

## The usefulness of metrics in the quantification of speech rhythm

Amalia Arvaniti

[aarvaniti@ucsd.edu](mailto:aarvaniti@ucsd.edu), Tel: +1 858 534 8409, Fax: +1 858 534 4789

Dept. of Linguistics, University of California, San Diego, La Jolla, CA 92093-0108, USA

### Abstract

The performance of the rhythm metrics  $\Delta C$ , %V, PVI and Varcos, said to quantify rhythm class distinctions, was tested using English, German, Greek, Italian, Korean and Spanish. Eight participants per language produced speech using three elicitation methods, spontaneous speech, story reading and reading a set of sentences divided into “uncontrolled” sentences from original works of each language, and sentences devised to maximize or minimize syllable structure complexity (“stress-timed” and “syllable-timed” sets respectively). Rhythm classifications based on pooled data were inconsistent across metrics, while cross-linguistic differences in scores were often statistically non-significant even for comparisons between prototypical languages like English and Spanish. Metrics showed substantial inter-speaker variation and proved very sensitive to elicitation method and syllable complexity, so that the size of both effects was large and often comparable to that of language. These results suggest that any cross-linguistic differences captured by metrics are not robust; metric scores range substantially within a language and are readily affected by a variety of methodological decisions, making cross-linguistic comparisons and rhythmic classifications based on metrics unsafe at best.

### Keywords

Rhythm, rhythm metrics, English, German, Spanish, Italian, Greek, Korean

## The usefulness of metrics in the quantification of speech rhythm

### 1.0 Introduction

The notion that languages can be classified for speech rhythm into a small set of classes, stress-, syllable- and mora-timing, dates from the early 20<sup>th</sup> c. and has remained popular despite the fact that, as noted by Bertinetto (1989: 100) “[...] no other phenomenon of phonology is so widely accepted, with so little supporting evidence.” The popularity of rhythm classes seemed to wane by the early 1990s, due to the lack of empirical support alluded to by Bertinetto, but it received a new boost with the advent of rhythm metrics, formulas that aim at quantifying the timing characteristics of distinct rhythm classes (among others, Ramus, Nespors & Mehler, 1999; Frota & Vigário, 2001; Grabe & Low, 2002; Wagner & Dellwo, 2004; Dellwo, 2006; White & Mattys, 2007).

The present study tests the most widely employed metrics – those proposed by Ramus et al. (1999), Grabe & Low (2002) and Dellwo (2006) – using a large number of speakers from six languages, and a variety of materials and elicitation methods. The aim was to examine the extent to which discrepancies between previous studies that used metrics can be attributed to their different methodological choices: regular differences could plausibly be attributed to methodology and could be constrained; random discrepancies would point to fundamental problems with metrics.

#### 1.1 Brief historical overview

Traditional descriptions of speech rhythm have relied on the notion of isochrony, that is, the idea that rhythm rests on regulating the duration of particular units in speech, syllables in syllable-timed languages, stress feet in stress-timed languages, and moras in mora-timed languages. Thus in this view, rhythm is based exclusively in durational patterns or *timing* (indeed the terms *rhythm* and *timing* have often been used as synonyms in this literature; for a discussion see Arvaniti 2009).<sup>1</sup>

Ideas of this sort can be found in Jones [1972: 237, 242 (1918)], and were first tested by Classé (1939) who attempted to find isochrony in English stress feet and concluded that isochrony is possible only for phonetically and syntactically homogeneous feet. The idea of a rhythmic typology in particular was first presented in Lloyd James (1940: 25) who described English, Arabic and Persian as having “Morse code rhythm” and French and Telugu as having “machine-gun rhythm.” Shortly after, Pike (1945: 34-35) coined the terms *stress-timing* and *syllable-timing* which he used to juxtapose the rhythm of English to that of Spanish. According to Pike, English is stress-timed because rhythm units “tend to follow one another in such a way that the lapse of time between the beginning of their prominent syllables is somewhat uniform;” Spanish is syllable-timed because “it is the syllables, instead of the stresses, which tend to come at more-or-less evenly recurrent intervals.” Jinbo [1980 (1927)], cited in Warner & Arai (2001), appears to be the first reference to the mora-timing of Japanese, suggesting that moras have approximately equal duration. The notion of rhythm as isochrony was most strongly expressed in Abercrombie (1967: 97) who recognized two rhythmic classes, stress- and syllable-timing, and proposed that all languages belong to one or the other class.

Attempts to find isochrony in production have proven unsuccessful time and again. A number of early experiments measuring feet in English have shown that foot duration is proportional to the number of syllables they contain (Shen & Peterson, 1962; Bolinger, 1965; Uldall, 1971; see Lehiste, 1977, for a review of early studies of this topic). This applied also to reiterant speech, used by Nakatani, O’Connor & Aston (1981) to test isochrony. Studies of

---

<sup>1</sup> Arvaniti (2009) briefly presents preliminary results based on part of the present corpus before data collection and analysis were completed. Although the results in the studies differ only in minor details, the present quantitative results which are based on the complete corpus supersede those in Arvaniti (2009).

languages other than English have also found no support for isochrony. Studies of syllable-timed languages show no evidence that syllable duration is kept constant (e.g., Wenk & Wioland, 1982, on French; Pointon, 1980, on Spanish), while studies of both stress- and syllable-timed languages that measured both syllable and foot durations conclude that isochrony is absent from both (e.g., Balasubramanian, 1980, on Tamil; Major, 1981, on Brazilian Portuguese; Borzone de Manrique & Signorini, 1983, and Pointon, 1995, on Spanish). In addition, studies that compared stress- and syllable-timed languages note more similarities than differences between languages said to be belong to different classes (e.g., Roach, 1982, on French, Telugu, Yoruba, English, Russian and Arabic; Dauer, 1983, on English, Italian, Spanish, Greek and Thai; Bertrán, 1999, on English, Russian, Spanish, Catalan, Portuguese, French and Italian). Similarly, Warner & Arai (2001) also conclude that there is little evidence in support of Japanese mora-timing.

Fewer studies have examined rhythm perception with respect to the rhythm class hypothesis, but those that have provide equally inconclusive results. Lehiste (1977) suggested that isochrony is probably a perceptual phenomenon that is due to the tendency of listeners to underestimate durational differences in speech more than in non-speech stimuli. She thus suggested that the Just Noticeable Differences (JNDs) established by psychophysical experiments with non-speech stimuli might be shorter than those pertaining to speech, which she estimated at 10% of the foot duration for feet of 300-500 ms. This idea is generally supported by both Lehiste's own results and those of more recent perceptual studies (e.g., McAuley & Riess Jones 2003). However, it remains the case that many studies (including Lehiste, 1977) show durational differences between feet that are substantially larger than (Lehiste's own estimate of) JND.

Studies in which listeners were asked, either directly or indirectly, to classify languages into rhythm classes or to discriminate between languages of different class have yielded mixed results. The study of Scott, Isard and de Boysson-Bardies (1985) – who indirectly tested listeners' responses to rhythm – indicates that isochrony may not relate to rhythmic class at all. Scott et al. asked English and French listeners to tap to word-initial consonants (corresponding to beats) in both French and English utterances. They expected English listeners to tap more isochronously than French listeners to both French and English stimuli. Their results, however, showed that French listeners were more isochronous in their tapping to stimuli of both languages, though in both groups inter-tap intervals were more even than the beats in the stimuli.

Miller (1984) – the only study in which listeners were directly asked to classify languages as stress- or syllable-timed – found that the task was practically impossible even when the participants were phonetically trained. Specifically, Miller asked English and French phoneticians and non-phoneticians to rhythmically classify Arabic, Finnish, Indonesian, Japanese, Polish, Spanish and Yoruba. The only classification all groups of listeners agreed on was that Arabic is stress-timed; in addition, English and French phoneticians classified Yoruba as syllable-timed. More tellingly perhaps, both French groups of listeners and English non-phoneticians classified Spanish as stress-timed. Generally, non-phoneticians showed less of a tendency to place languages in different classes than phoneticians did, a result Miller attributes to the possibility that non-phoneticians were not all attending to the same cues, while phoneticians “might [have been] influenced by received ideas” (p. 82), especially if they could recognize the languages of the experiment.

Results similarly unresponsive of rhythm classes are reported by Arvaniti (in press) who asked listeners to rate modified utterances of English, German, Greek, Italian, Korean and Spanish for similarity to a series of non-speech trochees, hypothesizing that stress-timed languages would be rated more similar to trochees (since their rhythm, said to be based on foot-initial prominences, is more akin to trochees than the assumed cadence of syllable-timed languages). Responses varied depending on stimulus modification: when low-pass filtered speech was used to modify stimuli, listeners rated all languages more similar to trochees than English; when *flat sasasa* was used – in which consonantal intervals are replaced by [s] and vocalic intervals by [a] – listeners rated English and German but also Spanish as more similar to trochees

than Italian, Greek and Korean. In short, neither experiment showed a strong listener tendency to classify languages along the lines expected by rhythm classes, while the effect that stimulus modification had on responses casts doubt on the idea of timing as the sole (and perceptually independent) exponent of speech rhythm.

Mixed results have emerged also from studies using the oddball or AAX paradigm in which listeners hear two impoverished stimuli from the same language (AA) and judge whether a third stimulus (X) belongs to the same or a different language. Ramus, Dupoux & Mehler (2003) used this paradigm with *flat sasasa*, hypothesizing that only languages belonging to different rhythm classes would be discriminated. Instead, they found that some languages such as Polish – previously classed as stress-timed (Ramus et al., 1999) – are discriminated from both English and Spanish. Moon-Hwan (2004) used the same paradigm to determine Korean rhythm and concluded that Korean is mora-timed, since Korean and Italian listeners could discriminate between Korean and Italian and between Korean and English but not between Korean and Japanese; this conclusion, however, is at odds with all other studies of Korean which lean towards syllable-timing (see section 2.1).<sup>2</sup>

An alternative approach to speech rhythm was taken by Dauer (1983, 1987). Dauer (1983) expressed doubts regarding the viability of syllable-timing as a possible basis for speech rhythm and proposed instead that rhythm is based on stress in all languages. In her view, the difference between languages like French and English does not lie in the choice of temporal interval to be kept constant but in the fact that stressed syllables are very prominent in English but much less so in French, a view harking back to Lloyd James's remarks on the "punch" of English stresses (Lloyd James, 1940: 24). Thus, Dauer decoupled rhythm from timing and advocated that languages do not fall into distinct rhythm classes but form a continuum ranging from least to most stress-based (and *not* from syllable- to stress-timing). Following up on this idea, Dauer (1987) provided a list of criteria that could be used to determine the salience of stressed syllables in a given linguistic system, but stopped short of testing the extent to which her criteria could place languages on a rhythmic continuum with any degree of consistency. Indeed, Barry and Andreeva (2001) have since shown that features such as vowel reduction – one of Dauer's criteria – apply equally to languages described as stressed-timed and languages described as syllable-timed. In addition, the results of Barry et al. (2003) suggest that Dauer's criteria may not be amenable to simple binary oppositions, such as presence vs. absence of a particular feature (see also Arvaniti, 2009, for a discussion of inconsistencies in Dauer's scheme).

Despite the lack of evidence to support it, the notion of rhythmic classes has remained popular and has been relied upon in research in phonology (e.g., Nespor & Vogel, 1989; Nespor, 1990; Coetzee and Wissing, 2007) and especially in research in language acquisition and speech processing. In particular, several studies report that infants can discriminate between languages that belong to different rhythmic classes, such as English and Italian, but not between languages that belong to the same class, such as English and Dutch (among others, Nazzi, Bertoni & Mehler, 1998; Nazzi, Jusczyk & Johnson, 2000; Nazzi & Ramus, 2003). These findings have led to proposals that language acquisition relies on speech rhythm and, in particular, on infants' ability to determine the rhythm class of their ambient language (primarily on the basis of the

---

<sup>2</sup> It is possible that discrimination using the AAX paradigm reflects differences in tempo and interactions between timing and F0 information. Rodriguez & Arvaniti (2011) show that with *flat sasasa* stimuli languages are discriminated based on tempo rather than rhythm class: e.g., fast spoken Greek and Polish are discriminated from slower-tempo English, while Danish and Korean (which have similar tempo to English) are not. Discrimination between English and Polish or Greek becomes impossible if the same stimuli are manipulated so as to eliminate tempo differences. The presence of F0 modulation can aid discrimination, especially when large differences in F0 patterns are present, as in English vs. Korean. Such results point to an interconnectedness between the processing of F0 and timing information and cast further doubt on the view of rhythm as timing (cf. Kohler, 2008; Yu, 2010; Arvaniti, in press).

characteristics of vocalic intervals). In turn, rhythm class helps infants select the unit they should use for further speech analysis and segmentation (Ramus et al., 1999). This idea has been supported by results showing that adults can discriminate better between languages of different rhythmic classes (e.g. Ramus et al., 1999; Ramus et al., 2003; but see also the earlier discussion of discrimination results and footnote 2 for a possible explanation). Support also comes from a number of studies which suggest that speech processing relies on the prosodic unit – mora, syllable or foot – on which the listeners’ native language rhythm is also based (e.g., Cutler et al., 1986, 1992; Otake et al., 1993; Cutler & Otake, 1994; Nazzi et al., 2006; Murty, Otake & Cutler, 2007; Kim, Davis & Cutler, 2008; for an alternative interpretation of some of these results, see Mattys & Melhorn, 2005).

### *1.2 Rhythm metrics*

The impetus for at least part of the psycholinguistic research that is based on the notion of rhythm classes came largely from a quantification of Dauer’s ideas as implemented first in the rhythm metrics proposed by Ramus et al. (1999). Unlike Dauer (1983), Ramus et al. presupposed the existence of rhythm classes and set out to find a set of metrics that would differentiate languages according to their traditional rhythm classifications. To do so, they focused on two of the eight diagnostic criteria proposed by Dauer (1987), syllable structure and vowel reduction, and assumed that stress-timed languages are characterized by more complex syllable structures and greater vowel reduction than syllable-timed languages. Further, they hypothesized that these two features have direct consequences on the duration of consonantal and vocalic intervals, thus reverting back to a conception of timing as the sole exponent of rhythm (a view that Dauer had advocated against). Specifically, Ramus et al. assumed that the greater syllable complexity of stress-timed languages would result in more variable consonant interval durations than in syllable-timed languages. Second, they hypothesized that vowel reduction would also result in more variability in vocalic duration for stress-timed languages than syllable-timed languages. Finally, they hypothesized that the greater complexity of consonantal clustering in combination with stress-related variability in vocalic intervals would also result in vocalic intervals occupying a smaller percentage of the signal in stress-timed than syllable-timed languages.

To test these hypotheses, Ramus et al. selected languages that had been consistently assigned to a rhythm class, stress-timed Dutch and English, syllable-timed French, Italian and Spanish, and mora-timed Japanese, as well as Catalan and Polish, languages that had been previously described as having mixed rhythm (among others, Wheeler, 2005, for Catalan; Rubach & Booij, 1985, for Polish). They concluded that the metrics best representing rhythm are  $\Delta C$ , the standard deviation of consonantal intervals in an utterance, and %V, the percentage of the utterance duration taken up by vocalic intervals. Their decision was based on the finding that these two measures best reflected the accepted classification of the languages under investigation when plotted together, creating a “rhythm space” in which languages of one rhythm type are clustered together and separately from those of the other (see e.g., Figs. 3-5 for such representations). Further, Ramus et al. found that Japanese is set apart from the other languages, validating the idea of a separate mora-timing class. Crucially, they also found that Polish and Catalan are grouped with stress- and syllable-timed languages respectively, suggesting that languages fall into distinct classes, rather than forming a continuum (contra Dauer, 1983) or having “mixed rhythm” (contra Nespors, 1990).

Grabe & Low (2002) presented a different metric, the Pairwise Variability Index (PVI), based on ideas first expounded in Low, Grabe and Nolan (2000). *Raw* PVI (rPVI) is the sum of the absolute differences between pairs of consecutive intervals (either vocalic or consonantal) divided by the number of pairs in the speech sample (see 1 below). This measure can be normalized (hence nPVI) by dividing each absolute difference between consecutive intervals by their mean; in this case, the score is multiplied by 100 to produce values comparable to those of rPVI (Grabe & Low, 2002; see 2 below). Grabe & Low assumed that consonants are less

sensitive to changes in tempo than vowels and thus proposed that the raw measure *rPVI* be used to measure variability in consonantal intervals and the normalized measure *nPVI* be used to measure variability in vocalic intervals.

$$(1) \quad rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}| / (m - 1)$$

$$(2) \quad nPVI = 100 \times \left[ \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m - 1) \right]$$

The results of Grabe & Low, which are based on data from one speaker for each of the 18 languages they tested, were not as stark as those of Ramus et al. (1999). Grabe & Low did find that prototypical stress- and syllable-timed languages, such as English, Dutch, Spanish and French, are separated in the rhythm space defined by their metrics, but they also found that many other languages – at least Greek, Malay, Romanian, Tamil, and Welsh – were placed roughly in the middle of this space, making their classification difficult if not impossible. In addition, while Ramus et al. had found that Japanese is not grouped with either syllable- or stress-timed languages, in Grabe & Low Japanese was grouped with the syllable-timed set. As a consequence of these results, Grabe & Low tentatively concluded that languages are “weakly” categorized into stress- and syllable-timing, possibly forming a continuum between the two.

The disagreement between Ramus et al. (1999) and Grabe & Low (2002) regarding the existence of distinct rhythm classes was not the only one between the two studies. Grabe & Low uncovered additional classification problems. By calculating not only PVIs but also the  $\Delta C\text{-}\%V$  metrics of Ramus et al. (1999) for their data, they found that in many cases, the two pairs of metrics classified the same dataset in different ways. Thus, PVIs classified Thai and Tamil as stress-timed but  $\Delta C\text{-}\%V$  grouped them with the syllable-timed languages; the reverse classification obtained for Luxembourgish. In addition, Greek, Catalan and Welsh, which were placed between stress- and syllable-timed languages by PVIs, were placed well within the stress-timed group by  $\Delta C\text{-}\%V$  (see Arvaniti, 2009, for a full review).

Despite such discrepancies, both  $\Delta C\text{-}\%V$  and the PVIs have remained popular. In the past decade they have been used as a means of providing rhythmic classification for a variety of languages, including Bulgarian (Barry et al., 2003), Latvian (Bond, Markus & Stockmal, 2007), Hawaiian (Parker Jones, 2006), Tamil (Keane, 2006), Greek (Grabe & Low, 2002; Tsiartsioni, 2003; Baltazani, 2007), Mandarin (Lin & Wang, 2007; Mok, 2009), Czech (Dankovicová & Dellwo, 2007), Korean (Jeon, 2006; Mok & Lee, 2008) and Cantonese (Mok, 2009). The results of many of these studies have not proved more consistent than those of Grabe & Low (2002), however. When it comes to rhythmically non-prototypical languages, metric scores are not such that they can successfully and unequivocally classify the languages in question. Some authors admit that this is so (Bond et al., 2005; Keane, 2006; Dankovicová & Dellwo, 2007; Mok & Lee, 2008), while others, such as Lin & Wang (2007), defend the accepted classification of a language despite metric scores that contradict it. In addition, even studies of prototypical languages often find that score differences between classic examples of stress- and syllable-timing, such as French and English, are not statistically significant (Grabe & Low, 2002; White & Mattys, 2007). When multiple studies of the same language are available, their results often disagree to such an extent that authors propose different classifications for the same language: e.g., Baltazani (2007) concludes that Greek has mixed rhythm, Tsiartsioni (2003) that it is syllable-timed, and Grabe & Low (2002) that is unclassifiable.

These discrepancies among studies and the difficulties in classification have led to the development of additional metrics. These are often variants of existing metrics normalized in some way. Frota & Vigário (2001) proposed the use of standard deviations of normalized percentages for vocalic and consonantal intervals to deal with languages of mixed rhythm.

Wagner & Dellwo (2004) proposed YARD (*Yet Another Rhythm Determination*), a measure similar to the PVIs in which  $z$ -transformed syllable durations (rather than raw intervals) are used for calculation. Dellwo (2006) proposed that normalized standard deviation measures of vocalic and consonantal intervals (standard deviation divided by the mean, or *Varco*) be used instead. Metrics that rely on the duration of prosodic units, rather than segments, have also been proposed: e.g., a syllable-based PVI measure was employed by Barry et al. (2003), while Nolan & Asu (2009) calculated nSPVI and nFPVI which are similar to the normalized PVI of Grabe & Low (2002) but rely on the duration of the syllable and the foot respectively.

Despite their widespread use and attempts to improve their performance, the issues with metrics discussed above have been increasingly noted by several researchers (e.g., Cummins, 2002; Barry et al., 2003; Bond et al., 2005; Keane, 2006; Dankovicová & Dellwo, 2007; Arvaniti, 2009; Barry, Andreeva & Koreman, 2009; Kohler, 2009a, 2009b; Wiget et al., 2010). Nevertheless, rhythm metrics continue to be a popular means of rhythmically classifying languages, and are also relied upon in other areas of research, including the study of language acquisition (e.g., Grabe, Watson & Post, 1999; Payne et al., 2011), bilingualism (e.g., Whitworth, 2002; Bunta & Ingram, 2007; Lleó, Rakow & Kehoe, 2007; Mok, 2011), second language learning (e.g., Bond et al., 2007; White & Mattys, 2007; Mok & Dellwo, 2008; Mok & Lee, 2008) and speech pathology (Liss et al., 2009).

Precisely because of the metrics' popularity, it is important to examine why there are discrepancies between studies of the same language and why, after the initial success with mainly Germanic and Romance languages, the classification of many other languages has proved so difficult to attain.

One possible interpretation of the variability in the results discussed above is that metrics are very sensitive to the effects that methodological choices have on the durations of consonantal and vocalic intervals and that such intra-language variability in timing may be more extensive than previously suspected. This idea is supported by the results of Wiget et al. (2010) who found that variability in the metric scores of individual sentences of English outweighed both differences due to segmentation practices among annotators and inter-speaker variation. Given results like these, it is not unreasonable to assume that sensitivity may extend to other factors, such as the way in which the data are elicited. However, it is not possible to ascertain that it is so based on existing evidence, because of the limited nature of the studies undertaken so far; e.g., in some only one speaker per language is recorded, others examine only one language, while others still rely on only one style of speech.

The aim of the present study is to address these issues together, and in particular to probe the sensitivity of metrics to various choices that are inevitable in rhythm research – such as the choice of a limited set of materials and speakers – and compare it to the differences found across languages. Documenting the reasons behind well-known discrepancies is important if one wishes to understand and ultimately constrain such differences in future metric-based research. It is also crucial because current practices rest on the assumption that metric scores represent some immutable quality of each language that can be reliably inferred from any speech sample and thus that it is possible to compare metric scores across languages and studies. By providing a measure of the sensitivity of metrics and by comparing them to one another, the results presented here can be used as a guide for making reliable comparisons across studies that use different methodologies or metrics. Finally, from a typological perspective it is important to determine how the sensitivity of metrics to extemporaneous yet inevitable methodological choices compares to cross-language effects. If languages of each rhythm class share a number of features, the effects of these methodological choices should be smaller than the cross-linguistic differences metrics intend to capture.

In order to address these issues, a sample of over 1.5 hrs of speech from six languages, English, German, Greek, Italian, Korean and Spanish, was collected; this set includes both prototypical and hard to classify languages in order to examine whether larger speech samples

would provide more stable metric scores and thus make the classification of non-prototypical languages easier. In addition, data were elicited from eight speakers of each language in order to examine the extent of inter-speaker variation in metric scores. Data were collected in three different ways (isolated sentence reading, story reading, spontaneous speech) to determine the extent to which the choice of elicitation method affects metric scores. In addition, the syllable composition of the sentences in the sentence corpus was manipulated in order to probe the sensitivity of metrics to intra-language variation and compare it to inter-language differences. Finally, since earlier research showed that scores obtained using different metrics do not always agree with one another, here comparisons are made between  $\Delta C$ , %V, PVI and Varcos – all popular metrics which have been employed in the past sometimes separately and sometimes in combination – so as to test the sensitivity of each to the methodological choices made here.

## **2.0 Methods**

### *2.1 Languages*

The prototypical languages in the study were English, German, Spanish and Italian. English and German are considered stress-timed, a classification supported by studies using metrics (Ramus et al., 1999, and White & Mattys, 2007, for English; Grabe & Low, 2002, for English and German). Spanish and Italian have been described as syllable-timed (e.g., Pike, 1945, for Spanish; Barry and Andreeva, 2001, for Italian); this classification largely agrees with reported metric scores (Ramus et al., 1999, for Spanish and Italian; Grabe & Low, 2002, and White & Mattys, 2007, for Spanish). For German and Italian, data were elicited from speakers of standard varieties, and thus the results were expected to be similar to those of previous studies. For English and Spanish, data were elicited from Southern Californian English and Standard Mexican Spanish respectively, rather than Standard British English and Standard Peninsular Spanish, the varieties studied in the past. Although there are undeniable accent differences between the varieties studied before and those used in the present study, rhythm metrics are said to reflect phonological properties, (phonologized vowel reduction and syllable structure complexity) which do not substantially differ across varieties. Note also that Pike's original distinction between stress- and syllable-timing is based on American English and Mexican Spanish. Thus, using different varieties of English and Spanish should not have a dramatic effect on the results.

In addition, the study included Greek and Korean, languages that have been difficult to classify. Lee et al. (1994) found that Korean may be changing from stress-timed among the older speakers to syllable-timed among the younger generation, while Yun (1998), unable to find strong evidence in favor of one or the other rhythm class, suggested that the rhythm of Korean is phoneme based. Two recent studies have used rhythm metrics to classify Korean, but without reaching an unequivocal classification: Jeon (2006: 38) concludes that "Seoul Korean is likely to be more syllable-timed than Southern British English, though Korean cannot be definitely categorized as a syllable-timed language;" similarly, Mok & Lee (2008) concluded that Korean has mixed rhythm but is probably closer to syllable-timing than to stress-timing. In contrast, Kim et al. (2008) concluded that Korean is clearly syllable-timed, on the basis of perceptual data showing that Korean speakers process French and Korean materials in a similar syllable-based manner (see Kim et al., 2008, for a review of additional studies on Korean rhythm). As mentioned, however, Moon-Hwan (2004), reaches an entirely different conclusion on the basis of AAX experiments, namely that Korean is mora-timed. The widely different classifications of these studies may have to do with the fact that some aspects of Korean prosody, stress and vowel quantity in particular, are unclear. Specifically, de Jong (1994) and Jun (1995) suggest that Korean does not have stress, but Lee (1999) argues that Korean does have stress that is linked to vowel weight distinctions; according to some, these distinctions have disappeared from Seoul Korean (Jun, 2005), while others find them to be still active (Yoshida, Yoon & Kim, 2007).

The rhythmic classification of Greek is equally uncertain (for a review see Arvaniti, 2007). Dauer (1983) places Greek in the middle of her continuum, but somewhat closer to the

“most stressed-based” end, and notes that stress salience is substantial in Greek. Indeed, in Dauer (1980) phonetically naive native speakers of Greek and two trained phoneticians (one of whom did not speak the language) showed very good agreement concerning the placement of stresses in Greek running speech. On the other hand, Arvaniti (1994) points out that if all of Dauer’s criteria are taken into account, Greek should be placed towards the “least stress-based” end of her continuum. Barry & Andreeva (2001) treat Greek as a syllable-timed language, while Grabe & Low (2002) maintain it was unclassified before their study and conclude that it is essentially unclassifiable by PVIs (though, as noted, it is stress-timed by  $\Delta C-\%V$ ). Tsiartsioni (2003), on the other hand, reports low PVI scores for Greek and concludes it is syllable-timed, while Baltazani (2007) suggests it is placed between prototypical stress-timed languages like German and prototypical syllable-timed languages like Spanish and likely has mixed rhythm with high vocalic but low consonantal variation.

## 2.2 Elicitation methods and materials

As mentioned, partly the motivation for the study was to examine the effect that the choice of elicitation methods could have on rhythmic scores. For this reason, all the main methods of eliciting data in studies using metrics were employed here: reading a set of sentences (e.g., Ramus et al., 1999; Wagner & Dellwo, 2004; White & Mattys, 2007), reading a story (e.g., Grabe & Low, 2002; Keane, 2006) and producing spontaneous speech (e.g., Lin & Wang, 2007). For clarity, these three ways of eliciting data are referred to as *elicitation methods*.

For the read running speech part, the story of “The North Wind and the Sun” (henceforth *story*) was selected. The versions used were those available in IPA illustrations of the six languages of the study, namely Ladefoged (1999) and Hillenbrand (2003) for English, Kohler (1999) and Fleisher & Schmid (2006) for German, Arvaniti (1999) for Greek, Rogers & d’Arcangeli (2004) for Italian, Lee (1999) for Korean, Martínez-Celdrán, Fernández-Planas & Carrera-Sabaté (2003) for Spanish. These can be found in Appendix A.

In order to elicit from the participants between one and two minutes of speech (henceforth *spontaneous speech*), a set of topics was developed. The first topic suggested to the speakers was their experiences with parking at the University of California, San Diego campus. Since parking is a source of frustration to everyone on campus, it was anticipated that this was a topic all members of the university community – who formed the bulk of the participants – would be likely to have an opinion or anecdotal story about. When this was not the case, the topics of public transportation, airport security or difficulties with roommates were used. In case a participant was unable to talk on any of the topics provided, they were asked to describe a set of three single boxes from Calvin and Hobbes comics. For the Greek speakers, who were recorded in Greece, these topics were not appropriate, so they were asked to talk about themselves and their experience of living in Athens, a request they had no difficulty complying with.

The set of sentences was designed with two aims in mind. First, the sentences as a set were meant to be compared to the other elicitation methods, read running speech (*viz. story*) and spontaneous speech. As noted, the sentences were also designed to examine the extent to which metrics are sensitive to variability within a language sample (for a similar treatment that the present study actually predates see also Prieto et al., in press). The effect of sentence composition on metric scores is a crucial issue, since many studies are based on very small sentence corpora; e.g., both Ramus et al. (1999) and White & Mattys (2007) relied on five sentences per participant. It is not inconceivable that such small datasets may not provide stable information, and indeed, Wiget et al. (2010) have shown that the composition of sentences can significantly affect metric scores.

Here, three sets of five sentences each for each language were devised (for the full list, henceforth collectively referred to as *sentences*, see Appendix B; for examples, see Table 1). One set (henceforth the “uncontrolled” set) consisted of five sentences selected from original works of each language. The criteria used to select the “uncontrolled” sentences were that they be

meaningful out of context and between fifteen to twenty-five syllables in length (a relatively large variation that was, however, necessary in a sample of six languages). The other two sets (henceforth “stress-timed” and “syllable-timed” sets) contained sentences that were similar to the “uncontrolled” sentences in terms of length and structure, but were designed to enhance syllable complexity and simplicity respectively: “stress-timed” sentences were designed to incorporate as much variability as each language allowed; e.g., sentences included consonant clusters, geminates (where appropriate), instances of vowel hiatus, diphthongs and so on; “syllable-timed” sentences on the other hand, showed simple syllable structure and, to the extent this was possible, did not include combinations that would contribute to the durational variability of either consonantal or vocalic intervals; e.g., the Italian “syllable-timed” sentences included practically no geminates, while those of English contained as few consonant clusters as possible.

The three elicitation methods together yielded between 15-17 minutes of speech per language (approximately 2 minutes per speaker) for a total corpus of 96 minutes of speech. This corpus is substantially larger than those used in previous studies on metrics (which range from an estimated minimum of 1.5 min in Wiget et al., 2010, to an estimated maximum of 9 min in Grabe & Low, 2002).

**Table 1 English and Spanish examples of each sentence type**

Language	Sentence type	Sentence
English	“stress-timed”	<i>The production increased by three fifths in the last quarter of 2007.</i>
	“syllable-timed”	<i>Two-year-old Lucy has macaroni and cheese every day for diner.</i>
	“uncontrolled”	<i>Some little boys had come up on the steps and were looking into the hall.</i>
Spanish	“stress-timed”	<i>Un zoólogo estaba inspeccionando unos especimenes nuevos.</i> 'A zoologist was examining some new specimens.'
	“syllable-timed”	<i>No sé si mi jefe se relajará la próxima semana.</i> 'I don't know if my boss will relax next week.'
	“uncontrolled”	<i>Las oficinas estaban cerradas y oscuras por el día feriado.</i> 'The offices were closed and dark for the holiday.'

### 2.3 Speakers

Results are based on the data from eight participants of each language. Additional speakers were recruited for some languages, but their data were not included for several reasons. Technical problems resulted in severely degraded recordings for three German, four Korean and six Italian speakers. Spanish speakers from Puerto Rico and Colombia were initially recorded but excluded when it became possible to record speakers entirely from the same dialect, Standard Mexican Spanish, as the aim was to keep dialectal variation within each language to a minimum. Finally, five Greek speakers were not included in the study in order to keep the number of participants equal among languages (the excluded Greek speakers were those recorded last).

The English, German, Italian, Korean and Spanish speakers were recruited from the student population of UC San Diego; some were paid for their participation but most took part for course credit. The Greek speakers were recorded in Greece and were on average older than the speakers of the other languages (see Table 2); they all refused payment for their participation. All English participants were monolingual with no language other than English spoken in their homes; they were all natives of Southern California where they had lived their entire lives. The native speakers of Spanish were from Mexico, and the German speakers were from northern Germany. The German, Greek, Italian, Korean and Spanish speakers all reported speaking with a standard accent (e.g., Seoul Korean, Athenian Greek); they had all grown up in their respective countries where they had spoken only their native language in their homes. The native speakers of German, Italian, Korean and Spanish had all learned English as a second language in their home country and spoke it fluently and frequently (since they resided in the US), but they also used

their native language on a regular basis. The native Greek speakers were all college graduates and reported some familiarity with English, but did not speak it fluently or use it frequently. None of the speakers reported any history of speech or hearing disorders and they were all naïve as to the purposes of the experiment.

#### 2.4 Procedures

The English, German, Italian, Korean and Spanish participants were recorded at the UC San Diego Speech Lab. The Greek participants were recorded in Athens, Greece in a quiet room either at their home or place of work. Participants first signed Institutional Review Board consent forms and filled out a language background form. They were then asked to familiarize themselves with a printed copy of the sentences and the North Wind and the Sun story. The spontaneous speech portion of the recording was explained to them and they were asked to select a topic they would be comfortable speaking about.

For all languages except Korean, the reading portion of the experiment was displayed to the participants as a PowerPoint presentation on a computer monitor. For the sentences, participants saw one sentence at a time in 16 point Arial font, left-justified and centered vertically on the screen. For read running speech, the entire story was presented on the computer screen using the same font and size. (Due to problems with the proper display of Korean fonts, Korean materials were presented in the same size and orientation on cardstock, but otherwise the same procedure was followed as for the other languages.) The order of the sentences was pseudo-randomized so that no more than two sentences of the same type appeared in a row, and the order of the three elicitation methods (sentences, story, spontaneous speech) was counterbalanced across subjects within each language.

**Table 2 Participant demographics**

Language	Age Range (mean)	Years in U.S. (mean)	Females	Males
English	18-22 (20.1)	N.A.	5	3
German	22-32 (25.6)	0 – 7 (2.1)	4	4
Greek	36-48 (41.6)	N.A.	5	3
Italian	21-39 (27.3)	0 – 3 (0.9)	7	1
Korean	19-30 (22.8)	0 – 6 (2.4)	5	3
Spanish	18-25 (23.9)	0 – 11 (7.1)	5	3

For the read materials (sentences and story), the participants were asked to read aloud at their natural pace and advance the presentation using the spacebar when they were ready (Korean participants were asked to move to the next card). They were also told to repeat any disfluent sentences before moving on. Several speakers, however, were reluctant to do so and as a result some utterances from the two read corpora have pauses which, on occasion, include false starts. Pauses and disfluent parts were excluded from measurement (and are not included in the calculation of the total corpus duration).

Speakers participating at UC San Diego were recorded directly to disk using an AD converter and Wavepad at 44.1 kHz sampling rate with 16-bit quantization. Participants recorded in Greece were recorded using Wavepad on a PC laptop at 20 kHz sampling rate with 16-bit quantization.

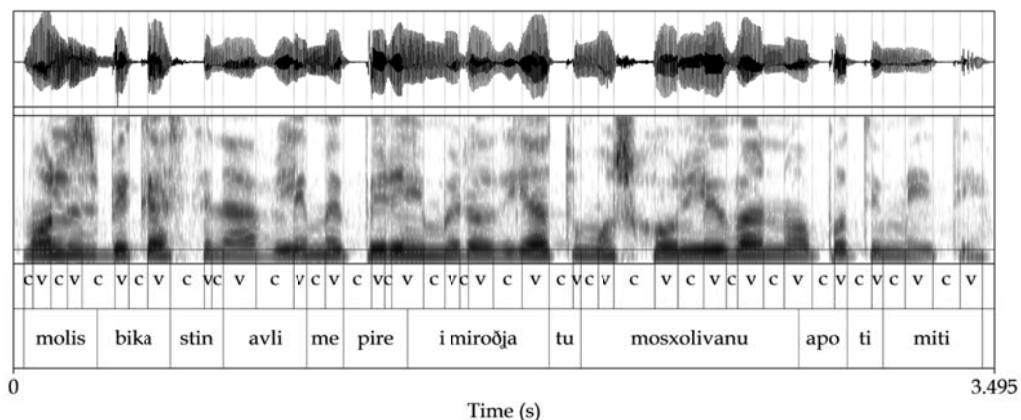
#### 2.5 Measurements

Before measurements were made, recordings were separated into smaller files containing one utterance each. This step was taken to ensure that speaker means were based on scores from individual utterances for the data from all elicitation methods. For the story, a standard segmentation into six to eight utterances was devised for each language based on a native

speaker’s intuition of where speakers were likely to pause. This was necessary because all versions of the story contained multi-clause sentences which speakers could treat as distinct utterances. When speakers did not pause as expected, the speaker’s own prosodic phrasing was followed instead.

The durations of the spontaneous speech portions of the recordings ranged from 45 seconds to two minutes. In order to standardize the duration of the recordings across participants, the first minute of each recording was used, unless the recording was shorter, in which case it was used in its entirety. The separation of utterances was done on the basis of pause placement so not all utterances were complete sentences. Filled pauses were excluded from analysis.

Measurements of consonantal and vocalic intervals were made by simultaneous inspection of spectrograms and waveforms using Praat and following standard segmentation criteria. Two additional considerations guided the measurements: first, a reliance on phonetic criteria rather than the phonological function of segments and a desire to accurately represent the durational profile of each language. The reasoning behind these considerations was as follows: metric scores are used to validate the idea that infants determine the rhythm class of their language and use this information at an early stage of acquisition that precedes the acquisition of syllable structure details; if so, then infants must rely on purely phonetic criteria to distinguish vocalic and consonantal intervals (see Nolan & Asu, 2009, for similar arguments). Based on this, syllabic consonants were included in consonantal intervals and glides were classified based on their phonetic profile: they were included in consonantal intervals if they showed evidence of frication, but in vocalic intervals if they did not (see Fig. 1 for an example). Second, it was decided that measurements should neither exclude intervals nor include intervals that could not be accurately measured. In order to satisfy this criterion, three practices were established: (i) prepausal intervals were not excluded from measurement (unlike previous studies, such as Grabe & Low, 2002); segments separated by a pause were treated as two distinct intervals, since they would be more likely to be perceived as such (unlike Grabe & Low, 2002, and White & Mattys, 2007); utterance-initial voiceless stops, voiced stops without a clearly visible voice bar and phrase-final unreleased stops were not measured. One exception was made regarding pre-pausal intervals: any such intervals were excluded from measurement in data from the sentence corpus that showed utterance-internal pauses. This was based on the fact that in the vast majority of cases (76.5% of the total sentence corpus) the sentences were produced without pauses as intended. Thus, including prepausal intervals in the rest of the data would result in extraneous differences between sentences that included pauses and those that did not being included in their metric scores.



**Fig. 1** Illustration of segmentation criteria using one of the “uncontrolled” Greek sentences (‘As soon as I entered the yard, I could strongly smell the frankincense’ by Sp2); note that the speaker did not apply the apocope indicated in the spelling of the original text (Appendix B)

For each sentence,  $\Delta C$ , %V, rPVI, nPVI, VarcoC and VarcoV were calculated. As mentioned,  $\Delta C$  is the standard deviation of consonantal interval durations across an utterance, and %V is the percentage of the utterance duration taken up by vocalic intervals (Ramus et al. 1999);<sup>3</sup> the PVI measures, rPVI and nPVI, are raw and normalized measures used to measure consonantal and vocalic interval variability respectively; for clarity, in the remainder of the paper they are referred to as rPVI-C and nPVI-V respectively (for the calculation of PVIs, see equations 1 and 2 in section 1.2). Finally, as shown in equation (3) VarcoC and VarcoV are both normalized standard deviations, that is standard deviations ( $\Delta C$ ) divided by the mean; scores are multiplied by 100 to create values comparable to those of other metrics (Dellwo, 2006).

$$(3) \text{Varco}\Delta C = \Delta C * 100 / \text{mean}C$$

Scores were calculated separately for each sentence in the isolated sentence set and for each utterance in the story and spontaneous speech sets. To avoid different weighting of the means of the three elicitation methods (since each set contributed a slightly different number of observations), speaker means were calculated from their means for each elicitation method. Mean language scores were then calculated from the three mean scores of each speaker.

### 2.6 Statistical analysis

As mentioned in the introduction, one of the purposes of the study was to compare metrics to each other. If metrics have a consistent relation to each other, they would correlate. Establishing such correlations would facilitate the comparison of results across studies that employ different metrics. To this effect, correlations between metrics were run on the story data and the “uncontrolled” sentences, arguably the most stable parts of the overall corpus. Consonantal metrics were correlated to each other, and the same applied to vocalic metrics. The correlations were run on speaker averages for the story (48 observations), and the scores of the individual sentences in the “uncontrolled” sentence set (240 observations). They were calculated both for data pooled across languages and for each language separately.

Since the results of these correlations did not show consistent relationships between metrics (see section 3.1), metric scores were further analysed by means of analyses of variance (ANOVAs). Specifically, mean speaker scores for each metric from each elicitation method were subjected to repeated-measures ANOVAs with language as the categorical predictor and elicitation method (sentences, story, spontaneous speech) as a repeated-measures factor. In addition, the mean scores from the three sentence types were subjected to separate repeated-measures ANOVAs with sentence type (“stress-timed”, “syllable-timed”, “uncontrolled”) as a repeated-measures factor and language as a categorical predictor. Pairwise comparisons of main effects and significant interactions were examined by means of Fischer LSD post-hoc tests. The Fischer LSD was chosen because it is suitable for complex designs (Cohen & Cohen 1983: 172-176; Davis & Gaito, 1984) and relatively liberal: since it is known from previous studies (e.g., White & Mattys, 2007) that effect sizes for metrics tend to be small, relying on LSD increased the chances of recording significant differences between languages, a desired effect in order not to “stack the deck” against rhythm metrics (as less liberal post-hoc tests would). Reported differences for pairwise comparisons based on the Fischer LSD are significant at  $p < 0.05$ .

In addition to the above analyses, ANOVAs were also run in which the pairs of vocalic and consonantal metrics most frequently used together in the literature (%V and  $\Delta C$ , nPVI-V and rPVI-C, VarcoV and VarcoC, and %V and VarcoC) were treated together, as two levels of a repeated-measures factor (with language as categorical predictor and elicitation or sentence type

---

<sup>3</sup> The scale of  $\Delta C$  values varies depending on whether the intervals are measured in milliseconds or seconds; here all interval durations were converted to milliseconds, so  $\Delta C$  values are comparable to those of %V and the other metrics.

as a second repeated-measures factor). The aim for these analyses – which do not differ in other respects from the analyses described above either in design or results – was to see whether the combined effect size of each pair of a consonantal and a vocalic metric would be greater than that of each metric alone, i.e. whether using two metrics together would enhance the language effect.

As an estimate of effect size partial  $\eta^2$  was calculated for language and the two main manipulations of the study, sentence type and elicitation method, both for each metric separately and for each pair of metrics, so as to see, as noted, if the performance of metrics is enhanced when they are used in tandem. (For elicitation in particular, partial  $\eta^2$  was also calculated from ANOVAs that excluded the “stress-timed” and “syllable-timed” sentences to see if the omission of the most “skewed” materials would enhance the language effect size or reduce that of elicitation.) Strictly speaking, partial  $\eta^2$  does not allow one to extrapolate from the present study to the general population but it does allow us to compare the size of the different effects in the study, by showing the percentage of the variance of the dependent variable (*viz.* the scores of each metric or pair of metrics) that is attributed to this effect and its associated error.

Finally, in order to aid in the analysis of the results, consonantal and vocalic scores (%V and  $\Delta C$ , nPVI-V and rPVI-C, VarcoV and VarcoC, and %V and VarcoC) were used to calculate Euclidean distances between individual languages, elicitation methods, sentence types and speakers from different reference points. The use of Euclidean distances is based on the fact that, as mentioned in section 1.2, it is standard practice for pairs of vocalic and consonantal metric scores to be used as coordinates in order to determine the rhythm class of a language on the basis of its position in the space defined by the two metrics. Despite the prevalence of this practice in the literature, the actual distances between languages in rhythm space are often not quantified (e.g., Grabe & Low, 2002; White & Mattys, 2007; for a discussion of the pitfalls of this practice, see Arvaniti, 2009). Euclidean distances provide precisely this quantification.

### 2.7 Predictions

It was expected that the scores of some metrics may correlate with those of others, but, given the variability among studies so far (such as the different overall results for PVIs and  $\Delta C$ -%V reported in Grabe & Low, 2002) these correlations were not expected to be strong.

The following hypotheses were made with respect to the experimental manipulations. Pooled metric scores were expected to show a separation of English and German from Spanish and Italian in metric space. Given previous results (Grabe & Low, 2002; Tsiartsioni, 2003; Jeon, 2006; Baltazani, 2007; Mok & Lee, 2008, *inter alia*) there was no expectation that placing Greek and Korean within one or the other class would be entirely consistent either within or across metrics, but there was an expectation that the larger sample would render the results less variable than those of previous, smaller-scale studies.

Regarding the three elicitation methods, increasing variability (that is higher scores) for all metrics except %V was expected with increased similarity to natural speech: thus, in general, it was expected that isolated sentences would show less interval variability than read running speech and that read running speech would in turn show less variability than spontaneous speech. With respect to sentence type, it was hypothesized that the “stress-timed” sets would show higher scores (i.e. more variability) and that the “syllable-timed” sets would show lower scores (with the exception of %V for which the trend was expected to be the reverse). Uncontrolled sentences were expected to have values intermediate between the other two sets.

## 3. Results

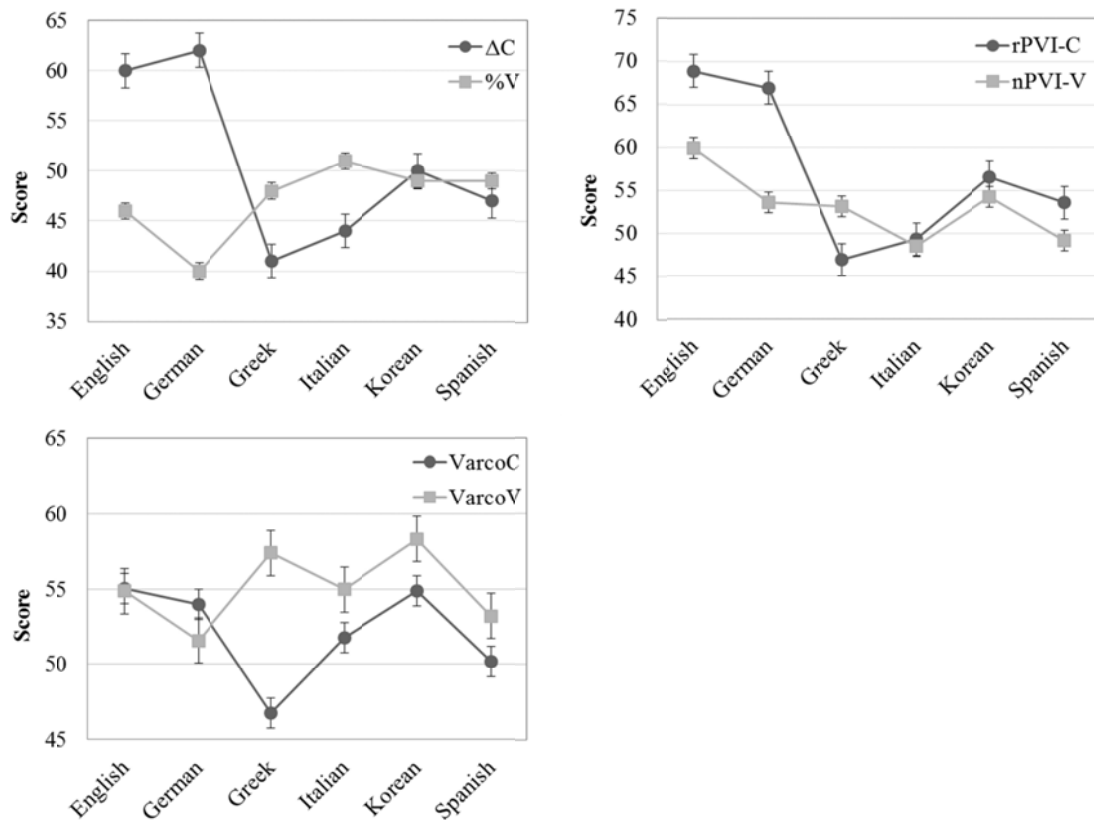
### 3.1 Correlating metrics

Regarding the story data, a strong correlation was found between  $\Delta C$  and rPVI-C for the pooled data, with more modest correlations between  $\Delta C$  and VarcoC and between rPVI-C and VarcoC. The strong correlation between  $\Delta C$  and rPVI-C was replicated in the data for English, German, Italian and Korean, but the results of Greek and Spanish did not reach significance; no individual

language results reached significance for the correlations involving VarcoC, except for Spanish in the correlation between  $\Delta C$  and VarcoC (see Table 3). For the vocalic scores, the pooled data showed only a weak negative correlation between %V and nPVI-V but no other correlations were significant either for pooled data or individual languages (see Table 3).

Regarding the “uncontrolled” data set, modest to strong correlations were found between all three consonantal metrics for the pooled data and most within-language comparisons. For the vocalic metrics, on the other hand, only nPVI-V and VarcoV showed a modest correlation that applied to the pooled data and all languages except English (see Table 3).

Since the correlations between metrics were neither consistent nor consistently strong, the scores of all metrics were further analyzed statistically by means of ANOVAs.



**Fig. 2** Mean language scores and standard errors for  $\Delta C$  and %V (panel a), rPVI-C and nPVI-V (panel b) and VarcoC and VarcoV (panel c)

### 3.2 Language differences

All metrics showed a statistically significant main effect of language [for %V,  $F(5,42) = 27.4$ ;  $p < 0.0001$ ; for  $\Delta C$ ,  $F(5,42) = 24.4$ ;  $p < 0.0001$ ; for nPVI-V,  $F(5,42) = 12.03$ ;  $p < 0.0001$ ; for rPVI-C,  $F(5,42) = 23.9$ ;  $p < 0.0001$ ; for VarcoV,  $F(5,42) = 2.7$ ;  $p < 0.04$ ; for VarcoC,  $F(5,42) = 10.4$ ;  $p < 0.0001$ ]. Effect size was quite variable, however, largest for %V,  $\Delta C$  and rPVI-C and smallest and rather weak for VarcoV; effect size mostly decreased when the “stress-timed” and “syllable-timed” data were omitted from analysis (see partial  $\eta^2$  values in Table 4).

Pairwise comparisons showed that differences between languages were not consistent across all metrics. As illustrated in Fig. 2a and 2b, for the consonantal metrics  $\Delta C$  and rPVI-C the

scores of English and German were similar to each other and significantly higher than the scores of the other languages, which were lowest for Greek and Italian and intermediate for Korean and Spanish (i.e. English, German > Korean, Spanish > Greek, Italian). VarcoC, on the other hand, showed no differences between the scores of English, German, Italian and Korean (except English > Italian), and only significantly lower scores for Greek and Spanish compared to the other four languages (see Fig. 2c). Thus, the picture across consonantal metrics is not altogether consistent: e.g., while Italian had the second lowest  $\Delta C$  and rPVI-C scores and is classed with Greek by these metrics, its VarcoC score was significantly higher than those of either Greek or Spanish and on a par with Korean and German.

The results from the vocalic metrics were more variable. As illustrated in Fig. 2a, for %V, German and English had lower scores than the other languages, and German %V was significantly lower than English; Greek also showed a significantly lower score than Italian, Korean and Spanish among which there were no differences (i.e. German < English < Greek < Italian, Korean, Spanish). For nPVI-V, however, English showed a higher score than German, the score of which was not significantly different from those of Greek and Korean; Italian and Spanish had significantly lower scores than the other languages (i.e. English > German, Greek, Korean > Italian, Spanish; see Fig. 2b). Far fewer differences were found in pairwise comparisons of VarcoV scores and the overall range of values was very small (52 to 58 points): German had a significantly lower score than Greek and Korean, while the Korean score was also significantly higher than that of Spanish; all other pairwise comparisons did not reach significance (Fig. 2c).

These different effects of language on consonantal and vocalic scores are also reflected in Euclidean distances of language means from English, presented in Table 5. As Euclidean distances show, according to  $\Delta C$ -%V and PVIs, English and German are closer to each other than the other languages, but this does not quite hold for Varcos or for VarcoC-%V which place Italian and Korean respectively closer to English than they place German. For the other languages as well, with the notable exception of Greek which is consistently furthest from English, the *relative* distances are not always stable when one compares across metrics. The Euclidean distances indicate that using two metrics, one to measure consonantal and the other to measure vocalic variability, does not improve consistency and performance. As can be seen in Table 4, this conclusion is corroborated by the effect size of language when vocalic and consonantal metrics are treated as a repeated-measures factor: these effect sizes were not substantially different from those of each metric alone and in some cases they were in fact lower.

**Table 3 Correlations between metrics for the story corpus and (separately) for the “uncontrolled” sentence set; for story, results are based on average speaker values for all languages (POOLED = 48 observations) and for each language separately (8 observations per language); for “uncontrolled” sentences, results are based on individual sentence scores for all languages (POOLED = 240 observations) and for each language separately (40 observations per language); statistically significant correlations (at  $p < 0.016$  for pooled data and  $p < 0.0028$  for within language comparisons, after Bonferroni correction) are shown in bold**

Language	Consonantal metrics						Vocalic metrics					
	Variables		Models for story		Models for “uncontrolled” sentences		Variables		Models for story		Models for “uncontrolled” sentences	
	Y	X	r	Regression eq.	r	Regression eq.	Y	X	r	Regression eq.	r	Regression eq.
English	$\Delta C$	rPVI-C	<b>0.97</b>	<b>y=3.4+0.81x</b>	<b>0.82</b>	<b>y=3.3+0.85x</b>	%V	nPVI-V	0.01	y=43.7+0.007x	0.29	y=33.3+0.19x
German	$\Delta C$	rPVI-C	<b>0.96</b>	<b>y=-6.8+1.04x</b>	<b>0.79</b>	<b>y=8.7+0.82x</b>	%V	nPVI-V	0.03	y=37.1+0.02x	0.19	y=38.1+0.06x
Greek	$\Delta C$	rPVI-C	0.86	y=-1.1+0.88x	<b>0.69</b>	<b>y=14.2+0.50x</b>	%V	nPVI-V	-0.09	y=48.9-0.02x	0.43	y=41.0+0.09x
Italian	$\Delta C$	rPVI-C	<b>0.96</b>	<b>y=6.6+0.74x</b>	<b>0.61</b>	<b>y=26.7+0.37x</b>	%V	nPVI-V	0.19	y=45.7+0.1x	0.10	y=46.1+0.07x
Korean	$\Delta C$	rPVI-C	<b>0.94</b>	<b>y=2.9+0.81x</b>	<b>0.90</b>	<b>y=5.5+0.79x</b>	%V	nPVI-V	0.23	y=37.2+0.2x	0.002	y=49.2+0.001x
Spanish	$\Delta C$	rPVI-C	0.87	y=0.8+0.82x	<b>0.71</b>	<b>y=20.7+0.48x</b>	%V	nPVI-V	-0.4	y=69.6-0.44x	0.36	y=37.6+0.23x
POOLED	$\Delta C$	rPVI-C	<b>0.96</b>	<b>y=-4.6+0.96x</b>	<b>0.80</b>	<b>y=7.9+0.75x</b>	%V	nPVI-V	<b>-0.38</b>	<b>y=61.1-0.29x</b>	0.05	y=45.3+0.02x
English	$\Delta C$	VarcoC	0.53	y=-8.8+1.24x	<b>0.71</b>	<b>y=11.7+0.79x</b>	%V	VarcoV	0.29	y=28.9+0.3x	-0.20	y=49.7-0.10x
German	$\Delta C$	VarcoC	0.79	y=13.8+0.91x	<b>0.87</b>	<b>y=-46.9+2.03x</b>	%V	VarcoV	-0.11	y=40.6-0.05x	0.11	y=39.0+0.04x
Greek	$\Delta C$	VarcoC	0.37	y=38.2+0.05x	<b>0.59</b>	<b>y=13.9+0.54x</b>	%V	VarcoV	-0.49	y=56.6-0.2x	0.28	y=43.4+0.05x
Italian	$\Delta C$	VarcoC	0.41	y=15.4+0.56x	<b>0.54</b>	<b>y=21.6+0.42x</b>	%V	VarcoV	-0.36	y=60.9-0.2x	-0.0007	y=49.2-0.0003x
Korean	$\Delta C$	VarcoC	0.86	y=-13.6+1.1x	<b>0.83</b>	<b>y=-12.2+1.15x</b>	%V	VarcoV	0.55	y=32.9+0.28x	-0.04	y=52.4-0.02x
Spanish	$\Delta C$	VarcoC	<b>0.91</b>	<b>y=-28.1+1.55x</b>	<b>0.90</b>	<b>y=9.1+0.72x</b>	%V	VarcoV	-0.09	y=52.4-0.07x	0.41	y=40.6+0.14x
POOLED	$\Delta C$	VarcoC	<b>0.72</b>	<b>y=-27+1.53x</b>	<b>0.70</b>	<b>y=3.2+0.87x</b>	%V	VarcoV	0.10	y=41.4+0.09x	0.12	y=43.7+0.05x
English	rPVI-C	VarcoC	0.66	y=-31.5+1.84x	0.45	y=34.3+0.49x	nPVI-V	VarcoV	0.38	y=25.5+0.67x	0.46	y=37.7+0.36x
German	rPVI-C	VarcoC	0.66	y=29.1+0.69x	<b>0.67</b>	<b>y=-19.8+1.52x</b>	nPVI-V	VarcoV	0.63	y=27.9+0.51x	<b>0.75</b>	<b>y=4.5+0.93x</b>
Greek	rPVI-C	VarcoC	-0.16	y=57.2-0.21x	0.35	y=26.2+0.43x	nPVI-V	VarcoV	0.77	y=-3.6+0.99x	<b>0.75</b>	<b>y=19.1+0.59x</b>
Italian	rPVI-C	VarcoC	0.50	y=6.3+0.88x	<b>0.66</b>	<b>y=2.7+0.85x</b>	nPVI-V	VarcoV	0.65	y=9.1+0.7x	<b>0.56</b>	<b>y=27.6+0.40x</b>
Korean	rPVI-C	VarcoC	0.74	y=-5.9+1.17x	<b>0.72</b>	<b>y=-5.1+1.13x</b>	nPVI-V	VarcoV	-0.04	y=55.1-0.02x	<b>0.57</b>	<b>y=17.2+0.64x</b>
Spanish	rPVI-C	VarcoC	0.79	y=-13.1+1.41x	<b>0.57</b>	<b>y=18.4+0.67x</b>	nPVI-V	VarcoV	-0.004	y=47.4-0.004x	<b>0.49</b>	<b>y=33.2+0.26x</b>
POOLED	rPVI-C	VarcoC	<b>0.68</b>	<b>y=-16+1.44x</b>	<b>0.58</b>	<b>y=13.5+0.79x</b>	nPVI-V	VarcoC	0.32	y=31.7+0.39x	<b>0.56</b>	<b>y=26.1+0.47x</b>

**Table 4** *Partial  $\eta^2$  for language, elicitation and sentence type effects for each metric separately (top) and for pairs of vocalic and consonantal metrics (bottom); language effect size is presented separately for the ANOVAs on the pooled data and those on the sentence set; values in square brackets represent partial  $\eta^2$  for language and elicitation effects in ANOVAs from which the “stress-timed” and “syllable-timed” sentence sets were excluded*

		%V	$\Delta C$	nPVI-V	rPVI-C	VarcoV	VarcoC
ANOVAs on pooled data	Language	0.77 [0.73]	0.74 [0.72]	0.59 [0.58]	0.74 [0.70]	0.24 [0.19]	0.55 [0.54]
	Elicitation	0.38 [0.37]	0.42 [0.38]	0.43 [0.36]	0.23 [0.27]	0.75 [0.66]	0.56 [0.52]
ANOVAs on sentences	Language	0.70	0.73	0.49	0.65	0.22	0.68
	Sentence Type	0.84	0.69	0.08	0.75	0.36	0.09
		$\Delta C$ -%V	PVIIs		Varcos		VarcoC-%V
ANOVAs on pooled data	Language	0.61 [0.59]	0.76 [0.73]		0.32 [0.18]		0.57 [0.63]
	Elicitation	0.59 [0.68]	0.46 [0.45]		0.79 [0.73]		0.68 [0.66]
ANOVAs on sentences	Language		0.66	0.71	0.37		0.70
	Sentence Type		0.39	0.63	0.31		0.27

**Table 5** *Euclidean distances between English and the other languages in ascending order, separately for each pair of metrics*

$\Delta C$ -%V		PVIIs		Varcos		%V-VarcoC	
German	6.3	German	6.6	Italian	3.3	Korean	3.5
Korean	10.2	Korean	13.5	German	3.5	German	6.0
Spanish	14.0	Spanish	18.7	Korean	3.5	Spanish	6.1
Italian	17.3	Italian	22.7	Spanish	5.1	Italian	6.3
Greek	19.1	Greek	23.0	Greek	8.6	Greek	8.6

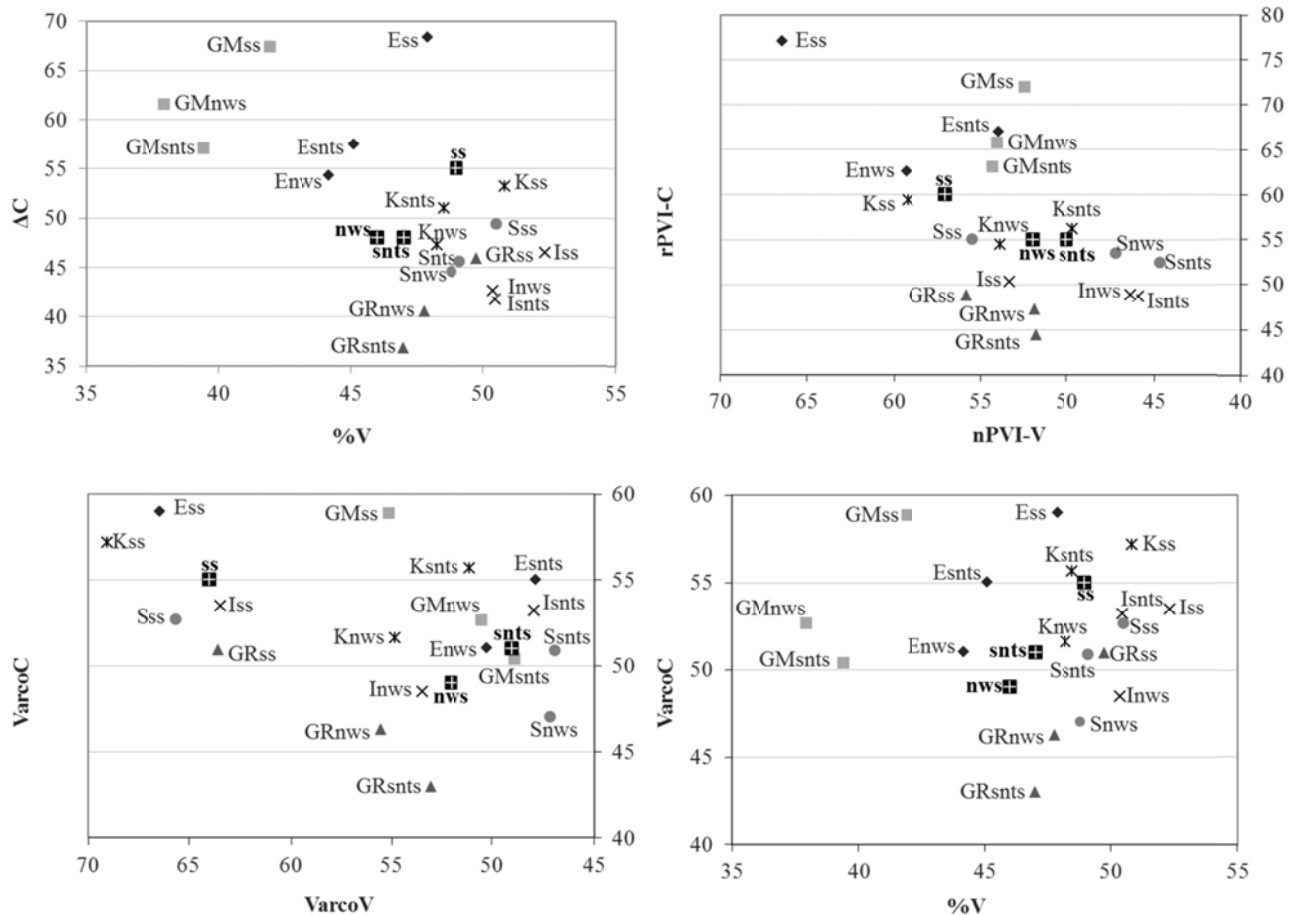
### 3.3. Elicitation effects: within language comparisons

Elicitation affected all metrics [for %V,  $F(2,84) = 25.8$ ;  $p < 0.0001$ ; for  $\Delta C$ ,  $F(2,84) = 30$ ;  $p < 0.0001$ ; for nPVI-V,  $F(2,84) = 31.8$ ;  $p < 0.0001$ ; for rPVI-C,  $F(2,84) = 12.3$ ;  $p < 0.0001$ ; for VarcoV,  $F(2,84) = 124.6$ ;  $p < 0.0001$ ; for VarcoC,  $F(2,84) = 54.3$ ;  $p < 0.0001$ ]. As partial  $\eta^2$  values show, the effect was modest to large for all metrics and often comparable in size to the effect of language; as shown in Table 4, effect size was not substantially affected by the exclusion of the “stress-timed” and “syllable-timed” sentences.

Pairwise comparisons between elicitation levels indicate that metrics were similarly affected by elicitation: in all cases, values from spontaneous data were significantly higher than those from the two spoken corpora (sentences and story). For %V,  $\Delta C$  and rPVI-C this was the only difference between elicitation levels, i.e. there were no statistically significant differences between sentences and story (for pooled means see Appendix C). VarcoV and nPVI-V, on the other hand, showed significant differences between all levels, with sentences having the lowest scores, spontaneous speech the highest and story showing intermediate values. For VarcoC as well, all pairwise comparisons were significant, but in this case, the story score was lower than that of sentences.

In addition, all metrics except %V and rPVI-C, showed an interaction between language and elicitation [for %V,  $F(10,84) < 1$ ; for  $\Delta C$ ,  $F(10,84) = 2.2$ ;  $p < 0.03$ ; for nPVI-V,  $F(10,84) = 3.1$ ;  $p < 0.002$ ; for rPVI-C,  $F(10,84) = 1.8$ ; *n.s.*; for VarcoV,  $F(2,84) = 3.3$ ;  $p < 0.001$ ; for VarcoC,  $F(10,84) = 4.5$ ;  $p < 0.0001$ ; see Appendix C for language means and standard errors separately for each elicitation level]. What these interactions suggest is that elicitation method did not affect all languages equally (see Fig. 3). Fischer LSD tests confirmed that this was the case: for example, for  $\Delta C$ , English, German and Greek followed the general pattern

(similar scores for sentences and story and significantly higher scores for spontaneous speech), but for Italian, Korean and Spanish, the sentences and spontaneous speech scores did not show statistically significant differences (see Fig. 3a). For nPVI-V the differences between sentences and story reported above on pooled means appear to be driven mainly by English (see Fig. 3b). Similarly, for VarcoV only Italian showed the difference between sentences and story reported above, while for VarcoC, Italian, Korean and Spanish showed no significant differences between sentences and spontaneous speech (see Fig. 3c). These inconsistencies across languages are also reflected in the Euclidean distances in Table 6. For example, in the  $\Delta C$ - $\%V$  space, the distances of the Korean story and spontaneous data from sentences were comparable, but for PVI-V and Varcos, spontaneous speech was much more distant from sentences than story was; for VarcoC- $\%V$ , on the other hand, the reverse pattern holds.



**Fig. 3** Mean metric scores of each language separately for each elicitation method; symbols in black represent mean scores per elicitation method pooled over all languages;  $\Delta C$ - $\%V$  in panel (a), PVI-V in panel (b), Varcos in panel (c) and VarcoC- $\%V$  in panel (d); ss = spontaneous speech, nws = The North Wind and the Sun, snts = sentences; E = English, G = German, GR = Greek, I = Italian, K = Korean, S = Spanish; note that the values in the x-axis of panels (b) and (c) are presented in reverse order to facilitate comparison with panels (a) and (d)

### 3.4. Elicitation effects: across language comparisons

The aim of the present experiment was not only to examine what effect elicitation would have on the scores of each language but also whether such effects could be large enough to alter the relationship between the scores of different languages, since it is often the case that scores from studies in which data were collected in different ways are compared to each other. To this purpose, pairwise comparisons on the interaction of language and

elicitation were examined to see whether the language differences discussed in section 3.2 above, hold for each metric within each level of elicitation and across elicitation levels.

**Table 6** *Euclidean distances of scores for spontaneous speech and story calculated from the sentence scores of each language*

	$\Delta C$ -%V		PVI <sub>s</sub>		Varcos		VarcoC-%V	
	spontaneous speech	story	spontaneous speech	story	spontaneous speech	story	spontaneous speech	story
English	11.2	3.3	16.0	6.9	19.0	4.6	4.8	4.1
German	10.7	4.8	9.1	2.7	10.5	2.8	8.8	2.7
Greek	9.5	3.8	5.9	2.8	13.2	4.1	8.4	3.4
Italian	5.0	0.8	7.6	0.5	15.5	7.3	1.9	4.7
Korean	3.2	3.8	10.1	4.6	18.0	5.5	2.8	4.1
Spanish	4.1	1.0	11.1	2.7	18.8	3.9	2.3	3.9

For %V and rPVI-C in particular, which did not show a two-way interaction, it is assumed that the differences across languages reported in section 3.2 hold for all three elicitation methods. For the other metrics, indicative results from the pairwise comparisons are given. For  $\Delta C$ , pairwise comparisons show that the English sentence corpus has a higher score than Greek, Italian and Spanish (as in the pooled data), but there is no difference between English and Korean  $\Delta C$ ; the same applies to the comparison of the English and Korean story scores, and the German and Korean sentence scores (see Fig. 3a). Similarly, pairwise comparisons for nPVI-V show that the significantly higher score of English with respect to German holds for the spontaneous data but not for the other two elicitation methods. Fischer LSD tests showed virtually no cross-linguistic differences in VarcoV scores, except for German spontaneous speech for which VarcoV was significantly lower than the equivalent scores of all the other languages in the corpus (see Fig. 3c). Finally, for VarcoC, the pooled data show a significantly higher score for English than Korean, Italian and Spanish, but pairwise comparisons within elicitation level show that the difference holds only for English vs. Korean and Spanish spontaneous speech; all other comparisons do not reach significance (see Fig. 3c/d).

Similar inconsistencies are present when languages are compared across elicitation levels. For example, sentence-based German VarcoC is comparable to that of the Greek and Spanish spontaneous speech and the Spanish sentences, although the pooled data show Greek and Spanish to have significantly lower VarcoC than German (see Fig. 3c/d).

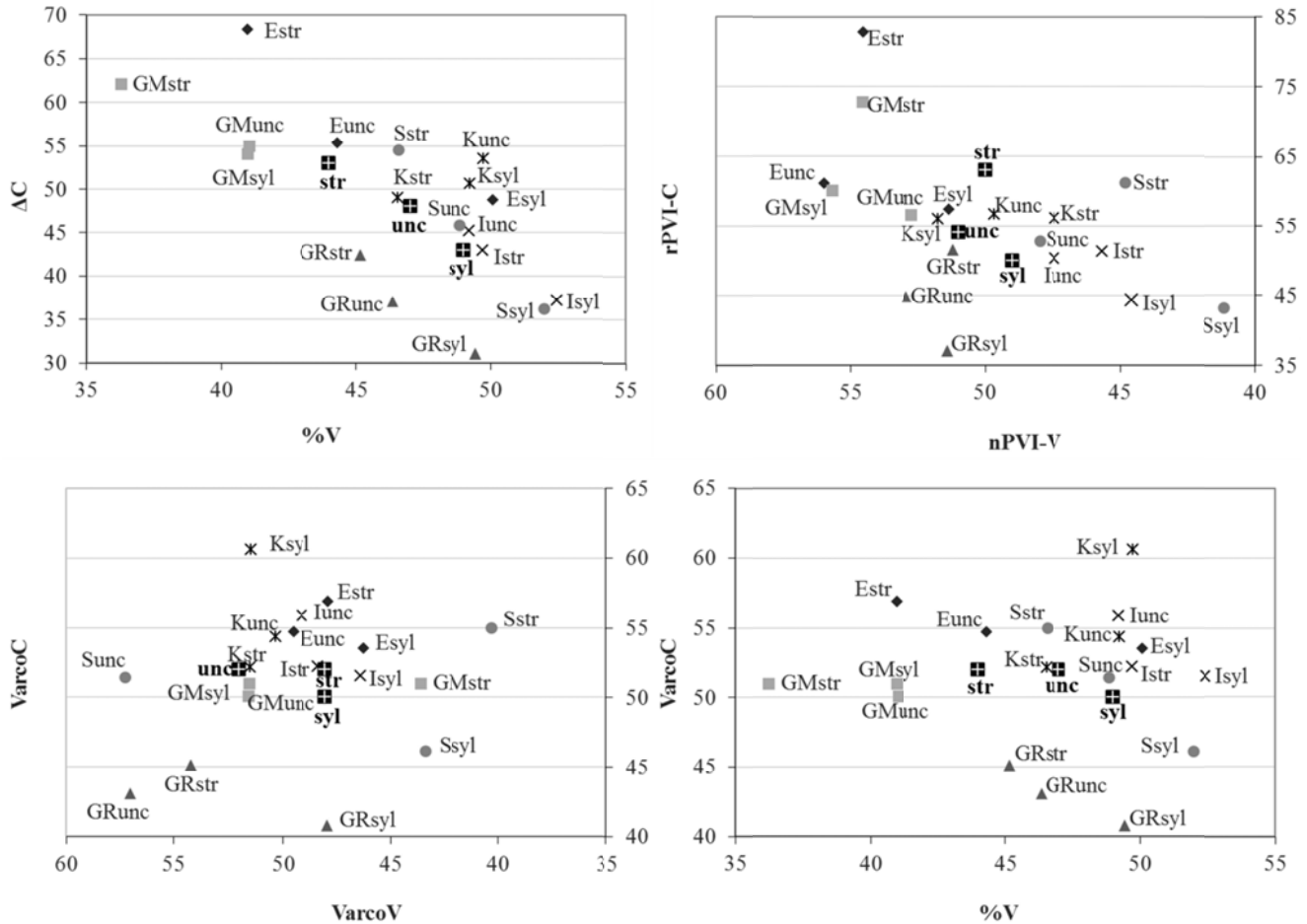
### 3.5. Sentence type effects: within language comparisons

In the sentence corpus, ANOVAs on all metrics except VarcoV showed a main effect of language: [for %V,  $F(5,42) = 19.4$ ;  $p < 0.0001$ ; for  $\Delta C$ ,  $F(5,42) = 22.7$ ;  $p < 0.0001$ ; for nPVI-V,  $F(5,42) = 8.0$ ;  $p < 0.0001$ ; for rPVI-C,  $F(5,42) = 15.7$ ;  $p < 0.0001$ ; for VarcoV,  $F(5,42) = 2.4$ ; *n.s.*; for VarcoC,  $F(5,42) = 4.0$ ;  $p < 0.02$ ; see Appendix C for language means and standard errors separately for each sentence type]. This effect is comparable to the language effect in the entire corpus presented in section 3.2 and is not discussed further (but see Table 4 for effect sizes).

All metrics also showed a main effect of sentence type: [for %V,  $F(2,84) = 224.4$ ;  $p < 0.0001$ ; for  $\Delta C$ ,  $F(2,84) = 91.8$ ;  $p < 0.0001$ ; for nPVI-V,  $F(2,84) = 3.5$ ;  $p < 0.03$ ; for rPVI-C,  $F(2,84) = 123.1$ ;  $p < 0.0001$ ; for VarcoV,  $F(2,84) = 23.5$ ;  $p < 0.0001$ ; for VarcoC,  $F(2,84) = 4.0$ ;  $p < 0.02$ ; see Appendix C for pooled means]. As shown in Table 4, the sentence effect was substantial for all metrics except nPVI-V and VarcoC, and in most cases larger than the language effect for the same set of data. These effect sizes indicate that variability within a language is as high as variability across languages.

Pairwise comparisons on the sentence type effect showed that the results conformed to the prediction that “stress-timed” sets would have higher scores than “syllable-timed” sets (with the reverse applying to %V), and that the “uncontrolled” sets would have intermediate scores.  $\Delta C$ , rPVI-C and %V all conform to this pattern. On the other hand, nPVI-V and VarcoC showed no difference between “stress-timed” sentences and the other

two sets, but did show significantly lower scores for “syllable-timed” sentences than “uncontrolled” sentences. Finally, VarcoV showed no differences between “stress-timed” and “syllable-timed” sentences but significantly higher scores for the “uncontrolled” set compared to the other two.



**Fig. 4** Mean metric scores of each language separately for each set of sentences; symbols in black represent mean scores per sentence type pooled over all languages;  $\Delta C$ - $\%V$  in panel (a), PVI in panel (b), Varcos in panel (c) and VarcoC- $\%V$  in panel (d); str = “stress-timed”, syl = “syllable-timed”, unc = “uncontrolled; language abbreviations as in Fig. 3; note that the values in the x-axis of panels (b) and (c) are presented in reverse order to facilitate comparison with panels (a) and (d)

In addition, all metrics showed an interaction between sentence type and language which suggests that, as with elicitation, the effects of sentence type were not consistent across languages and metrics [for  $\%V$ ,  $F(10,84) = 13.5$ ;  $p < 0.00001$ ; for  $\Delta C$ ,  $F(10,84) = 91.8$ ;  $p < 0.00001$ ; for nPVI-V,  $F(10,84) = 2.8$ ;  $p < 0.005$ ; for rPVI-C,  $F(10,84) = 13.7$ ;  $p < 0.00001$ ; for VarcoV,  $F(10,84) = 8.8$ ;  $p < 0.00001$ ; for VarcoC,  $F(10,84) = 4.0$ ;  $p < 0.02$ ]. The general patterns of higher scores for “stress-timed” than syllable-timed” sentences (lower for  $\%V$ ) largely holds however: thus,  $\%V$  showed highest scores for “syllable-timed” sentences and lowest for “stress-timed” sentences for all languages (see Fig. 4a); for nPVI-V, “syllable-timed” sentences had lower scores than “stress-timed” sentences in all languages except Korean and Spanish, while for VarcoV the effect was present in Greek and German (see Fig. 4b and 4c respectively). For the consonantal metrics, the difference between “stress-timed” and “syllable-timed” data mostly held as well: it applied to all languages with respect to  $\Delta C$ , to all languages except Korean for rPVI-C and to all languages except German and Italian for VarcoC (see Fig. 4a, 4b

and 4c/d respectively). “Uncontrolled” sentences generally showed intermediate scores, but the results for pairwise comparisons with “stress-timed” and “syllable-timed” sentences were far less consistent across languages and metrics and are not reported further.

The effects of sentence type are also reflected in the Euclidean distances of the “stress-timed” and “syllable-timed” sets from the “uncontrolled” set shown in Table 7. As with elicitation, relative distances were not consistent across metrics; e.g., while  $\Delta C$ -%V and PVI<sub>s</sub> suggest that Greek “stress-timed” sentences and “syllable-timed” sentences are approximately equidistant from the “uncontrolled” set, Varcos show a comparatively smaller distance for the former than the latter; in Korean, on the other hand, distances are comparable within  $\Delta C$ -%V and PVI<sub>s</sub>, but Varcos and VarcoC-%V show a much larger distance for the “syllable-timed” than the “stress-timed” set.

### 3.6 Sentence type effects: across language comparisons

As mentioned, one of the aims of the study was to examine whether the variability inherent in each language could lead to metric scores being affected by the choice of materials and by so doing obscure or exaggerate differences between languages. In order to test for such a possibility, pairwise comparisons were made between languages both within and across sentence sets. These comparisons confirmed that, depending on the syllable composition of the sentences used to calculate metrics, differences between languages can indeed be reduced or exaggerated.

**Table 7 Euclidean distances, for each sentence type calculated from “uncontrolled” sentences**

	$\Delta C$ -%V		PVI <sub>s</sub>		Varcos		VarcoC-%V	
	“stress-timed”	“syllable-timed”	“stress-timed”	“syllable-timed”	“stress-timed”	“syllable-timed”	“stress-timed”	“syllable-timed”
English	13.4	8.7	21.9	5.9	2.7	3.5	4.0	5.9
German	8.6	0.9	16.4	4.6	8.0	0.1	4.8	0.1
Greek	5.5	6.8	6.9	7.9	3.5	9.4	2.3	3.9
Italian	2.4	8.7	2.1	6.5	3.7	5.1	3.7	5.4
Korean	3.2	2.9	2.3	2.2	2.5	6.4	3.5	6.3
Spanish	9.0	10.2	8.9	11.7	17.4	14.9	4.2	6.1

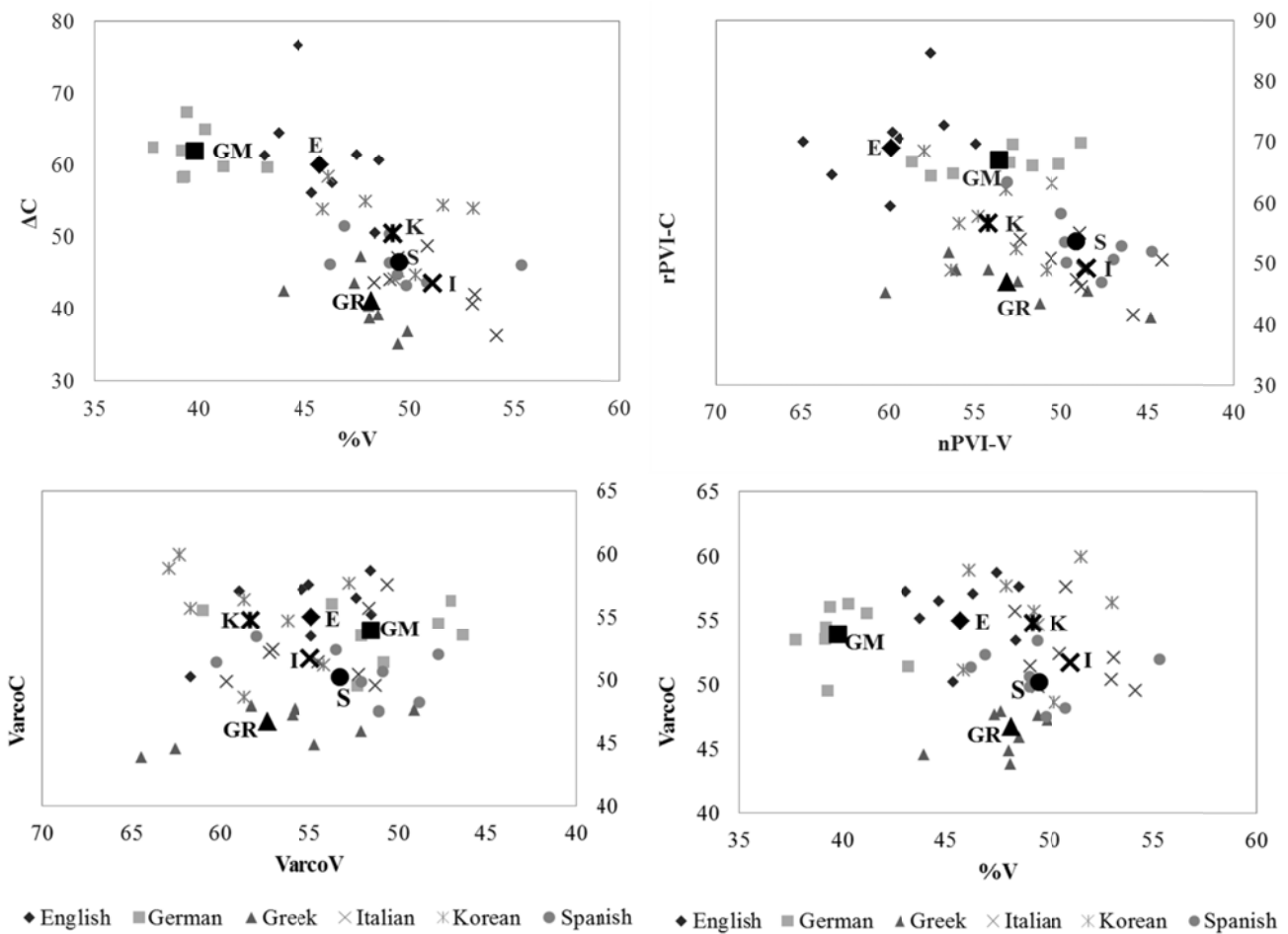
As can be seen in Fig. 4, the general patterns of differences between languages remain in place when the data are broken down by sentence set. Across sets, however, we find that this no longer holds; e.g., while the English  $\Delta C$  from the “stress-timed” set was higher than the Korean  $\Delta C$  of all sentence sets, the English “syllable-timed”  $\Delta C$  was not significantly different from any of the Korean  $\Delta C$  scores. Similar patterns are evident for %V: e.g., the English %V was significantly lower than that of the other languages (except German) only in the “stress-timed” set; no significant differences between English %V for the “syllable-timed” set and those of Greek, Italian, Korean or Spanish were found, while for the “uncontrolled” sentences, the English %V was not different from that of either German or Greek. Similarly, while pooled nPVI-V showed a significantly lower score for English than all the other languages, such a difference was not found between “stress-timed” English and “stress-timed” German or Greek, between “syllable-timed” English and Greek or Korean, or between “uncontrolled” English and German or Greek (see Fig. 4b). Finally, pooled VarcoV results showed a significantly lower score for German than all the other languages; the pairwise comparisons indicate that this was due to the “stress-timed” set only for which German VarcoV was significantly lower than that of Greek and Korean; the German VarcoV for the other two sentence sets did not differ significantly from the VarcoVs of any of the other languages except Spanish for the “syllable-timed” set (see Fig. 4c).

### 3.6 Inter-speaker variation

Since many studies that rely on rhythm metrics are based on small numbers of speakers, it was important to investigate the sensitivity of metrics to individual speaker variation and the extent to which such differences could affect comparisons across studies. Inter-speaker variation can be easily detected in Fig. 5 which plots average individual speaker scores together with pooled language scores (shown in black) for comparison. As is

evident, the speakers of each language do not form a discernible cluster, except possibly for the German and Greek speakers in the  $\Delta C$ -%V and VarcoC-%V plots respectively. In all other cases, the data of individual speakers of different languages are intermingled.

This great spread of values is reflected in the Euclidean distances in Table 8: while some speakers are very close to the average score for their language (e.g.,  $\Delta C$ -%V and PVI scores for Greek Sp3 or Spanish Sp2), others deviate markedly (e.g., English Sp8 according to Varcos and VarcoC-%V, and Korean Sp8 according to  $\Delta C$ -%V and PVIs). Crucially, however, the differences are not consistent across metrics: thus, although English Sp7 is the one closest to the English average, as defined by VarcoC-%V, she is the most distant from the mean in the  $\Delta C$ -%V and PVI spaces. Similarly, although according to Varcos and VarcoC-%V, Korean Sp6 could be seen as an outlier, her  $\Delta C$ -%V and PVI scores make her minimally different from the Korean mean. Overall, if one compares the minimum and maximum distances for each speaker within the data of a language, it is clear that these minima and maxima rarely coincide across metric sets.



**Fig. 5** Speaker average scores (pooled over elicitation methods) separately for each language; symbols in black represent the average score of each language;  $\Delta C$ -%V in panel (a), PVIs in panel (b), Varcos in panel (c) and VarcoC-%V in panel (d); note that the values in the x-axis of panels (b) and (c) are presented in reverse order to facilitate comparison with panels (a) and (d)

**Table 8 Euclidean distances for individual speakers, calculated from the mean of each language; minimum distances within each metric set and language are in light grey cells; maximum distances are shown in bold**

Language	Metric	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6	Sp7	Sp8
English	$\Delta C$ -%V	2.9	2.9	2.6	9.9	4.9	2.3	<b>16.6</b>	4.0
	PVIs	2.6	1.6	5.1	9.6	4.9	5.0	<b>15.8</b>	5.6
	Varcos	2.3	2.6	4.6	1.6	3.4	5.0	2.9	<b>8.3</b>
	VarcoC-%V	3.4	3.8	2.1	3.1	2.0	4.1	1.8	<b>4.8</b>
German	$\Delta C$ -%V	4.2	3.7	3.8	2.6	<b>5.4</b>	3.0	2.0	0.6
	PVIs	4.7	3.4	2.1	5.0	2.7	3.6	0.7	<b>5.6</b>
	Varcos	2.7	4.5	3.8	<b>9.6</b>	3.0	5.1	0.7	5.2
	VarcoC-%V	4.3	<b>4.5</b>	0.8	2.1	2.1	2.4	2.0	0.7
Greek	$\Delta C$ -%V	<b>6.3</b>	4.6	0.6	6.2	2.7	1.9	2.4	4.4
	PVIs	6.0	4.0	0.7	<b>10.2</b>	3.6	4.9	7.2	2.4
	Varcos	1.5	1.5	3.3	<b>8.4</b>	1.8	5.4	7.7	5.5
	VarcoC-%V	1.3	1.8	1.8	1.6	1.2	0.9	2.9	<b>4.7</b>
Italian	$\Delta C$ -%V	<b>8.0</b>	2.0	2.7	2.7	3.5	5.1	3.9	2.2
	PVIs	<b>8.2</b>	0.2	2.1	4.5	3.2	5.6	6.0	2.5
	Varcos	4.3	0.6	2.3	5.2	3.1	<b>7.3</b>	5.1	2.2
	VarcoC-%V	2.1	0.8	1.5	2.3	1.8	<b>2.7</b>	2.2	1.5
Korean	$\Delta C$ -%V	6.3	5.2	5.8	5.2	4.7	4.5	4.6	<b>8.4</b>
	PVIs	8.1	8.5	4.6	1.7	7.4	1.1	5.5	<b>12.3</b>
	Varcos	3.4	2.1	6.2	1.6	5.6	<b>6.5</b>	6.3	6.1
	VarcoC-%V	1.8	1.4	<b>2.5</b>	1.3	2.4	<b>2.5</b>	<b>2.5</b>	<b>2.5</b>
Spanish	$\Delta C$ -%V	3.1	0.4	<b>5.9</b>	5.6	3.3	3.9	1.7	3.3
	PVIs	3.7	0.6	4.7	<b>10.5</b>	6.9	4.6	3.6	2.7
	Varcos	4.9	2.4	5.8	2.2	3.4	1.2	5.7	<b>7.1</b>
	VarcoC-%V	2.2	1.6	2.4	1.5	1.9	1.1	2.4	<b>2.7</b>

#### 4.0 Discussion

By and large the results supported the study’s predictions: methodological choices had a substantial impact on metric scores. Scores showed considerable differences between read and spontaneous speech, while the syllable complexity of the materials significantly affected scores independently of rhythm class affiliation. In addition, metrics showed extensive inter-speaker variability. Overall, then, the present results suggest that metrics are very sensitive to inevitable “noise” in the data.

The study’s goal, however, was not simply to probe the sensitivity of metrics, but also to see if any discrepancies can be consistently attributed to specific experimental manipulations and thus effectively constrained. The results indicate that variability due to inter-speaker differences, elicitation method or syllable complexity is difficult if not impossible to constrain, because the effects are not consistent across metrics and languages. As an example, while spontaneous speech increased scores (compared to sentences) for most metrics and languages, effects were not present for Korean and Spanish for either  $\Delta C$  or VarcoC. Similarly, while in all languages %V significantly increased between the two read corpora and spontaneous speech, nPVI-V showed minimal effects that differed by language: there was no effect for German and no differences between sentences and spontaneous speech for Greek, Italian, Korean and Spanish, but significant differences between all three elicitation types for English. Similar inconsistencies emerged with respect to sentence type: e.g., while German %V was lower in “stress-timed” than “syllable-timed” sentences, nPVI-V showed no difference between the two sets, and VarcoV showed a significantly *lower* score for the “stress-timed” sentences. These discrepancies were replicated in the correlations between metrics for the same dataset, which show that although in some cases the values of one metric strongly correlate with those of another, this does not apply either to all metrics or to all languages for a given comparison. Finally, the inter-speaker variability was also inconsistent across metrics, thereby rendering futile any attempts to remove outliers. The participant who is an outlier according to one

metric can very well be close to the mean according to another, as the Euclidean distances in Table 8 demonstrate.

It would be tempting to attribute these discrepancies to external factors. For instance, one could argue that the reason why Korean was more resistant to the elicitation manipulation was that Korean speakers adopted a similar speaking style in all tasks. Similarly, one could argue that the syllable complexity was not as successfully manipulated in the Korean dataset as, say, in the Spanish set. Such arguments would be valid, if there were no strong inconsistencies between metric scores for the same data of each language. Thus, while the  $\Delta C$  and VarcoC Korean scores were largely unaffected by elicitation, %V, PVI and VarcoV were significantly higher in spontaneous speech than the two read styles, as in the other languages. Similarly, while rPVI-C showed no sentence effect for Korean,  $\Delta C$  and VarcoC did show the expected differences.

**Table 9 Mean metric scores for all languages in the present study (in bold); results from other studies are presented for comparison**

Language	Study <sup>a</sup>	Scores					
		$\Delta C$	%V	rPVI-C	nPVI-V	VarcoC	VarcoV
English	Ramus et al. 1999	53.5	40.1				
	Grabe & Low 2002	56.7	41.1	64.1	57.2		
	Dellwo & Wagner 2003 <sup>b</sup>	55.7	42.0				
	White & Mattys 2007	59.0	38.0	70.0	73.0	47.0	64.0
	<b>present study</b>	<b>60.0</b>	<b>45.7</b>	<b>68.9</b>	<b>59.9</b>	<b>55.0</b>	<b>54.8</b>
German	Grabe & Low 2002	52.6	46.4	55.3	59.7		
	Dellwo & Wagner 2003	71.7	42.8				
	Russo & Barry 2008 <sup>c</sup>	65.0	41.7	68.7	52.5		
	<b>present study</b>	<b>62.0</b>	<b>39.8</b>	<b>67.0</b>	<b>53.6</b>	<b>54.0</b>	<b>51.5</b>
Greek	Grabe & Low 2002	52.7	44.1	59.6	48.7		
	Tsiartsioni 2003			48.4	47.8		
	Baltazani 2007			68.0	45.0		
	<b>present study</b>	<b>41.1</b>	<b>48.2</b>	<b>46.9</b>	<b>53.2</b>	<b>46.8</b>	<b>57.4</b>
Italian	Ramus et al. 1999	48.1	45.2				
	<b>present study</b>	<b>43.6</b>	<b>51.1</b>	<b>49.3</b>	<b>48.5</b>	<b>51.7</b>	<b>55.0</b>
Korean	Jeon 2006	36.0	53.	41.5	49.2	49.5	46.2
	Mok & Lee 2008 <sup>d</sup>	53.2	54.9	57.9	61.2	59.0	61.8
	<b>present study</b>	<b>50.5</b>	<b>49.2</b>	<b>56.7</b>	<b>54.3</b>	<b>54.8</b>	<b>58.3</b>
Spanish	Ramus et al. 1999	47.4	43.8				
	Grabe & Low 2002	47.5	50.8	57.7	29.7		
	White & Mattys 2007	40.0	48.0	43.0	36.0	46.0	41.0
	<b>present study</b>	<b>46.6</b>	<b>49.5</b>	<b>53.7</b>	<b>49.1</b>	<b>50.2</b>	<b>53.3</b>

<sup>a</sup> For accuracy, the table is limited to studies that present scores in tables rather than figures.

<sup>b</sup> The values from Dellwo & Wagner (2003) are those obtained from normal speaking rate.

<sup>c</sup> The values from Russo & Barry (2008) are for medium speaking rate, defined on the basis of phones/sec.

<sup>d</sup> The values for Mok & Lee (2008) are averages of their scores for spontaneous speech and story reading. I am grateful to Peggy Mok for making these data available to me.

Such discrepancies are not unique to the present study. For example, Benton et al. (2007) reported lower and more uniform scores for both American English and Mandarin data elicited from news broadcasting than from spontaneous speech. Mok & Lee (2008), who compared readings of *The North Wind and the Sun* with semi-spontaneous retelling of the story by the same Korean speakers, found that the latter data had generally higher scores than the former. This trend could explain the much higher values reported by Mok & Lee compared to Jeon (2006), who relied on a small sentence corpus. Similarly, both Prieto et al. (in press), who

used a method similar to the present one to sample sentences, and Wiget et al. (2010), who sampled sentences randomly, found that the choice of sentences on which metrics are calculated can have a large albeit inconsistent impact on scores. Specifically, Wiget et al. report %V, VarcoV and nPVI-V scores for five British sentences and show that these do not correlate with one another: the sentence whose %V score is practically the same as the average %V for the set had the highest nPVI-V score (approximately 7 points above average) and at the same time the second lowest VarcoV score (approximately 4 points below average). Finally Renwick (2011) found that %V in English, Dutch, Spanish, Italian and Japanese correlates strongly with syllable structure, particularly with the presence of coda consonants in a sample, independently of rhythm class. Given these reports, it is perhaps unsurprising that some studies such as Ramus et al. (1999) have yielded results that are consistent with the idea of rhythm classes; there is sufficient variability in metric scores that they will occasionally or for some languages yield results in the expected direction, but such results do not appear to be readily replicable.

Overall, the random inconsistencies documented here suggest that any differences in the variability of consonantal and vocalic intervals captured by metrics are largely opaque. This should be hardly surprising given the many factors that influence durational variability in speech (see, e.g., Arvaniti, 2009, for a discussion). Nevertheless, some light may be shed on this issue by considering additional differences between the metrics examined here. Specifically, results from %V,  $\Delta C$  and rPVI-C, the metrics that do not normalize for speaking rate, were more consistent and showed a bigger language effect size than the three metrics that normalize for speaking rate, Varcos and nPVI-V (see Table 4). The need to control for speaking rate in rhythm studies was first argued for by Ramus et al. (1999), and differences in speaking rate across studies have been taken to be the cause of discrepancies among metrics (Ramus, 2002). However, the fact that normalized metrics are less sensitive to cross-linguistic differences as well suggests that what metrics measure is, to a large extent, the effect of speaking rate on the durational variability of segments. This conclusion is supported by several types of evidence. For instance, Loukina et al. (2009) found that adding speaking rate to their classifiers dramatically enhanced the ability of metrics to discriminate between languages (which was generally equally low for comparisons within and across rhythm classes). Similar results are reported by Brimhall, Horton & Morgan (2010) and Horton & Arvaniti (2012): both studies found that %V,  $\Delta C$  and rPVI-C yield more robust classification in both supervised learning using Naïve Bayes classifiers and in unsupervised clustering, while Horton & Arvaniti (2012) also show that scores from these metrics correlate much more strongly with tempo than those of Varcos and nPVI-V. From a production perspective, support is also found in the results of Dellwo & Wagner (2003) and Russo & Barry (2008), who calculated metrics separately for different speaking rates and report that scores go down as speaking rate increases. Note also that many languages classified as syllable-timed are spoken faster than typical stress-timed languages (see e.g., Dauer, 1983, Dellwo & Wagner, 2003, and Arvaniti, 2009, for some examples). For this reason, the contribution of tempo cannot be easily factored out of metric scores. For example, the suggestion of Ramus (2002) to make cross-linguistic comparisons only with data that share the same tempo is unrealistic: as Dellwo & Wagner (2003) note, in order for data of French and English to be comparable in terms of speaking rate, the English speakers must speak at what is for them a normal rate but French speakers must speak at what they would consider a slow rate.

The opacity discussed above does not seem to be the same for vocalic and consonantal metrics. Perhaps the most consistent finding of the present study was that consonantal scores were more regularly affected by sentence type and elicitation, correlated better with each other and showed more robust differences between languages (or, at least, between English and German on the other hand, and the rest of the languages on the other). One plausible explanation is that the relationship between variability in syllable structure, in particular in the types of consonantal clustering a language allows, is more straightforwardly reflected in duration and that consonantal metrics capture these differences relatively efficiently. On the other hand, it is clear that no such straightforward relationship exists between vocalic variability and reduction, the two effects that vocalic metrics are meant to capture. This lack of correlation between metrics meant to capture different aspects of vowel realization has been noted by Ramus et al. (1999), Barry & Andreeva (2001), Grabe & Low (2002), Bary et al. (2003) and Lin & Wang (2007) many of whom have reached similar conclusions those presented here (Ramus et al., 1999; Barry & Andreeva, 2001; Barry et al., 2003).

Given the sensitivity of metrics to various methodological choices and the inconsistent ways in which metrics are affected by these extemporaneous factors, the question that arises is whether metrics can be used to

rhythmically classify languages. The metric scores of the present pooled data do show a visual separation of German and English, on the one hand, and Spanish, Italian, Greek and Korean, on the other (see Fig. 5). Statistically, however, the differences do not hold for all metrics and, again, they are not consistent, even when prototypical examples of each rhythm class are compared. As shown in section 3.2, e.g., Italian and Spanish %V were significantly higher than both English and German %V, as predicted by rhythm class, but their VarcoV scores were comparable to those of both English and German. Once more, the results are not unique: as noted in the introduction, Grabe & Low (2002a) found that PVIs and  $\Delta C$ -%V classified several languages, including Thai, Greek and Japanese, in different ways, while White & Mattys (2007) found that in many instances metric score differences between English and French failed to reach statistical significance in their study. As a result of these trends, metrics do not always classify languages in the same fashion and this applies to the present study as well. For example, although Spanish and Italian appear more stress-timed than Korean and Greek in the rhythm space defined by PVIs, no such separation is possible in the  $\Delta C$ -%V space (cf. Figs. 6a and 6b). Similarly, although German appears more stress-timed than English in the  $\Delta C$ -%V rhythmic space, the opposite relationship obtains if one relies on PVIs (cf. Figs. 6a and 6b).

The classification problem was most notable for the two languages in the present study that have not been consistently classified in the past, Greek and Korean. Korean had very high Varco scores, a result that should unequivocally class it as stress-timed; but according to  $\Delta C$ -%V and PVIs, although Korean is closer to English than Spanish, Italian or Greek are, it is much closer to these three languages than to English or German (see Table 5). *Mutatis mutandis*, similar problems are present for Greek. Greek would be classified as syllable-timed on the basis of the present study but other studies present a very different picture. As can be seen in Table 9, rPVI-Cs show large variation in Greek, with scores having a range of 21.1 points across studies. The nPVI-V scores also show variation, albeit on a smaller scale (a range of 8.2 points across studies). A corollary of these problems is that Korean cannot be unambiguously classified for rhythm on the basis of the present results, while Greek can be classed as syllable-timed only if previous results are ignored. Thus, the present study clearly shows that the classification problems encountered in earlier work were not the outcome of limited speech samples or a small number of speakers: having a large sample elicited in different ways from a large number of speakers does not guarantee more stable metric results or a clear classification by rhythm class.

It is possible that such disparity among studies – at least of Korean – reflects a genuinely greater difficulty in classifying this language, perhaps because its prosodic system is changing. Recall, e.g., that there is disagreement regarding the vowel quantity contrast in Seoul Korean (cf. Jun, 2005, and Yoshida et al., 2007) and that a previous study suggests there is a rhythm change across generations (Lee et al., 1994). However, as Table 9 amply demonstrates, agreement between studies is not much greater for English, German, Italian and Spanish, which are said to be prototypes of stress- and syllable-timing.

All in all, this study has plainly shown that metric scores can differ quite substantially both within and across studies and metrics, even when exemplars of each rhythm class are examined. In turn this means that metric scores cannot be seen as fixed and immutable, reflecting some quintessential property of each language, equivalent, say, to its word order or its tolerance for onsetless syllables. Rather, metric scores from any given language are distributed over a range of values and this distribution can be quite wide. The present study uncovered some of the reasons that affect it: spontaneous speech is more variable in its timing patterns, as are utterances with more complex syllable structures; in addition speakers differ considerably from each other, possibly because of choices such as speaking rate and the clarity of their speech (which could affect the degree of vowel reduction and the duration of consonant clusters). Although evidently more needs to be done if the whole gamut of variability is to be documented, it is clear than comparing metric scores across studies, metrics and languages, much as one would compare shoe sizes, is inadvisable.

The problems discussed above have serious implications for the practice of using metrics to categorize languages for rhythm class, especially languages with intermediate metric scores, the classification of which can be easily swayed by methodological choices. Comparing a pooled metric score of the language under study to some norm, such as a previously published score for English (as is often done), can lead to dramatic differences in classification, depending on the norm used for comparison, the metric(s) chosen and the point in the whole distribution of a language's scores that the pooled mean represents. Crucially, such fluctuation does not merely

affect a language's typological classification into one rhythm class or another, but has important repercussions for how the acquisition and processing of a language are further studied and understood.

The fact that a language can be classified as stress- or syllable-timed because of the impact that methodology can have on metric scores points to the problems associated with the lack of an independent measure that could be used to compare the validity of metrics. This problem is reflected in the present study as well: some metrics, such as rPVI-C, were less affected by elicitation method, while others, such as nPVI-V and VarcoC, were less affected by sentence type. However, in the absence of independent criteria, it is impossible to tell whether the smaller effect sizes found in these cases were due to the robustness of these metrics to external manipulations or to their lack of sensitivity. Opacity makes it impossible to surmise why the same metric may show a substantial effect size for sentence type but not for elicitation (or vice versa), while the lack of independent validation makes it difficult to distinguish between robustness and insensitivity.

Nevertheless, the assumption that cross-linguistic differences exist has been an axiom in the metrics-related literature. For example, in contrast to the present study, in which vocalic metrics were the least stable and were not good predictors of rhythm class, Grabe & Low (2002), White & Mattys (2007) and Wiget et al. (2010) found that most statistically significant cross-linguistic differences were reflected in vocalic metrics. As a result, Grabe & Low (2002: 523) conclude that nPVI-V "provides a better separation of languages than the rPVI[-C]", while Wiget et al. (2010: 1561) characterize %V and VarcoV as "particularly useful" for similar reasons. In both cases, sensitivity to presumed language differences is taken to be an advantage of a metric and is used in turn to support the typology the metrics were meant to test. But given that the present results showed the opposite pattern, it is not possible to decide which metric provides a more accurate quantification of a language's rhythm, unless one a priori accepts the separation of languages into rhythm classes,<sup>4</sup> the very same separation that metrics were designed to bolster (Arvaniti, 2009, and Kohler, 2009a).

Finally, these problems with metrics have larger theoretical consequences as well. As noted in the introduction, the notion of rhythm classes was largely abandoned by the early 1990s due to the lack of empirical support. But in the past decade or so, metrics have been said to have provided the evidence needed to support the rhythm typology. The results of the present study, however, confirm on a large scale problems with metrics hinted at by many previous studies. They show that metrics are sensitive to low-level and inevitable data noise to such an extent that these effects can be comparable, if not larger, than the language effects metrics are meant to capture. Since it is rather unlikely that metrics would be more sensitive to extemporaneous than cross-linguistic variability in timing, these effect sizes most likely indicate that durational variability is much more extensive within each language than previously thought; as a result, substantial cross-linguistic overlap in timing patterns is likely to be the norm, as suggested by the present data. If so, it is unclear that consistent rhythm differences exist cross-linguistically, *at least as conceived by the view of rhythm as timing quantified by metrics*, especially to such an extent that they could be used to guide acquisition and speech processing.

## 5. Conclusion

The present results show that metrics cannot be reliably used to classify languages into rhythmic classes. On the one hand, the results they provide depend largely on tangential factors, such as inter-speaker variation, elicitation, and the syllable composition of materials. On the other, the language differences that metrics are intended to measure appear to be modest relative to these external effects, possibly a corollary of extensive language-internal variability in durational patterns. Further, the effects that metrics capture are both opaque and erratic, characteristics that do not allow for consistent control in experimental settings. Because of these problems, rhythmic classification on the basis of metrics and comparisons of results across different studies is risky at best. Since, so far, little support for the division of languages into rhythm classes has been provided

---

<sup>4</sup> It is possible that in part these differences between the present study and those of Grabe & Low (2002), White & Mattys (2007) and Wiget et al. (2010) are due to measurement protocols: as noted, Grabe & Low excluded final intervals and combined intervals of the same type across a pause into one measurement; White & Mattys and Wiget et al. adopted the same protocol and in addition excluded sonorants from their materials. Clearly these choices can affect the durational variability and extent of vocalic intervals but this explanation does no more than provide additional evidence for the sensitivity of metrics to factors tangential to their purpose.

other than evidence from metrics, the sensitivity of metrics to extemporaneous variability casts further doubt on the idea of rhythm classes as a valid construct and to theories that support it.

### **Acknowledgements**

Special thanks are due to my students, Younah Chung, Page Piccinini and Nadav Sofer for crucial help with data annotation and feedback, to Ken de Jong for extensive editorial input and help with bibliography, to Sun-Ah Jun who, in addition to providing references for Korean, checked and corrected the Korean data and transcriptions, to Peggy Mok and Hae-Sung Jeon for additional help with references for Korean and to Sam Tilsen who alerted me to some annotation issues in an earlier version of the study. Thanks are also due to Tristie Ross, for crucial input at the early stages of this work, to Noah Girgis for his help with annotation, to Alex del Guidice, Naja Ferjan, Nancy Gil, Christina Lee, Jini Shim and Amanda Simons for their contribution to data preparation, collection and management, and to Kris Phillips and Yanni Arvanitis for technical support. The financial support of the University of California San Diego Committee on Research through grant no RI-201G to Amalia Arvaniti with Tristie Ross as GSR is hereby gratefully acknowledged.

## Appendix A

The story of the North Wind and the Sun in the six languages of the study.

### ENGLISH

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

### GERMAN

Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam. Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzulegen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da mußte der Nordwind zugeben, daß die Sonne von ihnen beiden der Stärkere war.

### GREEK

Ο βοριάς κι ο ήλιος μάλωναν για το ποιος απ' τους δυο είναι ο δυνατότερος, όταν έτυχε να περάσει από μπροστά τους ένας ταξιδιώτης που φορούσε κάπα. Όταν τον είδαν, ο βοριάς κι ο ήλιος συμφώνησαν ότι όποιος έκανε τον ταξιδιώτη να βγάλει την κάπα του θα θεωρούνταν ο πιο δυνατός. Ο βοριάς άρχισε τότε να φυσάει με μανία, αλλά όσο περισσότερο φυσούσε τόσο περισσότερο τυλιγόταν με την κάπα του ο ταξιδιώτης, ώσπου ο βοριάς κουράστηκε και σταμάτησε να φυσάει. Τότε ο ήλιος άρχισε με τη σειρά του να λάμπει δυνατά και γρήγορα ο ταξιδιώτης ζεστάθηκε κι έβγαλε την κάπα του. Έτσι ο βοριάς αναγκάστηκε να παραδεχτεί ότι ο ήλιος είναι πιο δυνατός απ' αυτόν.

### ITALIAN

Il vento del nord ed il sole stavano discutendo su chi, tra i due, fosse il più forte quando arrivò un viaggiatore avvolto in un mantello. I due decisero che il primo di loro che fosse riuscito a far togliere il mantello al viaggiatore sarebbe stato il più forte tra i due. Quindi il vento del nord soffiò più forte che mai, ma più lui soffiava più il viaggiatore si avvolgeva nel suo mantello; fin a quando il vento rinunciò. Allora il sole lo riscaldò con i suoi raggi e, immediatamente, il viaggiatore si tolse il mantello. Fu così che il vento del nord ammise che il sole era il più forte tra i due.

### KOREAN

북풍과 햇님이 서로 힘이 더 세다고 다투고 있을 때, 한 나그네가 따뜻한 외투를 입고 걸어 왔다. 그들은 누구든지 나그네의 외투를 먼저 벗기는 이가 힘이 더 세다고하기로 결정했다. 북풍은 힘껏 불었으나 불면 불수록 나그네는 외투를 단단히 여몄다. 그 때에 햇님이 뜨거운 햇빛을 가만히 내려쬐니, 나그네는 외투를 얼른 벗었다. 이리하여 북풍은 햇님이 둘중에 힘이 더 세다고 인정할수 없었다.

### SPANISH

El viento norte y el sol porfiaban sobre cuál de ellos era el más fuerte, cuando acertó a pasar un viajero envuelto en ancha capa. Convinieron en que quien antes lograra obligar al viajero a quitarse la capa sería considerado más poderoso. El viento norte sopló con gran furia, pero cuanto más soplabla, más se arrebujaba en su capa el viajero; por fin el viento norte abandonó la empresa. Entonces brilló el sol con ardor, e inmediatamente se despojó de su capa el viajero; por lo que el viento norte hubo de reconocer la superioridad del sol.

## Appendix B

### ENGLISH

#### “stress-timed”

*Andrew introduced McGivney to my best friends, Clare, Lindsey and Kris.  
The problem required quite a long of strange equations and wasn't very easy.  
It was pretty clear from his presentation that he didn't know the product well.  
The production increased by three fifths in the last quarter of 2007.  
I just called Trent to confirm the appointment we had scheduled last Monday.*

#### “syllable-timed”

*Lara saw Bobby when she was on the way to the photocopy room.  
Everyone got up to leave as soon as the teacher said to do so.  
Tina did better than anyone of us could hope to do in the race.  
Sally and I were at Annie's house today planning our party.  
Two-year-old Lucy has macaroni and cheese every day for diner.*

#### “uncontrolled”

*When a man gets killed I never like to get mixed up in it in any way.  
Through this twilight universe Daisy began to move again with the season.  
It was nine o'clock when we finished breakfast and went out on the porch.  
Some little boys had come up on the steps and were looking into the hall.  
I called Gatsby's house a few minutes later, but the line was busy.*

“Uncontrolled” sentences from F. Scott Fitzgerald's *The Great Gatsby* (1925).

### GERMAN

#### “stress-timed”

*Gerhard hat Salzkartoffeln und Schweinefleisch mit Pflaumen bestellt.  
Der schwerverletzte Mann wurde gestern ins Krankenhaus gebracht.  
Die Austauschstudentin wollte Englisch als Fremdsprache unterrichten.  
Wahrscheinlich könnte Michael nach zwölf Uhr nicht mehr fernsehen.  
Gerhards Schwiegermutter hat ihren Geschirrspüler kaputtgemacht.*

#### “syllable-timed”

*Er wollte seine Freundin Susanna zum Geburtstag ins Kino einladen.  
Meine Lehrerin hat mir gestern ein neues pinkes Heft gegeben.  
Tina ist gestern um neun Uhr Abend in London angekommen.  
Meine Mutti wollte schon um neun Uhr mit dem Zug nach Berlin fahren.  
Heute Abend um elf geht Martin mit seinem Freund Markus aus.*

#### “uncontrolled”

*Sabine hat viele Burgen und Schlösser auf den Bergen gesehen.  
Die beiden Frauen sind ein Stück mit dem U-Bahn gefahren.  
An den Wänden hängen viele Bilder von Musikgruppen und Pferden.  
Endlich hat Sabine ihren Vater und ihre Mutter wiedergesehen.  
Zuerst hat die Familie in einem Gasthaus zu Mittag gegessen.*

“Uncontrolled” sentences from Razma Lazda-Cazers & Helga Thorson (2005). *Neuer Wein und Zwiebelkuchen: A Cultural Reader*. McGraw Hill Publishers.

## GREEK

### “stress-timed”

Οι άσπρες γλαδιόλες που παραγγέλνουν απ' την Αυστραλία έφτασαν μαραμένες.

[i 'aspres ɣla'djoles pu para'jelnun ap tin afstra'lia 'eftasan mara'menes]

Τα ξενόγλωσσα βιβλία βιολογίας είναι πολύ ακριβά στην Ελλάδα.

[ta kse'noɣlosa vi'nlia violo'jias ine po'li akri'va stin e'laða]

Η Σταματία ξετρελλάθηκε με τα καινούργια σκι που της πήραν στα γενέθλιά της.

[i stama'tia ksetre'laθice me ta ce'nurja 'ski pu tis 'piran sta ɣe'neθli'a tis]

Ο Πέτρος αγόρασε ένα άσχημο αλλά πανάκριβο πορτατίφ για το γραφείο του.

[o 'petros a'ɣorase 'ena 'asχimo a'la pa'nakrivo porta'tif ɣa to ɣra'fio tu]

Στην Ερμού γίνεται πάντα στριμωξίδι κατά τη διάρκεια των εκπτώσεων.

[stin er'mu 'jinete 'pada strimo'ksiði kata ti ði'arcia ton ek'ptoseon]

### “syllable-timed”

Το καπέλο που φορούσε την έκανε να μοιάζει με πουλί του παραδείσου.

[to ka'pelo pu fo'ruse tin 'ekane na 'mɣazi me pu'li tu para'disu]

Το κοριτσάκι που παίζει στον κήπο είναι κόρη του Χαράλαμπου.

[to kori'tsaci pu 'pezi sto 'ɣipo 'ine 'kori tu xa'ralabu]

Οι μόνοι συγγενείς του Μανώλη είναι η γιαγιά του κι η μητέρα του.

[i 'moni siɣe'nis tu ma'noli 'ine i ɣa'ɣa tu ci mi'tera tu]

Το καλοκαίρι μ'αρέσει πολύ να πηγαίνω με το καράβι από νησί σε νησί.

[to kalo'ceri ma'resi po'li na pi'ɣeno me to ka'ravi apo ni'si se ni'si]

Σου παράγγειλα σαλάτα μαρούλι και μακαρόνια με κιμά.

[su pa'rajila sa'lata ma'ruli ce maka'rona me ci'ma]

### “uncontrolled”

Όλοι οι μεγάλοι νοσταλγούν τον παράδεισο της παιδικής τους ηλικίας.

[ 'oli i me'ɣali nostaλ'ɣun to ba'raðiso tis peði'cis tus ili'cias]

Καμιά φορά προσπαθώ να θυμηθώ τα χαρακτηριστικά του και δεν μπορώ.

[ka'mɣa fo'ra prospa'θo na θimi'θo ta xarakteristi'ka tu ce 'ðe bo'ro]

Ο Μιλτιάδης ήταν μόνιμος αξιωματικός του στρατού εν αποστρατεία.

[o milti'aðis 'itan 'monimos aksiomati'kos tu stra'tu en apostra'tia]

Στο σχολείο ήμουνά πάντα η πρώτη στην έκθεση ιδεών και στην ιστορία.

[sto sxo'lio 'imun 'pada i 'proti stin 'ekthesi iðe'on ce stin isto'ria]

Μόλις μπήκα στην αυλή με πήρε η μυρωδιά του μοσχολίβανου απ'τη μύτη.

[ 'molis 'bika stin a'vli me 'pire i miro'ðɣa tu mosxo'livanu ap ti 'miti]

“Uncontrolled” sentences from Kostas Tachtsis’ *To Trito Stefani (The Third Wedding)*, 1962.

## ITALIAN

### “stress-timed”

*Sembra che tutte le volte che la gallina entra qui quel gatto diventi pazzo.*

*Quell'uomo e quella donna folleggiavano da mattina a sera.*

*In spiaggia i bambini si schizzano con l'acqua mentre i nonni dormono.*

*Un governo internazionale è necessario per realizzare la fine della immigrazione illegale.*

*Ammiro il rapido movimento delle ali del pettirosso durante il suo volo.*

### “syllable-timed”

*Il cane vuole riposare vicino alle pecore durante l'estate.*

*Ieri sera ho venduto la tavola che stava vicino al muro.*

*Poco dopo che bevo la medicina mi sento più felice.*

*Dopodomani porto i giovani a vedere veramente cosa è “il lavoro”.*

*Davide ha cucinato le patate dopo che ha bevuto del vino.*

“uncontrolled”

*L'eruzione continuò in modo spettacolare per un mese intero.*

*Nel quadro che abbiamo visto il pittore volle raffigurare una vista naturale.*

*La città nasce quando ciascuno di noi non è più sufficiente a se stesso, ma ha bisogno di molti altri.*

*Al piede di molte carte geografiche vi sono dei simboli.*

*Il capoluogo della regione non dovrebbe essere il centro industriale.*

“Uncontrolled” sentences from *La Nuova Geografia Loescher*. Loescher: Rome, 1995.

### Korean

“stress-timed”

유나는 굉장히 바빴고 식욕이 없어서 살이 빠졌다.

[junanin kwendzanhi pap\*atk\*o jigjogi ʌps\*ʌsʌ sari p\*adzʌtt\*a]

난 눈만 감았다 떴는데 시합이 벌써 끝나버렸다.

[nan nunman kamatt\*a t\*ʌninde jihabi pʌls\*ʌ k\*innabʌrjʌtt\*a]

난 수업을 많이 결석해서 4년안에 졸업을 할 수 없다.

[nan suʌbil mani kjʌls\*ʌkʰesʌ sanjanane tʃorʌbuʌl hals\*u ʌpt\*a]

아이들 10명이나 납치한 유괴범은 아직도 잡히지 않았다.

[aidil jʌlmjʌnina naptʃʰihan jugwebʌmin atʃikt\*o tʃapʰidzi anatt\*a]

편입생은 8학점 따는데 2년이나 걸렸다면 억울해했다.

[pʰjanips\*enjin pʰalhaktʃ\*ʌm t\*ʌninde injʌnina kalljʌt\*amjʌ ʌgu.lhehett\*a]

“syllable-timed”

어머니는 치마보다 바지를 더 입으세요.

[ʌmʌninin tʃʰimaboda padziril tʌ ibisejo]

미라는 얼굴도 예쁘고 노래도 잘 하니깐 인기가 많다.

[miranin ʌguldo jep\*igo noredo tʃʌlhanik\*a ink\*iga mantʰa]

선아가 보고싶어도 시간이 없고 바빠서 만날 수 없다.

[sanaga pogoʃipʰʌdo ʃigani ʌpk\*o pap\*ʌsʌ mannals\*u ʌpt\*a]

날씨가 너무 더우니까 수영하고 싶다.

[nalʃ\*iga namu tʌunik\*a sujʌnhago ʃipt\*a]

나는 매일 10시간이나 자도 피곤해서 큰 문제다.

[nanin meil jʌlʃ\*iganina tʃado pʰigonhesʌ kʰin mundzeda]

“uncontrolled”

이 날엔 소녀가 징검다리 한가운데 앉아 세수를 하고 있었다.

[inaren sonjʌga tʃingʌmdari hangaunde andʒa sesuril hago is\*ʌtt\*a]

분홍 스웨터 소매를 걷어올린 팔과 목덜미가 마냥 희었다.<sup>5</sup>

[punhoŋ siwetʰʌ someril kʌdʌollin makt\*ʌmiga manjaŋ hiʌtt\*a]

내일 소녀네가 양평읍으로 이사 간다는 것이었다.

[neil sonjanega jaŋpʰjʌŋibiŋo isa kandanin kʌʃiʌtt\*a]

그 날 밤, 소년은 자리에 누워서도 같은 생각뿐이었다.

[kinalp\*am sonjanin tʃarie nuwʌsʌdo katʰin seŋgak p\*uniʌtt\*a]

저 꽃을 보니까 등나무 밑에서 놀던 동무들 생각이 난다.

[tʃʌ k\*otʃʰil ponik\*a tiŋnamu mitʰesʌ noldʌn toŋmudil seŋgagi nanda]

“Uncontrolled” sentences from Soon-Won Hwang story *소나기 (The Rain shower)*, 1959.

<sup>5</sup> This sentence was inadvertently incomplete in the cards the speakers used for reading, in that 팔 과 was missing. The sentence was included in the calculation of metric scores, since the speakers used regular prosody to produce it. I am grateful to Sun-Ah Jun for pointing out this problem.

## Spanish

“stress-timed”

*Un zoólogo estaba inspeccionando unos especímenes nuevos.  
Daniel, Enrique y Juan van a viajar a Japón por un mes.  
A los doctores les gusta caminar por el parque central de La Paz.  
El ingeniero siempre parecía bastante amable.  
Nunca había visto El Jirón de la Unión tan desierto y oscuro.*

“syllable-timed”

*El muchacho le da una rosa a su hermana cada sábado.  
Sara dice que la playa es muy bonita durante el verano.  
Mañana iré al mercado para comprar una papaya.  
La casa de la profesora no parece pequeña.  
No sé si mi jefe se relajará la próxima semana.*

“uncontrolled”

*Se había ido sin escándalo, de común acuerdo del esposo.  
Es la primera vez que te oigo decir algo que no debías.  
Las oficinas estaban cerradas y a oscuras por el día feriado.  
Esto es pecado quemarlo, con tanta gente que no tiene ni que comer.  
Las visitas empezaron a adquirir una incomoda amplitud familiar.*

“Uncontrolled” sentences from Gabriel García Márquez’s *El amor en los tiempos del cólera* (1985).

## Appendix C

Scores and standard errors (in brackets) for each language and metric (and pooled across languages); at the top, results are given separately for sentences (sents), reading of the North Wind and the Sun (story) and spontaneous speech (SS); at the bottom, results are given separately for each set of sentences.

	$\Delta C$			rPVI-C			VarcoC		
	sents	story	SS	sents	story	SS	sents	story	SS
English	57 (2)	54 (2)	68 (3)	67 (2)	63 (2)	77 (3)	55 (1)	51 (1)	59 (1)
German	57 (2)	62 (2)	67 (3)	63 (2)	66 (2)	72 (3)	50 (1)	53 (1)	59 (1)
Greek	37 (2)	41 (2)	46 (3)	44 (2)	47 (2)	49 (3)	43 (1)	46 (1)	51 (1)
Italian	42 (2)	43 (2)	46 (3)	49 (2)	49 (2)	50 (3)	53 (1)	49 (1)	53 (1)
Korean	51 (2)	47 (2)	53 (3)	56 (2)	54 (2)	59 (3)	56 (1)	52 (1)	57 (1)
Spanish	46 (2)	45 (2)	49 (3)	52 (2)	54 (2)	55 (3)	51 (1)	47 (1)	53 (1)
<b>Pooled</b>	<b>48 (2)</b>	<b>48 (2)</b>	<b>55 (3)</b>	<b>55 (2)</b>	<b>55 (2)</b>	<b>60 (3)</b>	<b>51 (1)</b>	<b>49 (1)</b>	<b>55 (1)</b>
	%V			nPVI-V			VarcoV		
	sents	story	SS	sents	story	SS	sents	story	SS
English	45 (1)	44 (1)	48 (1)	54 (1)	59 (1)	66 (2)	48 (2)	50 (2)	66 (3)
German	39 (1)	38 (1)	42 (1)	54 (1)	54 (1)	52 (2)	49 (2)	51 (2)	55 (3)
Greek	47 (1)	48 (1)	50 (1)	52 (1)	52 (1)	56 (2)	53 (2)	56 (2)	64 (3)
Italian	50 (1)	50 (1)	52 (1)	46 (1)	46 (1)	53 (2)	48 (2)	53 (2)	63 (3)
Korean	48 (1)	48 (1)	51 (1)	50 (1)	54 (1)	59 (2)	51 (2)	55 (2)	69 (3)
Spanish	49 (1)	49 (1)	50 (1)	45 (1)	47 (1)	56 (2)	47 (2)	47 (2)	66 (3)
<b>Pooled</b>	<b>47 (1)</b>	<b>46 (1)</b>	<b>49 (1)</b>	<b>50 (1)</b>	<b>52 (1)</b>	<b>57 (2)</b>	<b>49 (2)</b>	<b>52 (2)</b>	<b>64 (3)</b>

	$\Delta C$			rPVI-C			VarcoC		
	“stress-timed”	“syllable-timed”	“un-controlled”	“stress-timed”	“syllable-timed”	“un-controlled”	“stress-timed”	“syllable-timed”	“un-controlled”
English	68 (2)	49 (2)	55 (2)	83 (2)	57 (2)	61 (2)	57 (1)	53 (1)	55 (1)
German	62 (2)	54 (2)	55 (2)	73 (2)	60 (2)	56 (2)	51 (1)	50 (1)	50 (1)
Greek	42 (2)	31 (2)	37 (2)	51 (2)	44 (2)	50 (2)	45 (1)	41 (1)	43 (1)
Italian	43 (2)	37 (2)	45 (2)	48 (2)	42 (2)	48 (2)	52 (1)	52 (1)	56 (1)
Korean	49 (2)	53 (2)	51 (2)	56 (2)	56 (2)	57 (2)	52 (1)	61 (1)	54 (1)
Spanish	55 (2)	36 (2)	46 (2)	61 (2)	43 (2)	53 (2)	55 (1)	46 (1)	51 (1)
<b>Pooled</b>	<b>53 (2)</b>	<b>43 (2)</b>	<b>48 (2)</b>	<b>63 (2)</b>	<b>50 (2)</b>	<b>54 (2)</b>	<b>52 (1)</b>	<b>50 (1)</b>	<b>52 (1)</b>
	%V			nPVI-V			VarcoV		
	sents	story	SS	sents	story	SS	sents	story	SS
English	41 (1)	50 (1)	44 (1)	55 (2)	51 (2)	56 (2)	48 (2)	46 (2)	50 (2)
German	36 (1)	41 (1)	41 (1)	55 (2)	56 (2)	53 (2)	44 (2)	52 (2)	52 (2)
Greek	45 (1)	49 (1)	46 (1)	51 (2)	51 (2)	53 (2)	54 (2)	48 (2)	57 (2)
Italian	50 (1)	52 (1)	49 (1)	47 (2)	45 (2)	47 (2)	48 (2)	46 (2)	49 (2)
Korean	47 (1)	50 (1)	49 (1)	47 (2)	52 (2)	50 (2)	52 (2)	51 (2)	50 (2)
Spanish	47 (1)	52 (1)	49 (1)	45 (2)	41 (2)	48 (2)	40 (2)	43 (2)	57 (2)
<b>Pooled</b>	<b>44 (1)</b>	<b>49 (1)</b>	<b>47 (1)</b>	<b>50 (2)</b>	<b>49 (2)</b>	<b>51 (2)</b>	<b>48 (2)</b>	<b>48 (2)</b>	<b>52 (2)</b>

## Reference List

- Abercrobie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.
- Arvaniti, A. (1994). Acoustic features of Greek rhythmic structure. *Journal of Phonetics*, 22, 239-268.
- Arvaniti, A. (1999). Standard Modern Greek. *Journal of the International Phonetic Association*, 29(2), 167-172.
- Arvaniti, A. (2007). Greek phonetics: The state of the art. *Journal of Greek Linguistics*, 8, 97-208.
- Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66, 46-63.
- Arvaniti, A. (in press) Rhythm classes and speech perception. In O. Niebuhr, & H. Pfitzinger (Eds.), *Prosodies. Context, Function, and Communication*. Walter de Gruyter.
- Balasubramanian, T. (1980). Timing in Tamil. *Journal of Phonetics*, 8, 449-467.
- Baltazani, M. (2007). Prosodic rhythm and the status of vowel reduction in Greek. *Selected Papers on Theoretical and Applied Linguistics from the 17<sup>th</sup> International Symposium on Theoretical & Applied Linguistics*, vol. 1 (pp. 31-43). Department of Theoretical and Applied Linguistics, Thessaloniki.
- Barry, W., & Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association*, 31 (1), 51-66.
- Barry, W., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? *Proceedings of XV<sup>th</sup> ICPHS*, Barcelona, Spain (pp. 2693-2696).
- Barry, W., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, 66, 78-94.
- Benton, M., Dockendorf, L., Jin, W., Liu, Y., & Edmondson, J. A. (2007). The continuum of speech rhythm: computational testing of speech rhythm of large corpora from natural Chinese and English speech. *Proceedings of XVI<sup>th</sup> ICPHS*, Saarbrücken, Germany (pp. 1269-1272).
- Bertinetto, P. M. (1989). Reflections on the dichotomy <<stress>> vs. <<syllable timing>>. *Revue de Phonétique Appliquée*, 91-92-93, 99-129.
- Bertrán, A. P. (1999). Prosodic typology: On the dichotomy between *stress*-timed and *syllable*-timed languages. *Language Design*, 2, 103-130.
- Bolinger, D. L. (1965). Pitch accent and sentence rhythm. In I. Abe, & T. Kanekiyo (Eds.), *Forms of English: Accent, Morpheme, Order* (pp. 139-180). Cambridge MA: Harvard University Press.
- Bond, Z. S., D. Markus & V. Stockmal (2007). Prosodic and rhythmic patterns produced by native and non-native speakers of a quantity-sensitive language. *Proceedings of XV<sup>th</sup> ICPHS*, Barcelona, Spain (pp. 527-530).
- Borzzone de Manrique, A. M., & Signorini, A. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics*, 11, 117-128.
- Brimhall, C., Horton, R., & Morgan, E. (2010). A machine learning approach to rhythmic classification of languages. *Journal of the Acoustics Society of America*, 128, 2478.
- Bunta, F., & Ingram, D. (2007). The acquisition of speech rhythm by bilingual Spanish- and English-speaking 4- and 5-year-old children. *Journal of Speech, Language, and Hearing Research*, 50, 999-1014.
- Classé, A. (1939). *The Rhythm of English Prose*. Oxford: Blackwell.
- Coetzee, A. W., & Wissing, D. P. (2007). Global and local durational properties in three varieties of South African English. *The Linguistic Review*, 24, 263-289.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioural sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Erlbaum.
- Cutler, A., Mehler, J., Norris, D., & Seguí, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385-400.
- Cutler, A., Mehler, J., Norris, D., & Seguí, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, 24(3), 381-410.
- Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, 33, 824-844.
- Dankovicová, J., & Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. *Proceedings of XVI<sup>th</sup> ICPHS*, Saarbrücken, Germany (pp. 1241-1244).
- Dauer, R. M. (1980). *Stress and Rhythm in Modern Greek*. Unpublished Ph.D. dissertation, University of Edinburgh.

- Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, 51-62.
- Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. *Proceedings of the XII<sup>th</sup> ICPHS*, Tallinn, Estonia, (pp. 447-449).
- Davis, C., & Gaito, J. (1984). Multiple comparison procedures within experimental research. *Canadian Psychology*, 25, 1-13.
- de Jong, K. (1994). Initial tones and prominence in Seoul Korean. *OSU Working Papers in Linguistics*, 43, 1-14.
- Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), *Language and Language-Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003* (pp. 231-241). Frankfurt am Main, Germany: Peter Lang.
- Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. *Proceedings of the XV<sup>th</sup> ICPHS*, Barcelona, Spain (pp. 471-474).
- Fleischer, J., & Schmid, S. (2006). Zurich German. *Journal of the International Phonetic Association*, 36(2), 243-253.
- Frota, S., & Vigário, M. (2001). On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. *Probus*, 13, 247-275.
- Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515-546). Berlin: Mouton de Gruyter.
- Grabe, E., Watson, E., & Post, B. (1999). The acquisition of rhythmic patterns in English and French. *Proceedings of XIV<sup>th</sup> ICPHS*, San Francisco, USA (pp. 1201-1204).
- Hillenbrand, J. M. (2003). American English: Southern Michigan. *Journal of the International Phonetic Association*, 33(1), 121-126.
- Horton, R. & Arvaniti, A. (2012) Clustering and classifying with rhythm metrics. UC San Diego manuscript.
- Jeon, H-S. (2006). Acoustic measure of speech rhythm: Korean learners of English. Unpublished M.Sc. dissertation, University of Edinburgh.
- Jinbo, K. (1980). Kokugo no onseijou no tokushitsu [The top phonetic characteristics of Japanese]. In T. Shibata, H. Kitamura, & H. Kindaichi (Eds.), *Nihon no gengogaku [Linguistics of Japan]* (pp. 5-15). Tokyo: Taishukan (originally published 1927).
- Jones, D. (1972). *An Outline of English phonetics* (9<sup>th</sup> ed.). Cambridge: Cambridge University Press [first published 1918].
- Jun, S-A. (1995). A phonetic study of stress in Korean. *Journal of the Acoustical Society of America*, 98(5), 2893.
- Jun, S-A. (2005). Korean intonational phonology and prosodic transcription. In S-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 201-229). Oxford: Oxford University Press.
- Keane, E. (2006). Rhythmic characteristics of colloquial and formal Tamil. *Language and Speech*, 49 (3), 299-332.
- Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and Speech*, 51(4), 343-359.
- Kohler, K. J. (1999). German. *Handbook of the International Phonetic Association* (pp. 86-89). Cambridge: Cambridge University Press.
- Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica*, 65, 257-269.
- Kohler, K. J. (2009a). Whither speech rhythm research? *Phonetica*, 66, 5-14.
- Kohler, K. J. (2009b). Rhythm in speech and language. A new research paradigm. *Phonetica*, 66, 29-45.
- Ladefoged, P. (1999). American English. *Handbook of the International Phonetic Association* (pp. 41-44). Cambridge: Cambridge University Press.
- Lee, H. B. (1999). Korean. *Handbook of the International Phonetic Association* (pp. 120-123). Cambridge: Cambridge University Press.
- Lee, J., Jin, N., Seong, C., Jung, I., & Lee, S. (1994). An experimental phonetic study of speech rhythm in standard Korean. *Proceedings of ICSLP 94*, Yokohama, Japan (pp. 1091-1094).
- Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5, 253-263.
- Lin, H., & Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of Chinese Language and*

*Computing*, 17 (3), 127-140.

- Liss, J. M., White, L., Mattys, S. L., Lansford, K., Spitzer, S., Lotto, A. J., & Caviness, J. N. (2009). Quantifying speech rhythm deficits in the dysarthrias. *Journal of Speech, Language, and Hearing Research*, 52(5), 1334-1352.
- Lleó, C., Rakow, M., & Kehoe, M. (2007). Acquiring rhythmically different languages in a bilingual context. *Proceedings of XVI<sup>th</sup> ICPPhS*, Saarbrücken, Germany (pp. 1545-1548).
- Lloyd James, A. (1940). *Speech signals in telephony*. London: Pitman & Sons.
- Loukina, A., Kochanski, G., Shih, C., Keane, E., Watson, I. (2009). Rhythm measures with language-independent segmentation. *Proceedings of Interspeech 2009*, pp. 1531-1534.
- Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43, 377-401.
- McAuley, J. D., & Riess Jones, M. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1102-1125.
- Major, R. C. (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics*, 9, 343-351.
- Martínez-Celdrán, E., Fernández-Planas A., & Carrera-Sabaté, J. (2003). Castilian Spanish. *Journal of the International Phonetic Association*, 33(2), 255-259.
- Mattys, S. L., & Melhorn, J. F. (2005). How do syllables contribute to the perception of spoken English? Insight from the migration paradigm. *Language and Speech*, 48(2), 223-253.
- Miller, M. (1984). On the perception of rhythm. *Journal of Phonetics*, 12, 75-83.
- Mok, P. (2009). On the syllable-timing of Cantonese and Beijing Mandarin. *Chinese Journal of Phonetics*, 2, 148-154.
- Mok, P. P. K. (2011). The acquisition of speech rhythm by three-year-old bilingual and monolingual children: Cantonese and English. *Bilingualism: Language and Cognition*, 14 (4), 458-472.
- Mok, P. P. K., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. *Proceedings of Speech Prosody 2008*, Campinas, Brazil (pp. 423-426).
- Mok, P. P. K., & Lee, S. I. (2008). Korean speech rhythm using rhythmic measures. Paper presented at the 18th International Congress of Linguists (CIL18). Seoul, Korea [available at <http://www.cuhk.edu.hk/lin/people/peggy/>]
- Moon-Hwan, C. (2004). Rhythm typology of Korean speech. *Cognitive Processing*, 5, 249-253.
- Murty, L., Otake, T., & Cutler, A. (2007). Perceptual tests of rhythmic similarity: I. Mora rhythm. *Language and Speech*, 50, 77-99.
- Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, 38, 84-106.
- Nazzi, T., Bertoincini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, 24, 756-766.
- Nazzi, T., Iakimova, G., Bertoincini, J., Frédonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54(3), 283-299.
- Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43 (1), 1-19.
- Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, 41, 233-243.
- Nespor, M., & Vogel, I. (1989). On clashes and lapses. *Phonology*, 6, 69-116.
- Nespor, M. (1990). On the rhythm parameter in phonology. In I. M. Roca (Ed.), *Logical Issues in Language Acquisition* (pp. 157-175). Dordrecht: Foris.
- Nolan, F., & Asu, E. L. (2009). The Pairwise Variability Index and coexisting rhythms in language. *Phonetica*, 66, 64-77.
- Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258-278.

- Parker, Jones, O. (2006). Durational variability and stress-timing in Hawaiian. In P. Warren & C. I. Watson (Eds.), *Proceedings of the 11th Australian International Conference on Speech & Science Technology* (pp. 417–420).
- Payne, E., Post, B., Astruc, L., Prieto, P., Vanrell, M. (2011). Measuring child rhythm. *Language and Speech*, 54, 1-27.
- Pike, K. (1945). *The Intonation of American English*. Ann-Arbor: University of Michigan Press.
- Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, 8, 293-304.
- Pointon, G. E. (1995). Rhythm and duration in Spanish. In J. W. Lewis (Ed.), *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor* (pp. 266-269). New York: Routledge.
- Prieto, P., Vanrell, M., Astruc, L., Payne, E., Post, B. (in press). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*.
- Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France.
- Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm class: Perceptual studies. *Proceedings of the XV<sup>th</sup> ICPHS*, Barcelona, Spain (pp. 337-340).
- Ramus, F., Nespors, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265-292.
- Renwick, M. E. L. (2011) Quantifying rhythm: Interspeaker variation in %V. *Journal of the Acoustical Society of America*, 130, 2567.
- Roach, P. (1982). On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. In D. Crystal, (Ed.) *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F. R. Palmer* (pp. 73-79). London: Edward Arnold.
- Rodriguez, T. & Arvaniti, A. (2011) Rhythm, tempo and F0 in language discrimination. *Journal of the Acoustical Society of America*, 130, 2567.
- Rogers, D., & d’Arcangeli, L. (2004). Italian. *Journal of the International Phonetic Association*, 34(1), 117–121.
- Rubach, J., & Booij, G. E. (1985). A grid theory of stress in Polish. *Lingua*, 66, 281-319.
- Russo, M., & Barry, W. J. (2008). Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo. *Proceedings of Speech Prosody 2008*, Campinas, Brazil, May 6-9, 2008 (pp. 419-422).
- Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics*, 13, 155-162.
- Shen, Y., Peterson, G. G. (1962). Isochronism in English. *University of Buffalo Studies in Linguistics Occasional Papers*, 9, 1-36.
- Tsiartsioni E. (2003). *The acquisition of features of rhythm and stop voicing in Greek and English L2*. Unpublished M.Phil. dissertation, Trinity College Dublin.
- Uldall, E. T. (1971). Isochronous stresses. In L. L. Hammerich, R. Jakobson, & E. Zwirner (Eds), *Form and substance: Phonetic and Linguistic Papers presented to Eli Fischer-Jorgensen* (pp. 205-210). Copenhagen: Akademisk Forlag.
- Wagner, P. S., & Dellwo, V. (2004). Introducing YARD (yet another rhythm determination) and re-introducing isochrony to rhythm research. *Proceedings of Speech Prosody 2004*, Nara, Japan.
- Warner, N., & Arai, T. (2001). Japanese mora timing: A review. *Phonetica*, 58, 1-25.
- Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, 10, 193-216.
- Wheeler, M. (2005). *Phonology of Catalan*, Oxford: Blackwell.
- White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35, 501-522.
- Whitworth, N. (2002). Speech rhythm production in three German-English bilingual families. In D. Nelson (Ed.), *Leeds Working Papers in Linguistics and Phonetics*, 9, 175-205.
- Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, 127, 1559-1569.

- Yoshida, K., Yoon, J., & Kim, H. (2007). Production and perception of word prosody in three dialects of Korean. *Proceedings of XVI<sup>th</sup> International Congress of Phonetic Sciences*, Saarbrücken, Germany (pp. 1169-1172).
- Yu, A. C. L. (2010). Tonal effects on perceived vowel duration. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (Eds.), *Laboratory Phonology 10* (pp. 151-168). Berlin/New York: De Gruyter Mouton.
- Yun, I. (1998). *A Study of Timing in Korean speech*. Unpublished Ph. D. dissertation, The University of Reading.