trend shows no signs of a change between the 16th century and the present day. These results are in line with our prediction that properties related to segments and syllables should show a constant pattern reflecting the syllable-timed basis of Portuguese.

PLEASE INSERT FIGURE 1 HERE

PLEASE INSERT FIGURE 2 HERE

Let us now turn to the p properties that were predicted to capture the prosodic change. To assess prosodic word (PW) shape indicators, we need to first examine the behaviour of the dominant PW shape in the language, the disyllable. The proportion of disyllables is quite stable over time, as depicted in Figure 3. The median values from the 16th to the 19th centuries are, respectively, 45.69, 43.83, 43.39 and 44.32. The next step is to inspect the variation in the proportion of the shorter word shape, the monosyllable, relative to the disyllable. An increase in the ratio between monosyllables and disyllables would be related to the emergence of stress-timing properties (section 2, Table 2). In the data, the proportion of monosyllables (PW1) relative to the disyllables (PW2) has a clear evolution pattern, with the 16th century deviating from all the others (Figure 4): the median values from the 16th to the

19th centuries are, respectively, 0.328, 0.421, 0.401 and 0.433. This increase in the proportion of monosyllables, which are short PWs, between the 16th and the 17th centuries and towards the present day values, is in line with the prediction that properties related to the PW are relevant to signal the hypothesized integration of stress-timing properties into the rhythm of Portuguese.

PLEASE INSERT FIGURE 3 HERE

PLEASE INSERT FIGURE 4 HERE

However, it should be noted that the 19th century data shows a conservative outlier (the narrative text written by Fronteira, born in 1802). This same conservative outlier will appear in all indicators related to PW shape, but not in the indicators related to word stress. Interestingly, the 16th century shows a data point that clearly lies in the interval of the 17th century data: it is the narrative text written by Brandão, born in 1584 and thus an author at the transition between centuries. This same text by Brandão behaves like its 16th century pairs if monomoraic PW shapes instead of monosyllables are considered, as we will see below. These two examples provide a first indication that p properties are independent from one another, as predicted by the derived approach to linguistic rhythm (section 2).

PLEASE INSERT FIGURE 5 HERE

The shortest words possible in a language are monosyllables that end with a vowel, that is that show a non-complex rhyme. These prosodic word shapes are known as monomoraic words or subminimal words (e.g., Demuth, 2006).[3] We examined the variation in the proportion of monomoraic words relative to the disyllable. Again, an increase in the ratio between monomoraic words and disyllables would be related to the emergence of stress-timing properties, by signalling an increase in the presence of shorter PW shapes. The proportion of monomoraic PW (PWm) relative to the disyllables shows a pattern similar to the PW1/PW2 ratio, with a change between the 16^{th} and the 17^{th} century (Figure 5). This result strengthens the trend towards the more frequent use of short words from the 17^{th} century onwards. It is revealing that in the case of monomoraic shapes there is no outlier in the 16^{th} century data: as noted above, the text by Brandão shows a post-16^{th} century behaviour with respect to monosyllables, but a 16^{th} century behaviour with regard to monomoraic words. This suggests that the shift towards a more frequent use of short words targeted monosyllables first and then monomoraic shapes, that is the shift started with the longer, monosyllabic forms and then extended into the shorter, monomoraic ones.

Besides prosodic word shape indicators, a second set of p properties was predicted to capture the prosodic change, namely those related to word stress. The predominant stress pattern in Modern Portuguese is penultimate stress. Final stress is the other common stress pattern, whereas antepenultimate stress is very rare (Mateus & Andrade, 2000: chap. 6; Frota, Vigário & Martins, 2006). Following the rationale underlying the data analysis, we inspected the variation in the proportion of final stress relative to penult stress in multisyllabic words. The results are shown in Figure 6. Again, the 16^{th} century exhibits a different behaviour from all the others with a specific distribution of stress patterns, with a median value clearly below 0.25 (the median values for the 16^{th} to the 19^{th} are, respectively, 0.208, 0.252, 0.272, and 0.285). This change in the relative use of stress patterns between the 16^{th} and the 17^{th} centuries, and towards the present day value, is in line with the prediction that properties related to stress are relevant to signal the hypothesized prosodic change. Importantly, both the PW shape and stress indicators point to the same timing for the change.

PLEASE INSERT FIGURE 6 HERE

The computation of stress patterns shown in Figure 6 was made on the basis of multisyllabic PW, and thus naturally excludes monosyllables. However, monosyllabic PW's also bear stress. In Vigário, Frota & Martins (2010) it is argued that monosyllabic words may add to the final stress pattern, as the syllables of monosyllables pattern like word-final syllables and proclitization may build disyllabic shapes with final stress. In other words, stressed monosyllables show the features of words with final stress. Along this line, we also examined the proportion of final stress plus monosyllables relative to penultimate stress (Figure 7). The results confirm the presence of a shift between the 16th and 17th centuries (the median values for the 4 centuries are, respectively, 0.437, 0.54, 0.547, and 0.579), while showing a picture very similar to that of the ratio between monosyllables and disyllables (see Figure 4 above).

PLEASE INSERT FIGURE 7 HERE

All the four potential rhythmic indicators converge in setting the 16th century apart, and grouping the 17th to 19th centuries together and close to the reference value for present day MP. This descriptive observation is strengthened by a joint analysis of all indicators. For this analysis, normalised values were obtained by dividing each data point for a given property by the respective median value for that property: for example, the normalised value for the ratio between monosyllables and disyllables for a given text was divided by the median value of this indicator computed over all the texts examined in this study. Besides the four potential rhythmic indicators, a similar computation was obtained for the ratio between consonants (C) and vowels (V), that had shown no change between the 16th and the 19th centuries. In this way, we are able to compare the evolution p properties that were predicted to capture the prosodic change with the evolution of a p property predicted to show a constant pattern. Results by historical text are given in Figure 8, plotted against the C/V ratio (full line), that is against a non-changing property. Three facts are worth noting: (1) the prosodic word shape and stress properties cluster below the full line before 1600 and show a different behaviour after 1600, confirming previous observations; (2) the data point 1802 (the narrative text by Fronteira) is a clear exception in all the 17th to 19th century period, for its sparse use of short words; (3) in line with the derived approach to rhythm, p properties do not necessarily change together or at the same time, as shown, for example, by the case of Fronteira where the PW shape indicators are conservative but the stress indicator is within the pace of its time.

PLEASE INSERT FIGURE 8 HERE

PLEASE INSERT FIGURE 9 HERE

A joint analysis of all potential rhythmic indicators was also performed using the grouping of the data by century, and including the data from the 20th century as a reference point. As shown in Figure 9, the divide between the 16th century and all the others emerges clearly, as well as the proximity of the 17th to 19th centuries' data to the 20th century.

The results reported so far point to the presence of a prosodic change at the beginning of the 17th century. This descriptive observation was tested by looking for significant differences between the behaviour of rhythmic indicators over time. This was done by means of a one-way ANOVA. The normalised values for the rhythmic indicators were taken as the raw data and time (defined as the century of birth of the text author) as a factor. The ANOVA showed a significant main effect of time ($F(3, 81) = 6{,}34$, $p = .001$). Post Hoc tests (Bonferroni) confirmed that only the 16th century was significantly different from the others (16^{th} / 17^{th} : $p < .05$; 16^{th} /18^{th} , 19^{th} : $p < .01$). We would like to observe, however, that there exist several restrictions for using ANOVA in our case, namely the impossibility of checking the normality of the sample distribution, and the evidence for heteroscedasticity.[4] In section 5, we propose a new statistical approach that could be useful to deal with our kind of data.

In summary, our findings clearly show the presence of a shift in rhythmic indicators supporting a prosodic change between the 16^{th} and 17^{th} centuries. The direction of the shift towards present day values of the same rhythmic indicators confirms the hypothesis that the change consisted in the integration of stress-timing properties into a Romance-like syllable-timed rhythm, contributing to the mixed rhythm of Modern Portuguese. The success of the specific p properties analysed as potential indicators of the change, namely prosodic word shape and word stress related properties, further confirms the relevance of these properties to the characterization of linguistic rhythm, especially in languages where different rhythm types are combined. However, two important questions remain to be answered. First, are the findings robust enough to allow a predictive analysis to be made on the basis of the rhythmic indicators? Specifically, given the evidence found what is the probability that the 16^{th} century data successfully predicts each of the other centuries data? If the 16^{th} century data is not able to predict the following centuries data, then we would have strong support for the difference between the 16^{th} century and the following centuries, and thus for the prosodic change at hand. Secondly, if our analysis is on the right track, we should be able to model our data showing a time line consistent with the direction of the prosodic shift towards the integration of stress-timing properties. Along this time line, some rhythmic variables may be better indicators of the prosodic change than others. We thus seek to offer a new approach that could be able to handle our kind of data, showing not only possible differences between the centuries, but also pointing to the direction of such differences. These questions are addressed in the following section by means of a predictive approach based on Bayesian statistics.

5. Modelling the prosodic change: a predictive approach

The use of large-scale written corpora in historical linguistics, which is becoming increasingly common, requires the application of statistical techniques (e.g., Hilbert and Gries, 2009). Such techniques may be crucial to reveal phenomena in the data that are not observable by eye, to confirm the robustness and significance of the findings, and to model changes in diachronic data. In other words, and although they involve a certain degree of technicality, statistical methods are useful for assessing language change. In this paper, we develop a statistical predictive approach, based on Bayesian statistics, to model our frequency data.

The usage of the Bayesian predictive is quite common in statistics. The concept has been explored in life testing models (e.g., Dunsmore, 1974; Hussaini, 2011) and applied in economic sciences and health care, for example (see, respectively, Gneiting & Raftery, 2007; Aven & Eidesen, 2007). In linguistics, a Bayesian approach similar to the one proposed in this section has been explored, for example, in speech processing and language comprehension studies (e.g., Johnson & Goldwater, 2009; Fine, Qian, Jaeger & Jacobs, 2010). Recently, the large family of Bayesian models has found applications in many areas within linguistics, including language learning and cognitive development (e.g., Perfors, Tenenbaum, Griffiths & Xu, 2010). Our goal is to assess the prosodic change by evaluating the difference and/or proximity shown by the data from each century under analysis. Lack of proximity or difference between the data from any two centuries can be quantified and when the difference is statistically significant, it is interpreted as change.

A statistical predictive analysis was developed based on the dataset from each of the centuries by rhythmic indicator (the information collected at time t), going forward or backward to another century (other information collected at time t'). This analysis, following a Bayesian approach (Gupta & Nadarajah, 2004; Robert, 2007), shows whether the data from each century successfully predicts the others, locates the data points in the time line, and allows a selection of the best rhythmic indicators for the prosodic change. The data from a given century successfully predicts the others if the properties it shows are similar enough to the properties of others, as measured by a rule of proximity described below. The same assessment of degree of proximity/difference enables the location of the data points in a time line defined by each of the centuries, where the other centuries are closer or further apart and appear in the conservative or the progressive region with respect to the prosodic change. In the following paragraphs, we describe the mathematical reasoning behind the predictive analysis.

Our predictive distribution was constructed for each temporal point t (or century). We thus obtained four predictive curves (16th century, 17th century, 18th century and 19th century). Each predictive is based on the number of successes, denoted by "y", observed at time t in n independent realizations of the experiment. If we know the number of realizations n' at time t', the predictive denoted here by Pred(Y'/ t,n,y,t',n') checks if the number of successes y' observed at time t' could be considered as a regular observation of the experiment related to time t. In our case, the number of successes will depend on the output of each of the four rhythmic indicators related to prosodic word shape and word stress.

The definition of Pred(Y'/ t,n,y,t',n') is based on information from two fonts: the data observation and the conjecture. The elements that compose the predictive are the following:

(1) The likelihood (named Binomial distribution) or the discrete distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p, denoted by Bin(n,p); (2) The prior distribution on p, that is the conjecture about p (proportion of success that we expect), which is a Beta distribution denoted by Beta(α,β); (3) The innovative information about the time t', which is the number of realizations n' and number of successes y'. The predictive Pred(Y'/ t,n,y,t',n') is an average between the likelihood and the prior distribution, introduced by (1) and (2). This predictive is thus a Beta-Binomial distribution and rigorously speaking it depends also on α and β (hyperparameters of the prior distribution), with which we can formulate several conjectures. In our approach, we assume the non-informative conjecture (α=β=1) according to which all the values of p have the same chance to occur, that is the probability of success is equal to the probability of failure. Finally, we characterize an observation y' of the random variable Y' as being predictable from Y, if the observed value y' is located within the credibility region defined by the mathematical formulation of the predictive.[5]

The variables in the predictive analysis were the four rhythmic indicators related to prosodic word shape and word stress described in the previous sections (PW1/PW2, PWm/PW2, Fin/Penult and Fin+PW1/Penult). We defined the event of interest as being the occurrence of the property described in the "Success" column in Table 5. We defined as failure the occurrence of the complementary property, described in the "Failure" column in Table 5. Thus, "Success" entails a result in line with the prosodic change, and "Failure" a result not in line with the change. The data, consisting of the outputs of the FreP tool as described in section 3, was collected by century. This was done through the use of weighted means (that is, each text contributes to the characterization of time t proportionally to its size), and provides the required big sample size for the analysis. In order to assure the independence between the trials assumed by the Binomial likelihood, we selected one word out of 5 for analysis, that is, from the prosodic words 1, 2,…, 12, we selected 1, 6, 11, and so on.

PLEASE INSERT TABLE 5 HERE

The results obtained are shown in Figures 10 and 11. In each graph, the horizontal lines represent the centuries. For each century, the predictions are displayed relative to the other centuries (e.g., the data points in the line for the 16th century show predictive values for the 17th, 18th and 19th centuries based on the data from the 16th century). The two vertical lines delimit the predictive region, that is the region within which the data from one century is able to predict the other data. If, in the horizontal line corresponding to century Y, the point corresponding to century X is located in this region, this means that Y predicts X (meaning that the properties of X are similar enough to the properties of Y). Note that the region to the right of the horizontal lines corresponds to observations with more "Success" than inside the interval, and the region to the left of the horizontal lines corresponds to observations with less "Success" than inside the interval. In other words, the region to the right is the progressive region in terms of the rhythmic change, and thus the data points of the 16th century are expected to be located to the left whereas the data points of the remaining centuries are expected to be placed to the right in the time line relative to those of the 16th century.

PLEASE INSERT FIGURE 10 HERE

PLEASE INSERT FIGURE 11 HERE

The results of the predictive analysis show that the data from the 16th century does not predict that of any of the following centuries. This is depicted by the horizontal lines representing the 16th century that, for all four rhythmic indicators, do not show data points within the predictive region. By contrast, the lines representing the other centuries all show data points within (or very close) to the predictive regions. In the lines of the 17th, 18th and 19th centuries, the only data point that is consistently farther away from the predictive regions is that of the 16th century. These results thus show that the properties of the 16th century are different enough from the properties of the other centuries, providing strong support for the presence of a prosodic change between the 16th and 17th centuries.

The inspection of the regions to the right and left of the predictive region shows that in the 16th century line, the data points of the other centuries are located to the right, that is in the progressive region. Conversely, the location of the 16th century in the lines of the other centuries is always in the region to the left of the predictive region, that is the conservative region. This result clearly confirms that the direction of the prosodic shift was towards the integration of stress-timing properties, manifested as a change in the distribution of prosodic word shapes and word stress patterns towards smaller PWs and more frequent final stress.

The relative position of the data points in the horizontal lines, besides the 16th century / other centuries divide, is consistent with the observations made in section 4. The prosodic word shape rhythmic indicators were more sensitive to authors/texts with specific behaviour, as shown by the relative position of the 19th century located before the 17th or 18th centuries due to the conservative behaviour of the narrative text written by Fronteira. By contrast, the word stress-related rhythmic indicator Fin/Penult nicely predicts not only the change between the 16^{th} and 17^{th} centuries, but also a time line where the relative location of the data points is more stable.

In summary, the predictive approach provided strong support for the prosodic change between the 16^{th} and 17^{th} centuries, and successfully modelled our data showing a time line consistent with the direction of the prosodic shift.

6. Final discussion and conclusion

In this paper, we presented a new approach to the study of rhythm in language change, which goes beyond the traditional sources usually confined to a few insightful comments by earlier grammarians, the study of rhythm in poetry, or the description of the speech of literary characters. Under the assumption that rhythm is a by-product of the presence/absence of a set of properties in a given linguistic system, we computed phonological properties from written texts of the 16^{th} to 19^{th} centuries to investigate both the presence and time of a prosodic change previously mentioned in the literature on the history of Portuguese. Given the mixed rhythmic properties of Modern European Portuguese in comparison to most Romance languages, which are typical syllable-timed languages (e.g., Italian, Spanish, French), our hypothesis was that the prosodic change consisted of the integration of stress-timing properties into Romance-like syllable-timed rhythm. Under this view, the inclusion of stress-timing properties was an innovation of MP. In keeping with this hypothesis, the prosodic change would be captured by the distributions of properties related to word stress and the prosodic word in the stress-timing profile, whereas the distributions of segments (consonants and vowels) and syllables in the texts should show a constant pattern reflecting the syllable-timed basis of Portuguese. Specifically, we predicted a change in prosodic word shapes towards the more frequent use of shorter words and a change in the role played by word boundaries in the phonology of the language towards a stronger role of word boundaries. Both of these properties have been shown to characterize stress-timed rhythm (Kleinhenz, 1997; Mehler & Nespor, 2002).

Using a large dataset taken from the Tycho Brahe Historical Corpus, we computed quantitative frequency information on the relevant rhythm-related properties by means of the electronic tool FreP (Frequency in Portuguese), which provided a phonological coding of Portuguese written texts. As predicted, our findings showed that the syllable types profile is Romance-like throughout the period observed and that the consonant/vowel distribution does not change. By contrast, prosodic word shape and word stress measures showed a change in the expected direction: towards a more frequent use of shorter words and more frequent word final stress. The descriptive observation of the data pointed to the presence of a prosodic change between the 16th and 17th centuries. The ANOVA results provided a first confirmation of this finding, by showing a significant main effect of time over the behaviour of the rhythmic indicators, an effect that was due to the significant difference between the 16th century and all the others (although the use of ANOVA with the dataset analysed could be contested, due to technical aspects). Finally, a predictive approach based on Bayesian statistics not only provided strong support for the prosodic change between the 16th and 17th centuries, but also successfully modelled our data showing a time line consistent with the direction of the prosodic shift: (i) the 16th century data was unable to predict each of the other centuries data (and conversely, each of the other centuries was unable to predict the 16th century), showing that the 16th century was indeed different from the following centuries; (ii) in the 16th century predictive line, the data points representing the other centuries were located in the progressive region, whereas in the other centuries lines the 16th century was always located in the conservative region. It is worth noting that this location in time of the prosodic change shows that it preceded the syntactic change reported to have occurred at the beginning of the 18th century (see section 1). This finding is thus compatible with hypotheses that the prosodic change may have triggered the syntactic change, along the lines of similar proposals relating prosody and syntactic changes in other languages (see Adams, 1987 for French).

The present investigation provided new empirical support for the derived approach to language rhythm from the perspective of language change. Our findings support the relevance of the set of properties proposed by Dauer (1983, 1987), Nespor (1990), or Nespor & Mehler (2002), for an account of language rhythm. Further, the Portuguese historical findings show that the set of phonological and phonetic (p) properties that co-occurs in a given language may evolve over time, and that p properties may change independently of one another (even if they usually cluster together in specific subsets), as it is predicted by the derived approach. It is revealing that our historical findings are in line with recent descriptions of Modern European Portuguese as a language whose prosody deviates from the common Romance pattern: MP shows stronger cues to the prosodic word than most Romance languages, has a strong contrast between stress and unstressed syllables and a preference for using shorter words, all properties that are common in stress-timed languages (Vigário, 2003; Vigário, Freitas & Frota, 2006). The prosodic word and word stress-related changes reported in our findings may thus be interpreted as signalling the shift towards these aspects of present day MP prosody. However, if the change in prosodic word shapes towards the more frequent use of shorter words offers a straightforward interpretation, the connection between the word stress change towards a more frequent use of final stress and the above mentioned stress-timing properties is less transparent. We interpret the change in word stress patterns as signalling a stronger role played by the word boundary in the language. Stress-timed languages exhibit strong cues to the delimitation of prosodic words, unlike syllable-timed languages (Kleinhenz, 1997; Vigário, 2003), and word stress has long been recognized as one such cue (Cutler & Norris, 1988). The change found in MP towards the more frequent use of final word stress favors the alignment between a word boundary and a prominent phonological feature, thus strengthening the cues available for the delimitation of prosodic words.

The connection between our findings and the scant descriptions of the prosodic change made by grammarians and scholars may be seen as follows. What grammarians and scholars have mentioned were mostly changes in the vowel system related with changes in the realization of stress, namely in the contrast between stressed and unstressed. Such changes, as suggested by Gonçalves Vianna (1892), have placed the rhythmic pattern of Modern Portuguese between that of Germanic Languages and that of Romance Languages. By computing different kinds of rhythm-related properties retrievable from the phonological coding of written texts (i.e., the distribution of major classes of segments, syllable types, word shapes and word stress patterns), we found a change in the rhythm-related word shape and word stress properties that supports the integration of stress-timing properties into MP Romance-like syllable-timed rhythm. In other words, our findings also place the rhythmic pattern of Modern Portuguese between that of Romance and Germanic Languages. While it seems clear that both the vowel quality changes mentioned by grammarians and earlier scholars and the changes in word shape and word stress patterns we found are part of the of the mixed rhythmic nature of MP, a closer inspection of the relation between these properties requires further investigation and is left for future research.

On the methodological side, the present research has shown that rhythm-related properties can be computed from texts written in prose in a meaningful away, by using computational methods and statistical models (namely the FreP tool and the predictive approach based on Bayesian statistics) applied to the study of prosodic changes over large empirical datasets. The methodology proposed in this paper holds the potential of being applied to other languages, especially in the case of other Romance languages for which similar computational tools are being developed. It is thus hoped that the present study may contribute to the broadening of the sources and data available to the study of prosodic changes, as well as to the methods used for the statistical validation of results in this area of research.

[1]. Garrett, who was born in 1799, was classified as an author of the 19th century.

[2]. A study of the effects of author or text type would require a set of materials different from that presently available: specifically, a set of texts of different genres by the same author, and a set of texts of the same genre by different authors. We leave this study by author and by text type for future research.

[3]. Monomoraic prosodic words exhibit the shortest format possible. It is known that this word shape is not allowed in many languages. Notice that this fact is a consequence of restrictions that are independent of the rhythmic group a language belongs to, as minimal word constraints are active in languages like English (stress-timed), Catalan (syllable-time, or with mixed properties between stress- and syllable-timed), and Japanese (mora-timed). The analysis of the evolution of the proportion of monomoraic words in comparison with the evolution of the proportion of monosyllables in general allows us to capture the relative contribution of the shortest format to the proportion of PW1.

[4]. The main objections that we point out are (a) the impossibility of checking the normality of the sample distribution, and (b) evidence of heteroscedasticity. It is known that if any of these assumptions is violated, the ANOVA results may be seriously biased or misleading. For (a) with relatively small data sets (that is our case with one value per rhythmic indicator per text), there isn't a good way to check normality from the data, which means that we need to assume normality. But we are working with ratio of proportions (like PW1/PW2=(PW1/(PW1+PW2))/(PW2/(PW1+PW2))) and proportions are approximately normal, which makes us suspect that normality fails (the ratio of standard normal independent variables is a Cauchy variable). For (b), the evidence is given by the difference between the intervals of variation inside each century. For illustration, compare the 17th century's variability with the 19th century's variability (Figure 5). We can expose for each rhythmic

variable two or more centuries with visible different dispersion and that is evidence against homoscedasticity. (a) and (b) are the main factors against the credibility of the ANOVA results. The Post Hoc Bonferroni is also affected by factor (a) above. A nonparametric ANOVA would show similar problems. The Kruskal-Wallis ANOVA, for example, requires the assumption of a common continuous distribution with the same shape for all the factors. The Welch's ANOVA also requires homoscedasticity (for how the violations of nonparametric ANOVA assumptions could affect the statistical inference based on the underlying statistic, see Fagerland & Sandvik, 2009). If our data cannot support the normality assumption and shows evidence of heteroscedasticity, we need to find an alternative methodology to detect a significant difference between the groups.

[5]. The mathematical formulation and rule of decision of the predictive are as follows. At time t', if for the new sample size n' the behavior of Y' is given by Bin(n',p), and considering that we observed at time t the value y in n independent realizations, the conditional expected value of Y', denoted by E, is given by the equation in (i). The conditional variance of Y', denoted by V, is given by (ii). When the Y' random variable is constructed using a big sample size n', as in the case under analysis, we have that $(Y'-E)/V^{1/2}$ has a distribution approximately standard Normal (as a consequence of the Central Limit Theorem).

(i) $E=E(Y'/ t, n, y, t',n', \alpha, \beta)$

$=n'(\alpha+y)/(\alpha+\beta+n)$

(ii) $V=Var(Y'/ t,n,y,t',n', \alpha, \beta)$

$=n'(\alpha+y)(\beta+n-y)(\alpha+\beta+n+n')/((\alpha+\beta+n)^2(\alpha+\beta+n+1))$

Given a level γ we characterize an observation y' of the random variable Y' as being predictable (at level $1-\gamma$) from Y (with the conditions t, n, y, t', n', α, β), if the observed value y' is located within the credibility region given by the Normal distribution, or if $(y'-E)/V^{1/2}$ is located within $[-z_\gamma, z_\gamma]$ (where z_γ is the 1-$\gamma/2$ quantile of the standard Normal distribution). In our analysis, we fix $\gamma=0.05$, then $z_{0.05}= 1.96$.

Acknowledgments

This work results from collaboration between Laboratório de Fonética (CLUL/FLUL) and the Tycho Brahe Project (IEL, Unicamp; MacLinC Project, USP), and it was supported by research grants from the Portuguese Foundation for Science and Technology (PTDC/LIN/70367/2006), the Brazilian Fapesp (04/03643-0) and the Brazilien CNPq (grants 485999/2007-2 and 303421/2004-5). We would like to thank the audiences of *XXIII Encontro Nacional da Associação Portuguesa de Linguística* (Évora, 2007), and of the *Frequency in Phonology Workshop* (Lisbon, 2010), for comments on different parts of this research. Special thanks are due to the anonymous reviewers for their most useful comments. Thanks are also due to Marisa Cruz and the research assistants of the FreP project for precious help with data collection, and to Joe Butler for help with the editing of the present version.

References

Abercrombie. David. (1967). *Elements of General Phonetics*. Chicago: Aldine.

Adams, Marianne. P. (1987). *Old French, null subjects and verb second phenomena*. Ph.D. Dissertation, University of California, Los Angeles.

Aven, Terje, & Eidesen, Karianne. (2007). A predictive Bayesian approach to risk analysis in health care. *Medical Research Methodology* 7:38. http://www.biomedcentral.com/1471-2288/7/38.

Beckman, Mary, & Edwards, Jan. (2000) Lexical frequency effects on young children's imitative productions. In Michael B. Broe & Janet Pierrehumbert (Eds.), *Papers in Laboratory Phonology V (Acquisition and the Lexicon)*. Cambridge: Cambridge University Press. 208-218.

Bybee, Joan. (2001). *Phonology and Language Use.* Cambridge: Cambridge University Press.

Bybee, Joan, & Hooper, Paul. (Eds.) (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam/Philadelphia: John Benjamins.

Castro, Ivo. (2006). *Introdução à história do Português*. Lisboa: Edições Colibri.

Dasher, R. & Bolinger, Dwight. (1982). On pre-accentual lengthening. *Journal of the International Phonetic Association* 12:58-69.

Chafe, Wallace. (1988). Punctuation and the Prosody of Written Language. *Written Communication* 5 (4):395-426.

Chomsky, Noam. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Cutler, Anne, & Norris, Dennis. (1988). The role of stressed syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14:113-121.

Dauer, Rebecca. (1983). Stress-timing and syllable-timing reanalized. *Journal of Phonetics* 11:51-62.

Dauer, Rebecca. (1987). Phonetic and Phonological Components of Language Rhythm. In *Proceedings of the XIth International Congress of Phonetic Sciences*. 268-274.

Demuth, Katherine. (2006). Crosslinguistic Perspectives on the Development of Prosodic Words. Introduction. *Language and Speech* 49(2):129-135.

Dimitrova, Snezhina. (1997). Bulgarian Speech Rhythm: Stress-Timed or Syllable-Timed? *Journal of the International Phonetic Association* 27:27-33.

Duarte, Denise, Galves, António, Lopes, Nancy & Maronna, Ricardo. (2001). The statistical analysis of acoustic correlates of speech rhythm. Paper presented at the Workshop on Rhythmic patterns, parameter setting and language change. ZiF, University of Bielefeld.

Dufresne, Monique. (1993). *L'articulation syntaxique et phonologique de la cliticisation: le cas des pronoms sujets en moyen français*. Ph.D. Dissertation, Université du Québec à Montréal.

Dunsmore, I. R. (1974). The Bayesian Predictive Distribution in Life Testing Models. *Technometrics* 16(3):455-460.

Edwards, Jan, Munson, Benjamin, & Beckman, Mary E. (2011). Lexicon-Phonology Relationships and Dynamics of Early Language Development. *Journal of Child Language* 38(1):35-40.

Fagerland, Morten W. & Sandvik, Leiv. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials* 30:490-496.

Fine, Alex, Qian, Ting, Jaeger, T Florian, & Jacobs, Robert. (2010) Is there syntactic adaptation in language comprehension? In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics. 18-26.

Frota, Sónia, & Vigário, Marina. (2001). On the correlates of rhythmic distinctions: the European/Brazilian Portuguese case. *Probus* 13(2):247-275.

Frota, Sónia, Vigário, Marina, & Martins, Fernando. (2002). Language Discrimination and Rhythm Classes: Evidence from Portuguese. In *Speech Prosody 2002 Proceedings*. Aix-en-Provence. 315-318.

Frota, Sónia, Vigário, Marina, & Martins, Fernando. (2006) FreP: An Electronic Tool for Extracting Frequency Information of Phonological Units from Portuguese Written Text. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa. CD-ROM.

Galves, António, & Galves, Charlotte. (1995). A case study of prosody driven language change. From ClP to EP. UNICAMP - USP, ms.

Galves, Charlotte, Britto, Helena, & Paixão de Sousa, M. Clara. (2005). The Change in Clitic Placement from Classical to Modern European Portuguese. *Journal of Portuguese Linguistics* 4(1):39-67.

Galves, Charlotte & Faria, Pablo. (2010). *Tycho Brahe Parsed Corpus of Historical Portuguese*. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Galves, Charlotte, & Paixão de Sousa, M. Clara. (2010). The loss of verb-second in the history of Portuguese: Subject position, Clitic placement and Prosody. Paper given at Digs XII, Cambridge.

Gneiting, Tilmann, & Raftery, Adrian E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359-378.

Gonçalves Vianna, A. R. (1892). *Exposição da pronúncia normal portuguesa, para uso dos Nacionais e Estrangeiros*. Lisboa: Imprensa Nacional.

Gupta, A. & Nadarajah, S. (2004). *Handbook of Beta Distribution and its Applications*. New York: Marcel Dekker, Inc.

Hendrickx, Iris, & Marquilhas, Rita. (2012). From old texts to modern spellings: an experiment in automatic normalisation. *Proceedings of the Workshop on Annotation of Corpora for Research in the Humanities*, Heideberg University, Germany.

Hussaini, Essam, & Hussein, M. (2011). Bayes Prediction of Future Observables from Exponentiated Populations with Fixed and Random Sample Size. *Open Journal of Statistics* 1. [doi:10.4236, http://www.scirp.org/journal/ojs/]

Johnson, Mark, & Goldwater, Sharon. (2009). Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American*

Chapter of the ACL. Boulder, Colorado: Association for Computational Linguistics. 317–325.

Kleinhenz, Ursula. (1997). Domain Typology at the Phonology-Syntax Interface. In Gabriela Matos, Matilde Miguel, Inês Duarte & Isabel Faria (Eds.), *Interfaces in Linguistic Theory*. Lisbon: APL/Colibri. 201-220.

Lightfoot, David. (2006). *How new languages emerge*. Cambridge: Cambridge University Press.

Lloyd, James. (1940). *Speech Signals in Telephony*. London.

Manrique, Ana M.B. & Signorini, Angela. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics* 11: 117-128.

Marquilhas, Rita. (2003). Mudança analógica e elevação das vogais pretónicas. In Castro, I. & Duarte, I. (eds.) *Razões e emoção*. *Miscelânea de estudos em homenagem a Maria Helena Mira Mateus*, 2. Lisboa: Imprensa Nacional-Casa da Moeda. 7-18.

Martins, Ana Maria. (1994). Clíticos na História do Português. Unpublished PhD. dissertation. Faculdade de Letras da Universidade de Lisboa.

Martins, Fernando, Vigário, Marina, & Frota, Sónia. (2009). FreP – Frequências no Português. Software. http://www.fl.ul.pt/ LaboratorioFonetica/frep

Mateus, Maria H. & Andrade, Ernesto. (2000). *The Phonology of Portuguese*. Oxford: Oxford University Press.

McEnery, Tony, & Wilson, Andrew. (2001) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mehler, Jacques, & Nespor, Marina. (2002). Linguistic rhythm and the acquisition of language. In Adriana Belleti (Ed.) *Structures and Beyond*. Oxford: Oxford University Press, Vol.3, 213-222.

Mehler, Jacques, Jusczyk, P., Dehaene-Lambertz, G., Bertoncini, N. & Amiel-Tison, C. (1988). A percursor of language acquisition in young infants. *Cognition* 29:143-178.

Mersad, Karima, Goyet, Louise, & Nazzi, Thierry. (2010). Cross-linguistic differences in early word form segmentation: a rhythmic-based account. *Journal of Portuguese Linguistics* 9(2)/10(1):37-65.

Milotte, Séverine, Morgan, James, Margules, Sylvie, Bernal, Savita, Dutat, Michel, & Christophe, Anne. (2010). Phrasal prosody constrains word segmentation in French 16-month-olds. *Journal of Portuguese Linguistics* 9(2)/10(1):67-86.

Morgan, J. L. (1986). *From simple input to complex grammar*. Cambridge, Mass.: MIT Press.

Morgan, J. L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of memory and Language* 35:666-688.

Nazzi, Thierry, Bertoncini, J & Mehler, J. (1998). Language discrimination by newborns: towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24(3):756-766.

Nespor, Marina. (1990). On the rhythm parameter in phonology. In Iggy Roca (Ed.), *Logical Issues in Language Acquisition*. Dordrecht: Foris. 157-175.

Paixão de Sousa, M. Clara. (2004). *Língua Barroca: Sintaxe e História do Português dos Seiscentos*. Ph.D. dissertation, UNICAMP.

Perfors, Amy, Tenenbaum, Joshua B., Griffiths, Thomas, & Xu, Fei. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition* 120:302-321.

Pike, Kenneth L (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.

Quednau, Laura. (2000). *O acento do Latim ao Português Arcaico*. PhD. dissertation, Pontifícia Universidade Católica do Rio Grande do Sul, Brazil.

Ramus, Franck, & Mehler, Jacques. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *JASA* 105:512-521.

Ramus, Franck, Nespor, Marina & Mehler, Jacques. (1999). Correlates of linguistic rhythm in speech. *Cognition* 73:265-292.

Robert, C. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. Dordrecht: Springer.

Roberts, Ian. (2007). *Diachronic Syntax*. Oxford: Oxford University Press.

Sandler, Wendy. (1999). The Medium and the Message: Prosodic Interpretation of Linguistic Content in Sign Language. *Sign Language and Linguistics* 2 (2):187-216.

Serrano, Francisca, Genard, Natalie, Sucena, Ana, Defior, Sylvia, Alegria, Jesus, Mousty, Philippe, Leybaert, Jacqueline, Castro, São L. & Seymour, Philip. (2010). Variations in reading and spelling acquisition in Portuguese, French and Spanish: A cross-linguistic comparison. *Journal of Portuguese Linguistics* 9(2)/10(1):183-204.

Speyer, Augustin. (2008). *Topicalization and Clash avoidance. On the interaction of prosody and syntax in the history of English with a few glances at German*. Ph.D. Dissertation, University of Pennsylvania.

Teyssier, Paul. (1980). *Histoire de la langue portugaise*. Paris: PUF.

Vigário, Marina. (2003). *The Prosodic Word in European Portuguese*. Berlin/New York: Mouton de Gruyter.

Vigário, Marina, Frota, Sónia & Martins, Fernando. (2010). A frequência que conta na aquisição da fonologia: types ou tokens? In Ana M. Brito, Fátima Silva, João Veloso & Alexandra Fiéis (Eds.), *XXV Encontro Nacional da Associação Portuguesa de Linguística: Textos Seleccionados*. Porto: Associação Portuguesa de Linguística. 749-767.

Williams, Edwin B. (1975) *Do Latim ao Português*. Rio de Janeiro: Tempo Brasileiro. (3rd edition).

Table 1. P-properties and the stress-timing/syllable timing profiles

Stress-timing	Syllable-timing
Not CV dominant	CV dominant
Unbalanced C / V ratio	Balanced C / V ratio
Shorter PW (+PW1)	Longer PW (-PW1)
Stronger PW, stronger stress	Weaker PW, weaker stress

Table 2. P-properties and the rhythmic change from Classical to Modern European Portuguese

Stress-timing		Syllable-timing
Not CV dominant		CV dominant
Unbalanced C / V ratio		Balanced C / V ratio
Shorter PW (+PW1)	→	Shorter PW (+PW1)
Stronger PW, stronger stress	→	Stronger PW, stronger stress

Table 3. The subset of the Tycho Brahe corpus used in our study

Author	Gândavo	Pinto	Sousa	Brandão	Vieira	Vieira	Chagas	Bernardes	Oliveira
Date	1502	1510	1556	1584	1608	1608	1631	1644	1702
Type	narr	narr	narr	narr	lett	serm	phil	phil	lett
Author	Aires	Costa	Alorna	Garrett	Garrett	Fronteira	Camilo	Ortigão	Total
Date	1705	1714	1750	1799	1799	1802	1826	1836	564,554
Type	phil	lett	lett	lett	theat	narr	narr	lett	words

Table 4. Getting phonological properties from written text by means of the FreP tool: an illustration with the text Não se pode duvidar que há muitas províncias ‘It cannot be doubted that there are many provinces’

Orthographic words	Intermediate phonological analysis	Phonological template
#Não#	#n´ãGN#	#C´VGN#
#se#	#se#	#CV#
#pode#	#p´o.de#	#C´V.CV#
#duvidar#	#du.vi.d´ar#	#CV.CV.C´VC#
#que#	#ke#	#CV#
#há#	#´á#	#´V#
#muitas#	#m´uGN.tas#	#C´VGN.CVC#
#províncias#	#pro.v´íN.cGas#	#CCV.C´VN.CGVC#

Table 5. Definition of the variables in the predictive analysis

Success	Failure	Universe
PW1	PW2	Prosodic words with 1 and two syllables
PWm ((C)V)	PW2	Prosodic words with 1 and two syllables
Fin	Penult	Prosodic words with two, three or more syllables
Fin or PW1	Penult	Prosodic words with one, two or more syllables

Figure 1. Ratio between simple (S) and complex (C) syllables from 1502 to 1990-2010. The dates on the x-axis represent a text written by an author born in that specific year.

Figure 2. Ratio between consonants (C) and vowels (V), by century (16th to 20th). The mixed text from 1990-2010 is plotted as belonging to the 20th century. Each data point corresponds to a text.

Figure 3. Proportion of disyllables (in percentage), by century (16th to 20th). Each data point corresponds to a text.

Figure 4. Ratio between monosyllables (PW1) and disyllables (PW2), by century (16th to 20th). Each data point corresponds to a text.

Figure 5. Ratio between monomoraic PW (PWm) and disyllables (PW2), by century (16th to 20th). Each data point corresponds to a text.

Figure 6. Ratio between the final stress pattern (Fin) and the penultimate stress pattern (Penult), by century (16th to 20th). Each data point corresponds to a text.

Figure 7. Ratio between the final stress pattern including monosyllables (Fin+PW1) and the penultimate stress pattern (Penult), by century (16th to 20th). Each data point corresponds to a text.

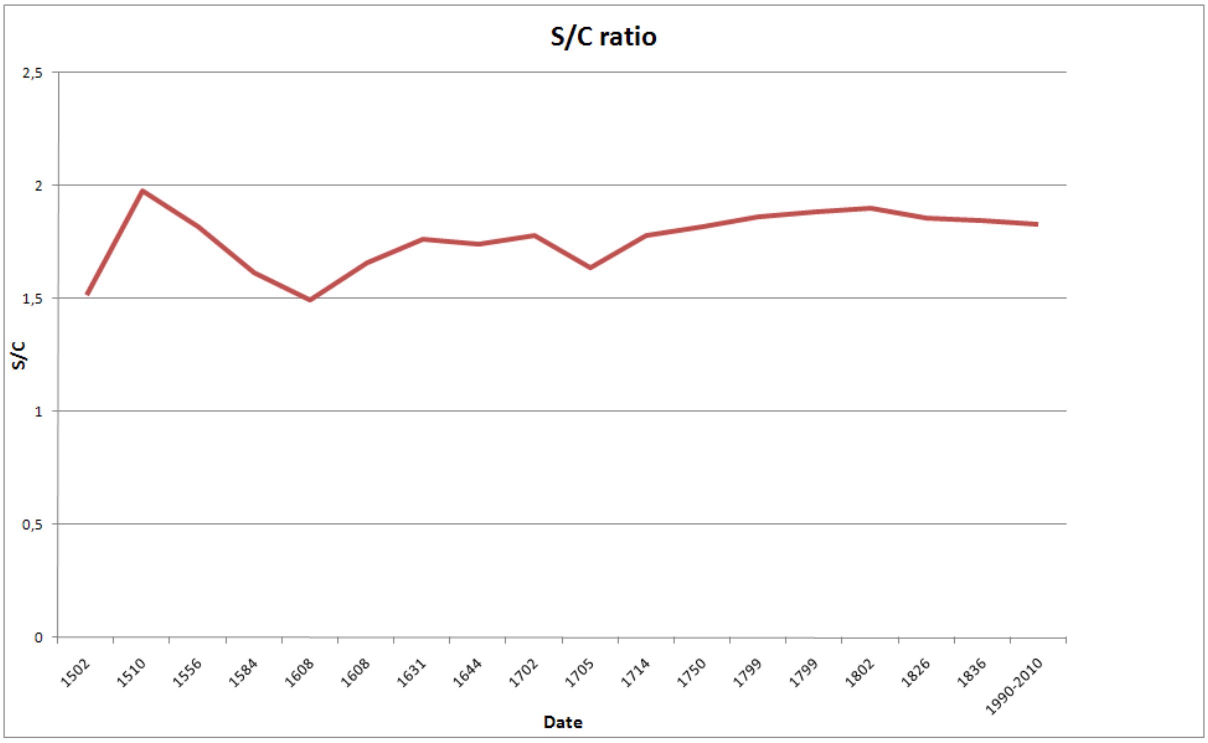
Figure 8. The four indicators studied (each value divided by the respective median value) plotted against the C/V ratio (full line).

Figure 9. The data points (divided by the respective median value) grouped by century (16^{th} to 20^{th}). The 1802 data point was excluded (see text).

Figure 10. Results for the prosodic word shape rhythmic indicators. The left panel shows the predictions assuming PW1 as success and PW2 as failure. The right panel shows the predictions assuming PWm as success and PW2 as failure .

Figure 11. Results for the stress-related rhythmic indicators. The left panel shows the predictions assuming Fin as success and Penult as failure. The right panel shows the predictions assuming Fin and PW1 as success and Penult as failure.

Figure 1



S/C ratio
2,5
2
1,5
1
0,5
0
S/C
1502
1510
1556
1584
1608
1608
1631
1644
1702
1705
1714
1750
1799
1799
1802
1826
1836
1990-2010
Date

Figure 2

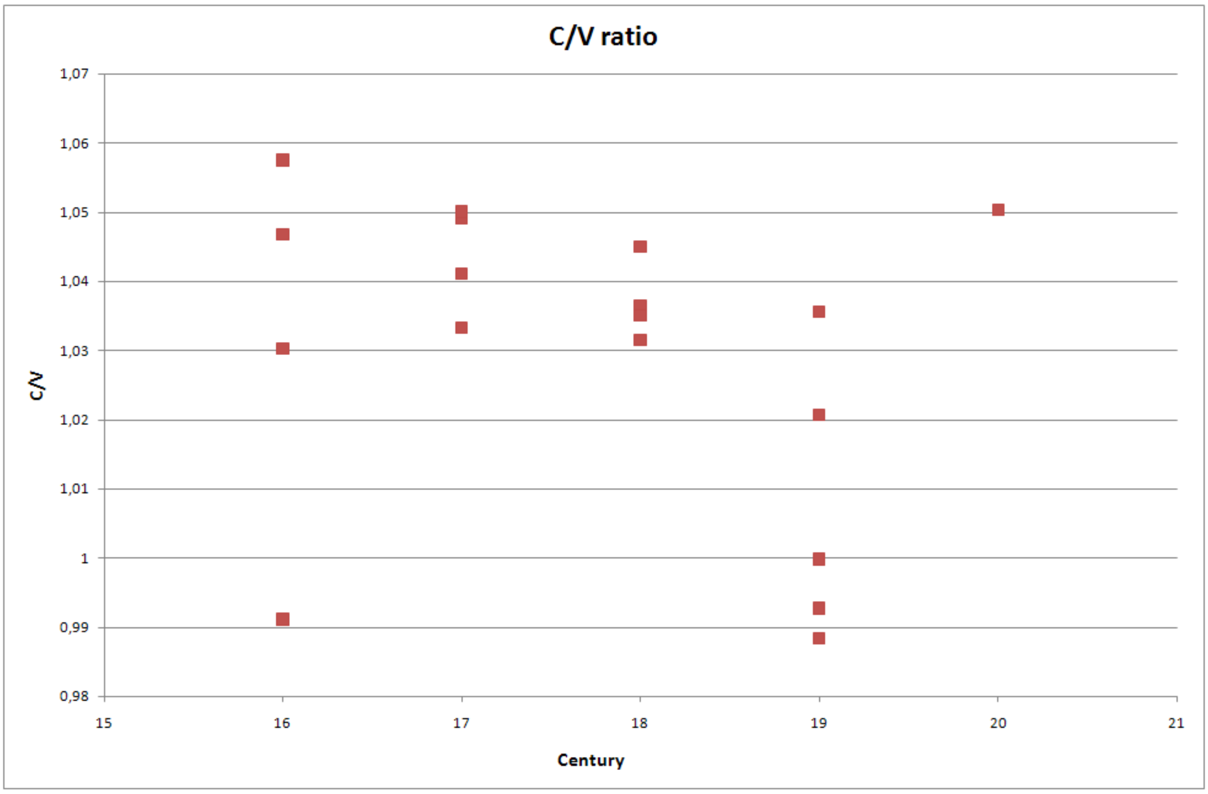

C/V ratio
1,07
1,06
1,05
1,04
1,03
1,02
1,01
1
0,99
0,98
C/V
15
16
17
18
19
20
21
Century

Figure 3

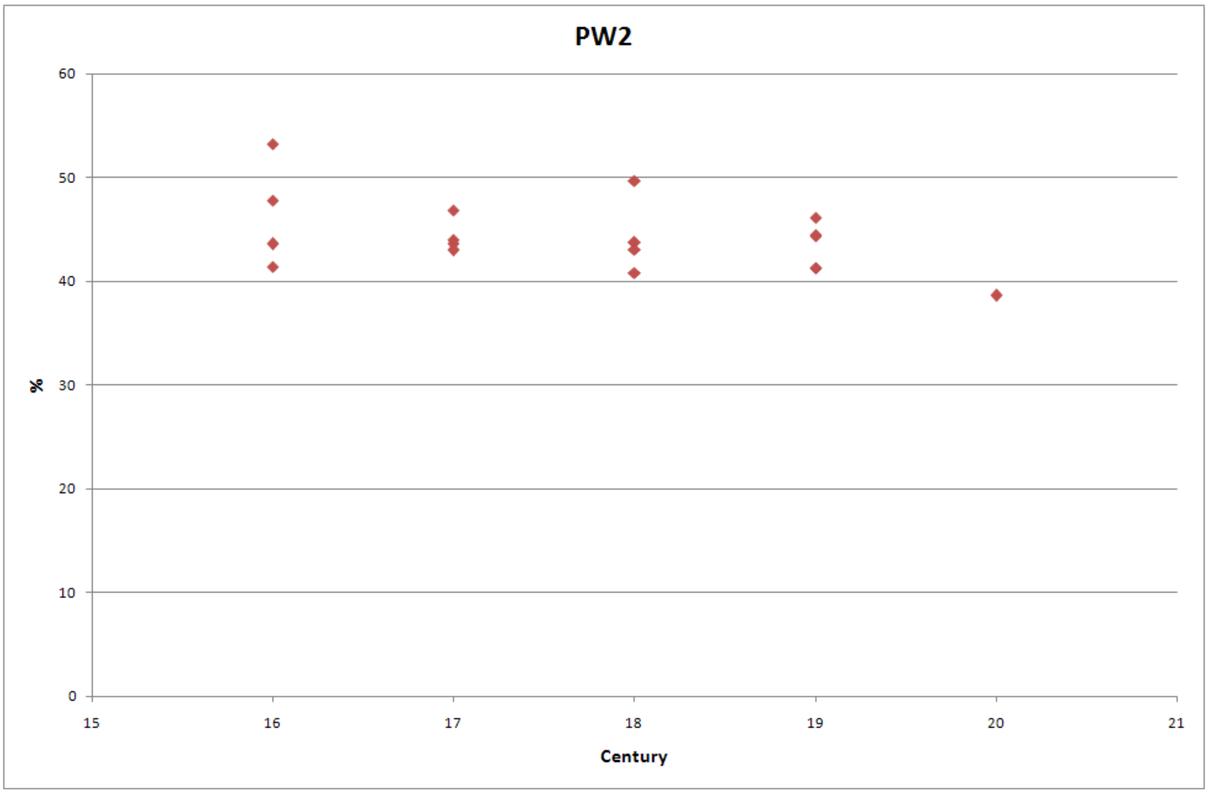


Figure 4

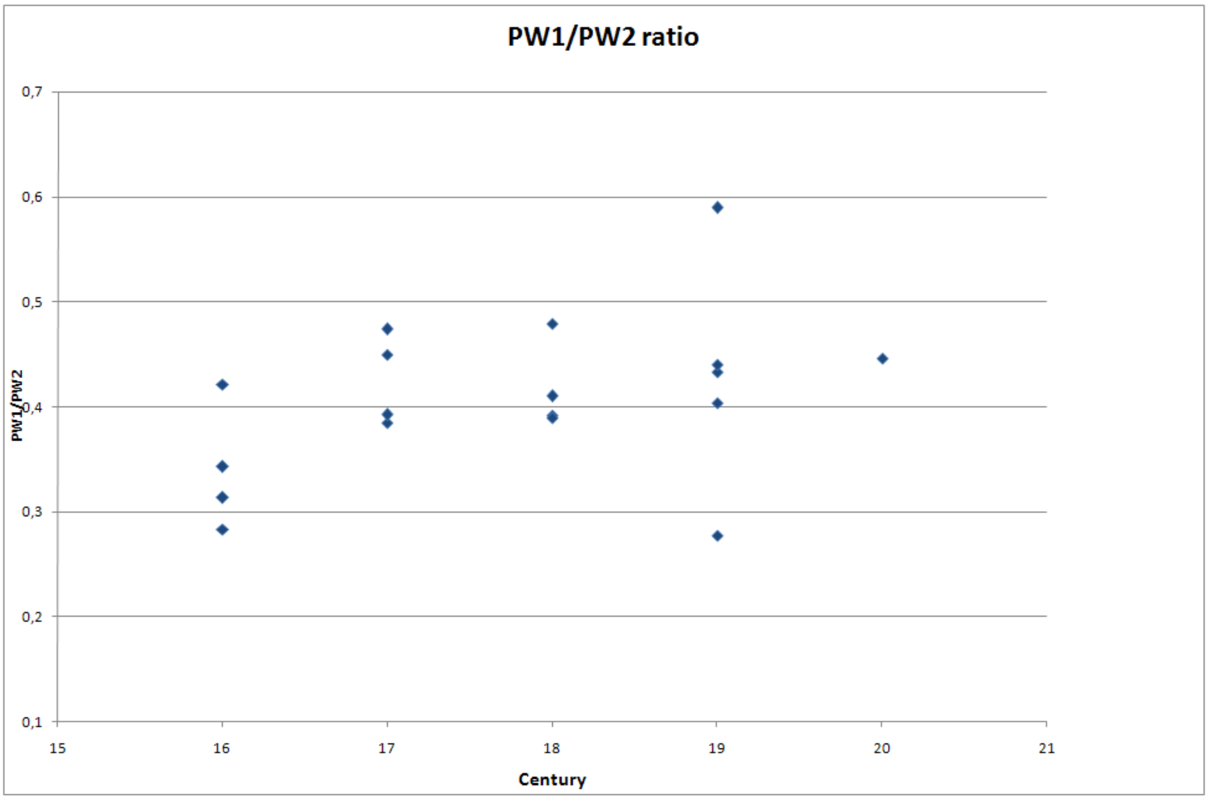

PW1/PW2 ratio
PW1/PW2
0,7
0,6
0,5
0,4
0,3
0,2
0,1
15
16
17
18
19
20
21
Century

Figure 5

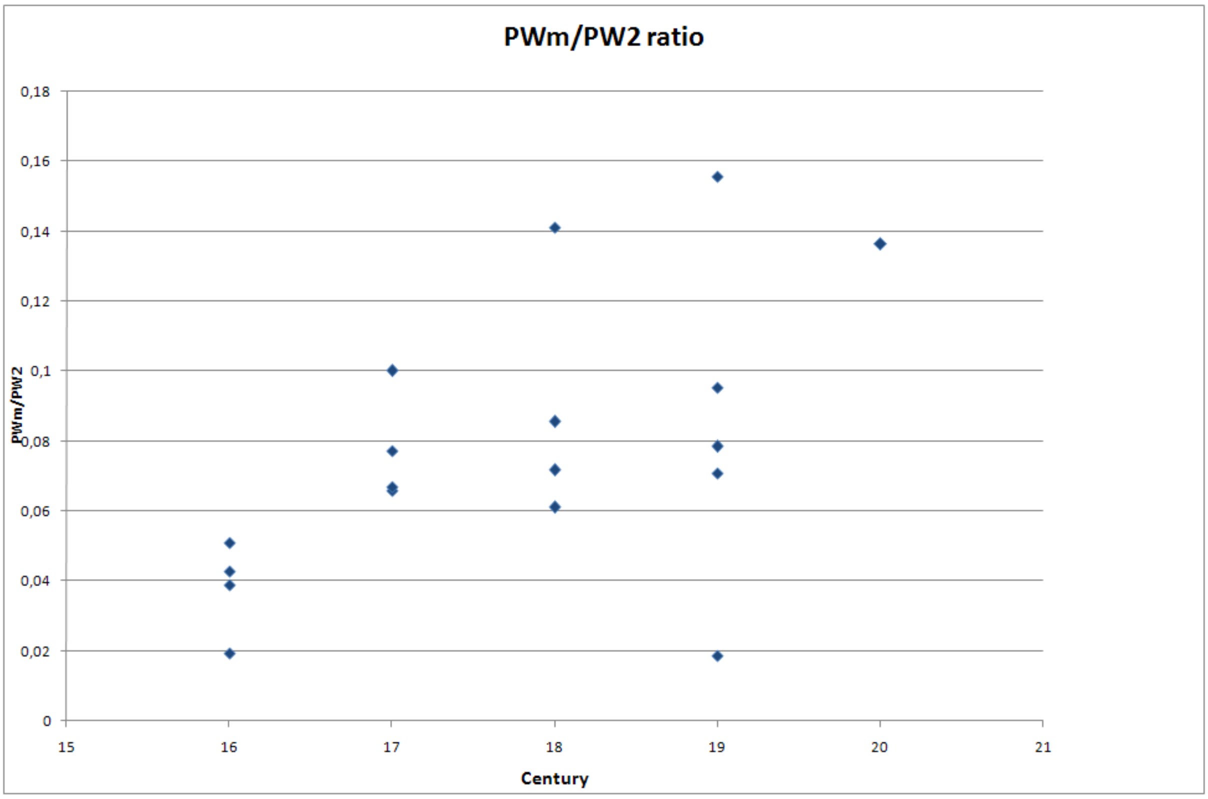

PWm/PW2 ratio
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
PWm/PW2
15
16
17
18
19
20
21
Century

Figure 6

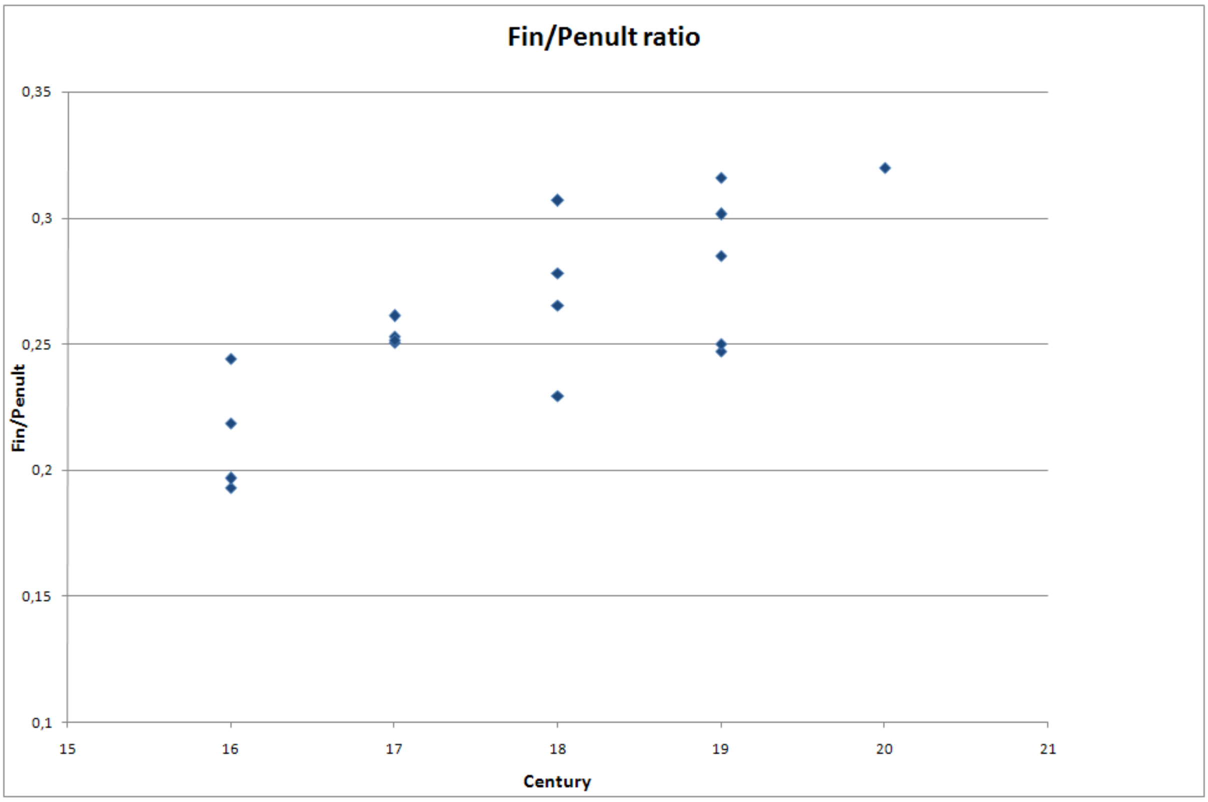

Fin/Penult ratio
0,35
0,3
0,25
0,2
0,15
0,1
Fin/Penult
15
16
17
18
19
20
21
Century

Figure 7

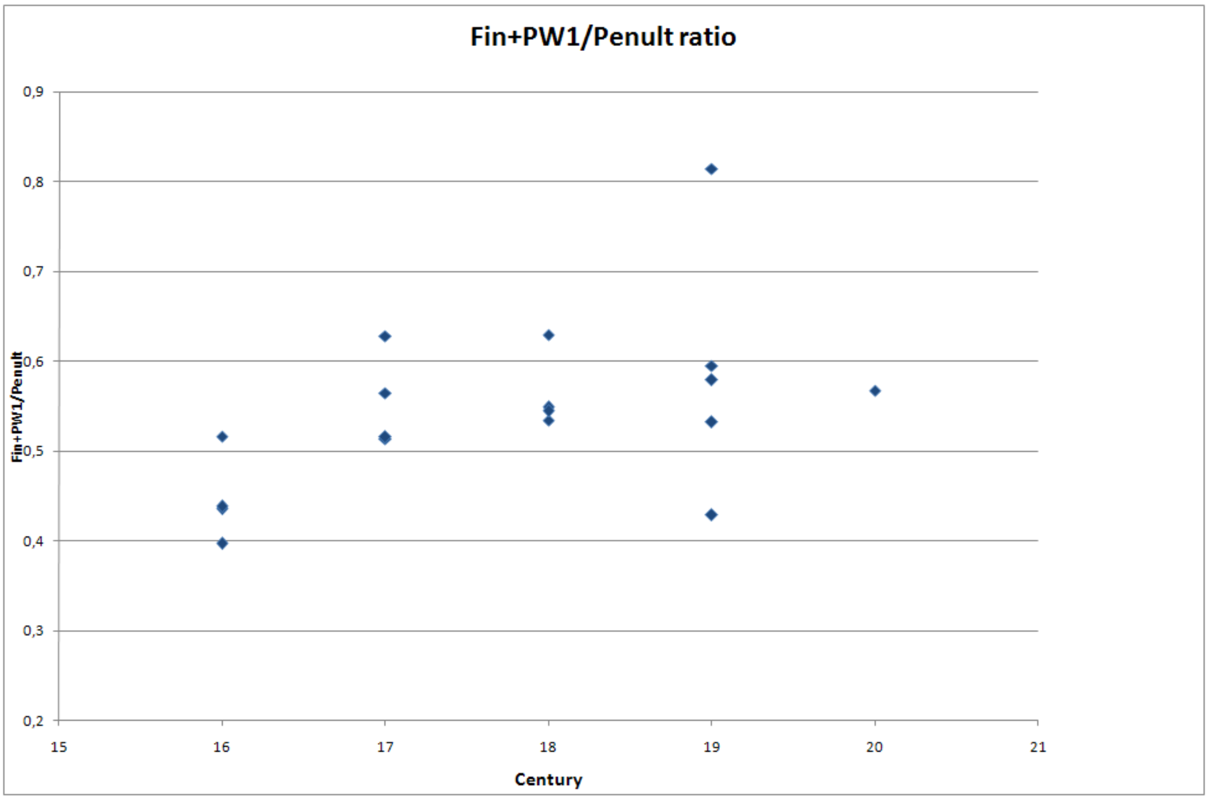

Fin+PW1/Penult ratio
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
Fin+PW1/Penult
15
16
17
18
19
20
21
Century

Figure 8

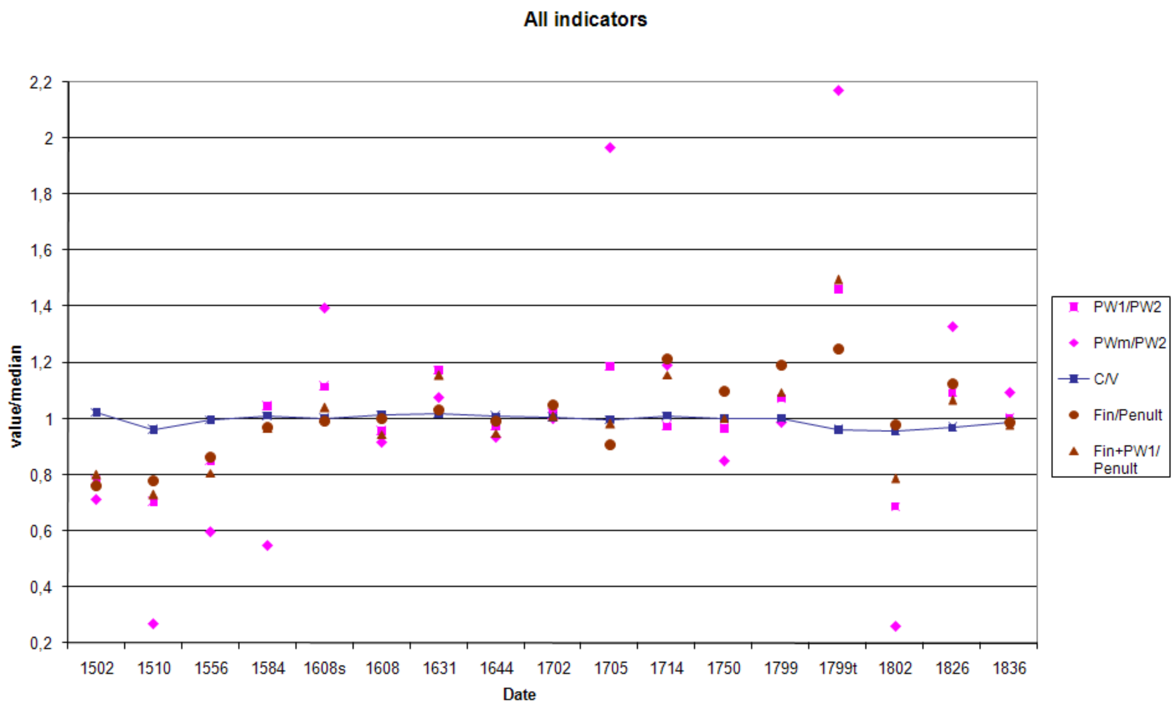

All indicators
value/median
2,2
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
1502
1510
1556
1584
1608s
1608
1631
1644
1702
1705
1714
1750
1799
1799t
1802
1826
1836
Date
PW1/PW2
PWm/PW2
C/V
Fin/Penult
Fin+PW1/ Penult

Figure 9

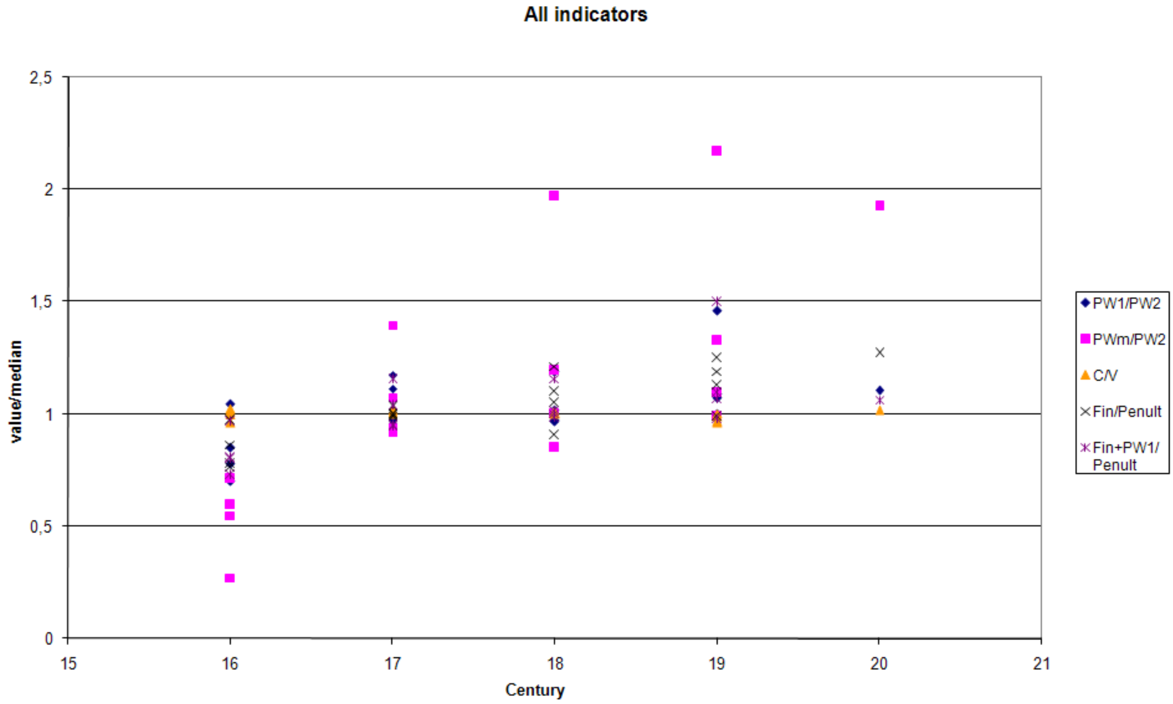

All indicators
value/median
Century
2,5
2
1,5
1
0,5
0
15
16
17
18
19
20
21
PW1/PW2
PWm/PW2
C/V
Fin/Penult
Fin+PW1/
Penult

Figure 10 (left panel)

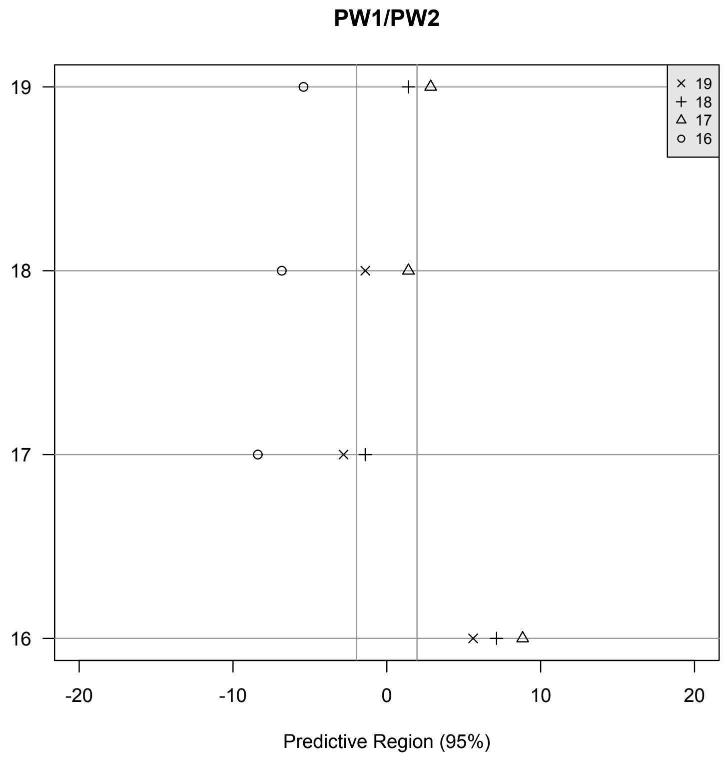


Figure 10 (right panel)

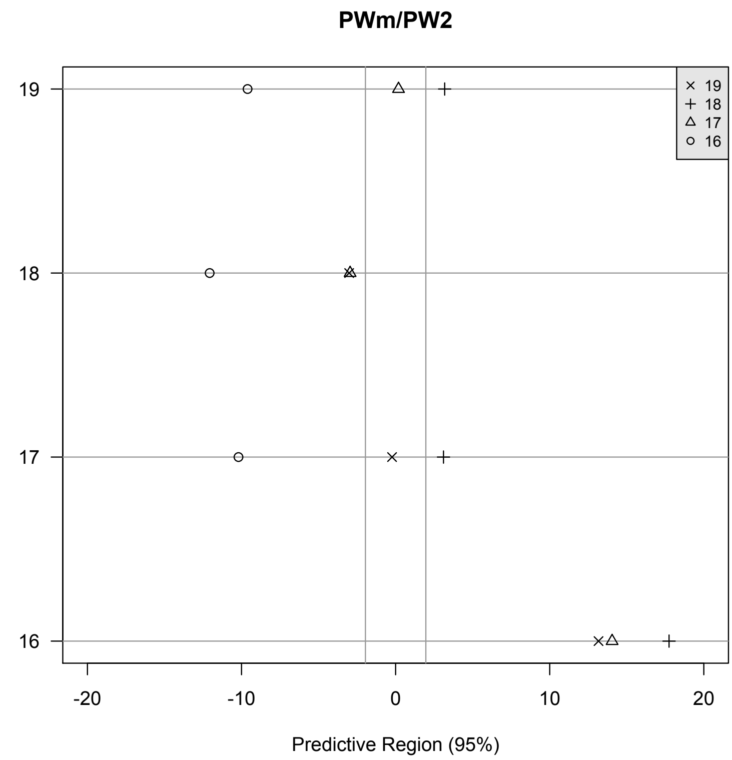


Figure 11 (left panel)

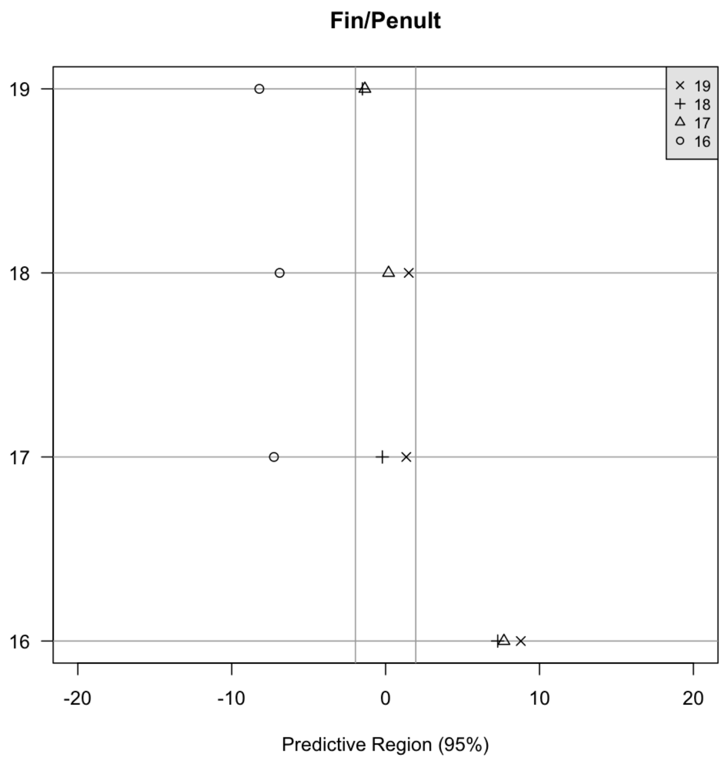


Figure 11 (right panel)