Running head: Correlates of linguistic rhythm

Correlates of linguistic rhythm in the speech signal

Franck Ramus[a,*], Marina Nespor[b], Jacques Mehler[a]

[a] *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS), 54 boulevard Raspail, 75006 Paris, France*
[b] *Holland Institute of generative Linguistics, University of Amsterdam, The Netherlands, and Facoltà di Lettere, Università di Ferrara, via Savonarola 27, 44100 Ferrara, Italy*

* Corresponding author. Email: ramus@lscp.ehess.fr

Abstract

Spoken languages have been classified by linguists according to their rhythmic properties, and psycholinguists have relied on this classification to account for infants' capacity to discriminate languages. Although researchers have measured many speech signal properties, they have failed to identify reliable acoustic characteristics for language classes. This paper presents instrumental measurements based on a consonant/vowel segmentation for eight languages. The measurements suggest that intuitive rhythm types reflect specific phonological properties, which in turn are signaled by the acoustic/phonetic properties of speech. The data support the notion of rhythm classes and also allow the simulation of infant language discrimination, consistent with the hypothesis that newborns rely on a coarse segmentation of speech. A hypothesis is proposed regarding the role of rhythm perception in language acquisition.

Keywords: speech rhythm; syllable structure; language discrimination; language acquisition; phonological bootstrapping.

1 Introduction

There is a clear difference between the prosody of languages such as Spanish or Italian on the one hand and that of languages like English or Dutch on the other hand. Lloyd James (1940) attributed this difference to rhythm and used the metaphor “machine-gun rhythm” for the first group of languages and “Morse code rhythm” for the second. In the two groups, different elements would recur at regular intervals establishing temporal organization: syllables in Spanish or Italian and stresses in English or Dutch. Pike (1945) thus renamed the two types of rhythms “syllable-timed” and “stress-timed”, and Abercrombie (1967) went a step further by claiming that linguistic rhythm was either based on the isochrony of syllables, or on the isochrony of interstress intervals, for all languages throughout the world. Further work generally classified Germanic and Slavonic languages, as well as Arabic, as stress-timed, Romance languages as syllable-timed, and hypothesized a third category of mora-timed languages, including Japanese and Tamil (Abercrombie, 1967; Bertinetto, 1989; Ladefoged, 1975; Pike, 1945; Port, Dalby, & O'Dell, 1987; Rubach & Booij, 1985; Steever, 1987).

Linguists are not alone in having committed themselves to these distinctions. Mehler, Dupoux, Nazzi and Dehaene-Lambertz (1996) relied on the syllable-timing/stress-timing dichotomy to explain how infants may learn part of the phonology of their native language. Indeed, they hypothesized that rhythm type should be correlated with the speech representation unit in any given language. That is, speakers of stress-timed languages should represent speech in feet, speakers of syllable-timed languages in syllables, and speakers of mora-timed languages in morae (Cutler, Mehler, Norris, & Segui, 1992; Cutler & Otake, 1994). Thus, precocious detection of the rhythm type of their native language might be a simple way for infants to decide which representation unit to use for further speech analysis (this view is further discussed in the general discussion).

Furthermore, this hypothesis also makes predictions for children exposed to a bilingual environment: if the two languages in question share the same representation unit, infants should have no trouble in selecting it; if the two languages do not, infants will be receiving contradictory input and should be confused, unless, of course, they are able to discriminate between the two languages without needing to segment speech: in this case they would be aware that two separate units are to be used. Mehler et al. (1996) therefore hypothesized that infants use rhythm to discriminate languages when they are exposed to languages of different rhythmic classes. This hypothesis was supported by Mehler et al. (1988), Bahrick & Pickens (1988), Moon, Cooper and Fifer (1993), Dehaene-Lambertz and Houston (1998), and Christophe and Morton (1998) who showed that young infants, including newborns (Mehler et al., 1988; Moon et al., 1993), can discriminate between sentences drawn from their mother tongue and sentences from a language belonging to another rhythmic class[1]. Moreover, in two of these studies (Dehaene-Lambertz & Houston, 1998; Mehler et al., 1988), discrimination was possible with low-pass filtered speech (at 400 Hz), suggesting prosodic cues alone can be used.

Finally, the most convincing support for the “rhythm-based language discrimination hypothesis” is provided by Nazzi, Bertoncini and Mehler (1998), who showed, using filtered speech exclusively, that French newborns can discriminate between English and Japanese sentences, but not between Dutch and English ones. Moreover, they also showed that newborns can perform discrimination at the more abstract level of the rhythmic class: they discriminated a set of English and Dutch sentences from a set of Spanish and Italian ones, but

failed to discriminate English and Spanish sentences from Dutch and Italian ones, strongly suggesting that rhythmic classes play an actual role in the infant's perception of speech. Thus, it seems that phoneticians' intuitions were right, and that the syllable-timing/stress-timing dichotomy may well be deeply anchored in the human perceptual system.

2 Current views on speech rhythm

2.1 Against the isochrony theory

Given the excellent reasons for believing in rhythmic classes, one would expect that these groups of languages should differ by readily identifiable acoustic or phonetic parameters. However, this is not the picture provided by past studies on rhythm.

A considerable amount of phonetic research has been carried out to test the physical reality of the isochrony theory on syllable- and stress-timed languages. Nevertheless this research has negated rather than confirmed the existence of different types of isochronous intervals in spoken language.

As far as stress-timed languages are concerned, it has been shown that the duration of interstress intervals in English is directly proportional to the number of syllables they contain (Bolinger, 1965; Lea, 1974; O'Connor, 1965; Shen & Peterson, 1962). Bolinger (1965) also showed that the duration of interstress intervals is influenced by the specific types of syllables they contain as well as by the position of the interval within the utterance. Interstress intervals thus don't seem to have a constant duration.

As regards syllable-timed languages, it has been shown that it is impossible to speak of isochronous syllables in French. Rather, larger rhythmic units - of the size roughly corresponding to the phonological phrase in prosodic phonology - which are characterized by final lengthening, would be responsible for rhythm in French (Wenk & Wiolland, 1982). For Spanish, it has been shown that syllable duration is not constant and that interstress intervals tend to cluster around an average duration (Borzone de Manrique & Signorini, 1983).

A study was also carried out involving six languages, of which, according to Abercrombie (1967), three are classified as syllable timed - French, Telegu and Yoruba - and three as stress timed - Arabic, English and Russian (Roach, 1982). The results of this research are a) that variation in syllable duration is similar in all six languages and b) that stress pulses are not more evenly spaced in the second group of languages than they are in the first.

Similar work was done by Dauer (1983) on English (stress-timed), and Spanish, Italian and Greek (syllable-timed). She concluded that a) the mean duration of interstress intervals for all the languages analyzed is proportional to the number of syllables in the interval, b) stresses recur no more regularly in English than in the other languages. These findings confirm Allen's (1975) observation that motor tasks are characterized by a preferred rate of performance that places beats at intervals of a limited range.

The studies mentioned above do not lead us to adhere to a strict theory of isochrony. The subjective perception of isochrony, if real, must therefore be based on a more abstract construct, possibly similar to that which governs the relation between underlying beats and surface rhythm in music (Cooper & Meyer, 1960; Lerdahl & Jackendoff, 1983; see also Drake & Palmer, 1993).

2.2 A new account of speech rhythm

A view of rhythm radically different from that of Abercrombie and Pike, among others, was first proposed by Bertinetto (1981) and Dasher and Bolinger (1982), according to whom the impression of different types of rhythm is the result of the coexistence of specific

phonological phenomena in a given system. The distinction between syllable-timed and stress-timed languages would thus not be a primitive of phonology, but rather a product of their respective phonological properties.
Along this line of research, Dauer (1983) observed that stress-timed and syllable-timed languages have a number of different distinctive phonetic and phonological properties, of which the two most important are:

- Syllable structure: stress-timed languages have a greater variety of syllable types than syllable-timed languages. As a result, they tend to have heavier syllables[2]. In addition, this feature is correlated with the fact that in stress-timed languages, stress most often falls on the heaviest syllables, while in syllable-timed languages stress and syllable weight tend to be independent.
- Vowel reduction: in stress-timed languages, unstressed syllables usually have a reduced vocalic system (sometimes reduced to just one vowel, schwa), and unstressed vowels are consistently shorter, or even absent.

These features combine to give the impression that some syllables are far more salient than others in stress-timed languages, and that all syllables tend to be equally salient in syllable-timed languages. This in turn, creates the impression that there are different types of rhythm.
In addition, Dauer (1987) suggested that the different properties mentioned above could be independent and cumulative: a language shall not be either stress-timed or syllable-timed with the corresponding properties; rather, the more typical stress-timed language properties a language presents, the more it is stress-timed, and the less it is syllable-timed. Dauer thus advocated a continuous unidimensional model of rhythm, with typical stress-timed and syllable-timed languages at either end of the continuum.

2.3 Existence of intermediate languages

Nespor (1990) supported this view with critical examples, arguing that indeed, there are languages whose features match neither those of typical stress-timed languages, nor those of typical syllable-timed languages. Even though they have most often been described respectively as syllable-timed and stress-timed, Catalan and Polish are such languages. Indeed, Catalan has the same syllabic structure and complexity as Spanish, and thus should be syllable-timed, but it also presents the vowel reduction phenomenon, which is consistently associated with stress-timed languages. Polish presents the opposite configuration: a great variety of syllable types and high syllabic complexity, like stress-timed languages, but no vowel reduction at normal speech rates. Thus these two languages would rate as intermediate on a rhythmic scale like the one proposed by Dauer (1987). As a matter of fact, no firm agreement on their rhythmic status had been arrived at by phonologists (Hayes & Puppel, 1985; Mascaró, 1976; Rubach & Booij, 1985; Wheeler, 1979).
It should be noticed that while Dauer (1987) proposed that languages may be scattered along a continuum, the fact that some languages fall between typically syllable-timed and stress-timed languages does not exclude the possibility that there are just more classes than those originally proposed. For instance, it has been proposed that twelve inventories of possible syllable types can be grouped into five classes according to their markedness (Levelt & van de Vijver, 1998). Three of these classes seem to correspond to the three rhythmic classes described in the literature. It might very well be that the other two classes - both containing less studied languages - have characteristic rhythms, pointing to the possibility that there are more rhythmic classes rather than a continuum.
Which of the two versions turns out to be correct is an empirical question which can be answered only after an investigation of many more languages belonging to unrelated families.

2.4 Perspectives

Dauer (1987) claimed that "numerous experiments have shown that a language cannot be assigned to one or the other category on the basis of instrumental measurements of interstress intervals or syllable durations" (p 447). Does this mean then that we should abandon instrumental measurements? We favor the view that we would do better to look for more effective instrumental measurements that could account for the perception of speech rhythm. The phonological account seems perfectly acceptable and sensible, but it does not explain how rhythm is extracted from the speech signal by the perceptual system.

Indeed, a purely phonological model of rhythm fails to make a number of predictions. As we mentioned above, infants' language discrimination behavior is interpreted as relying on the stress-timed/syllable-timed dichotomy. However, unsurprisingly, only well-classified languages have been used in those experiments. How would infants classify languages such as Polish or Catalan? Could they tell them apart from both stress- and syllable-timed languages, or from neither, or from one of these groups only (and then which one)? And could they tell them apart from one another? The phonological model cannot answer these questions because it is not implemented. It does not explicitly situate languages with respect to each other on the rhythm scale, because it does not say how much each phonological feature contributes to the perception of rhythm, and how features interact with each other. As a result, it is impossible to predict whether Polish is in the middle of the continuum or, say, whether its syllabic complexity overrides its lack of vowel reduction and pushes it towards the stress-timed end. But answers to these questions are necessary, if we are to understand how infants perceive speech rhythm, how they learn the phonology of their native language, and how they can deal with any kind of bilingual environment.

In the remainder of this paper, we propose an implementation of the phonological account of speech rhythm with the aim of clarifying how rhythm may be perceived and to make predictions as to how listeners classify languages according to their rhythm.

3 Instrumental measurements in 8 languages

3.1 Rationale

Dauer (1987) observed that:

> Neither "syllable" nor "stress" have general phonetic definitions, which from the start makes a purely phonetic definition of language rhythm impossible. All instrumental studies as well as all phonological studies have had to decide in advance where the stresses (if any) fall and what a syllable is in the language under investigation in order to proceed". (pp 447-448)

Since our main interest is to explain how infants come to perceive contrasting rhythms at birth, and since the infant cannot be expected to know anything specific a priori about the language to be learned, we would like to argue that a viable account of speech rhythm should not rely on a complex and language-dependent phonological concept such as stress. We will therefore attempt to provide a purely phonetic definition of language rhythm without appealing to those concepts.

Our starting point is a hypothesis about the perception of speech by the infant. Following Mehler et al. (1996), we propose that infant speech perception is centered on vowels, because "vowels carry most of the energy in the speech signal, they last longer than most consonants, and they have greater stability. They also carry accent and signal whether a syllable is strong or weak" (p 112). In addition, there is evidence that newborns pay more attention to vowels than to consonants (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988), and that

they are able to count the number of syllables (and therefore vowels) in a word, independently of syllable structure or weight (Bertoncini, Floccia, Nazzi, & Mehler, 1995; Bijeljac-Babic, Bertoncini, & Mehler, 1993; van Ooyen, Bertoncini, Sansavini, & Mehler, 1997).
Thus we assume that the infant primarily perceives speech as a succession of vowels of variable durations and intensities, alternating with periods of unanalyzed noise (i.e., consonants), or what Mehler et al. (1996) called a Time-Intensity Grid Representation (TIGRE).
Guided by this hypothesis, we will attempt to show that a simple segmentation of speech into consonants and vowels[3] can:

- Account for the standard stress- / syllable-timing dichotomy and investigate the possibility of other types of rhythm.
- Account for language discrimination behaviors observed in infants.
- Clarify how rhythm might be extracted from the speech signal.

3.2 Material

Sentences were selected from a multi-language corpus initially recorded by Nazzi et al. (1998) and augmented for the present study (Polish and Catalan[4]). Eight languages (English, Dutch, Polish, French, Spanish, Italian, Catalan, Japanese), 4 speakers per language and 5 sentences per speaker were chosen, constituting a set of 160 utterances. Sentences were short news-like declarative statements, initially written in French, and loosely translated into the target language by one of the speakers. They were matched across languages by number of syllables (from 15 to 19), and roughly matched for average duration (about 3 seconds). Sentences were read in a soundproof booth by female native speakers of each language, and were low-pass filtered, digitized at 16 kHz and recorded directly on hard disk.

3.3 Method

The first author marked the phonemes of each sentence with sound-editing software, using both auditory and visual cues. Segments were identified and located as precisely as possible, using the phoneme inventory of each language.
Phonemes were then classified as vowels or consonants. This classification was straightforward with the exception of glides, for which the following rule was applied: Pre- and inter- vocalic glides (as in English /kwiːn/ “queen” or /vawəl/ “vowel”) were treated as consonants, whereas post-vocalic glides (as in English /haw/ “how”) were treated as vowels.
Since we made the simplifying assumption that the infant only has access to the distinction between vowel and consonant (or vowel and other), we did not measure durations of individual phonemes. Instead, within each sentence we measured:

- The duration of sequences of consecutive vowels (from onset of the first vowel to offset of the last vowel of the sequence), that we will refer to as *vocalic intervals*.
- The duration of sequences of consecutive consonants, i.e., *consonantal intervals*, or if we continue to refer to vowels, *inter-vocalic intervals*.

As an example, the phrase "next Tuesday on" (phonetically transcribed as /nɛkstjuzdeiɔn/) only has 3 vocalic and 4 consonantal intervals: /n/ /ɛ/ /kstj/ /u/ /zd/ /eiɔ/ /n/.
From these measurements we derived three variables, each taking one value per sentence:

- the proportion of vocalic intervals in the sentence, or %V.
- the standard deviation of vocalic intervals within the sentence, or ΔV.
- the standard deviation of consonantal intervals within the sentence, or ΔC. [5]

3.4 Results

Table 1 presents, for each language, the number of measurements, the average proportion of vocalic intervals (%V), and the average standard deviations of consonantal (ΔC) and vocalic (ΔV) intervals. Languages are sorted depending on %V. As can be seen, they also seem to be sorted from most to least stress-timed, which is a first indication that these measurements reflect something about rhythmic structure.

It is now possible to locate the different languages in a tridimensional space. Figures 1, 2 & 3 show the projections of the data on the (%V, ΔC), (%V, ΔV) and (ΔV, ΔC) planes. The (%V, ΔC) projection clearly seems to fit best with the standard rhythm classes. How reliable is this account? We computed an ANOVA by introducing the "rhythm class" factor (Polish, English and Dutch as stress-timed, Japanese as mora-timed and the rest as syllable-timed). For both %V and ΔC, there was a significant effect of rhythm class ($p < 0.001$). Moreover, post-hoc comparisons with a Tukey test showed that each class was significantly different from the two others, both in %V (each comparison $p < 0.001$) and ΔC ($p \leq 0.001$). No significant class effect was found with ΔV.

Thus %V and ΔC seem to support the notion of stress-, syllable- and mora-timed languages. However, ΔV suggests that there may be more to speech rhythm than just these distinctions; this variable, although correlated with the two others, rather emphasizes differences between Polish and the other languages. We will come back to this point further below.

3.5 Discussion

As mentioned earlier, this study was meant to be an implementation of the phonological account of rhythm perception. The question now is whether our measurements can be related to the phonological model.

ΔC and %V seem to be directly related to syllabic structure. Indeed, a greater variety of syllable types means that some syllables are heavier (see footnote 2). And in most languages, syllables gain weight mainly by gaining consonants. Thus, more syllable types mean more variability in the number of consonants, more variability in their overall duration in the syllable, and thus a higher ΔC. They also imply a greater consonant/vowel ratio on average, i.e. a lower %V (hence the evident negative correlation between ΔC and %V). It is therefore not surprising to find English, Dutch and Polish (more than 15 syllable types) at one end of the ΔC and %V scales, and Japanese (4 syllable types) at the other. Thus, the nice fit between the (%V, ΔC) chart and the standard rhythm classes comes as an empirical validation of the hypothesis that rhythm contrasts are accounted for by differences in the variety of syllable structures.

Can the ΔV scale be interpreted as transparently as ΔC? Not really, because several phonological factors combine with each other and influence the variability of vocalic intervals:

- Vowel reduction (English, Dutch, Catalan);
- contrastive vowel length (Dutch and Japanese);
- vowel lengthening in specific contexts (Italian);
- in addition, English and French have certain vowels that are significantly longer than others (respectively falling diphthongs and nasal vowels).

Only vowel reduction and contrastive vowel length have been described as factors influencing rhythm (Dauer, 1987), but the present analysis suggests that the other factors may do so as well. In our measurements, ΔV reflects the sum of all phenomena. As a possible consequence, the ΔV scale seems less related to the usual rhythm classes. However, it still

can be interpreted: the two languages with the lowest ΔV, Spanish and Polish, are the ones which show no phenomenon likely to increase the variability of vocalic intervals. Thus ΔV still tells us something about the phonology of languages. It remains an empirical question whether it tells us something about rhythm perception.
Supposing it does, how does it influence the global picture on language rhythms? Taking the standard rhythm classes evident on the (%V, ΔC) chart as a reference, ΔV mostly adds one piece of information, suggesting that Polish is in some respects very different from the other stress-timed languages. This finding echoes the doubts raised by Nespor (1990) about its actual status and suggests that indeed, Polish should be considered neither stress- nor syllable- (nor mora-) timed. At this stage, new discrimination experiments are clearly needed to test whether ΔV plays a role in rhythm perception.

4 Confrontation with behavioral data

Following Bertinetto (1981), Dasher and Bolinger (1982) and Dauer (1983), we have assumed that the standard rhythm classes postulated by linguists may be the result of the presence and interaction of certain phonological features in languages. We have then shown that these phonological features have reliable phonetic correlates that can be measured in the speech signal, and that these correlates can predict the rhythm classes. It follows that at least some rhythmic properties of languages can be extracted by phonetic measurements on the signal, and this finding provides us with a computational model of how types of speech rhythm could be retrieved by the perceptual system. That is, we may assume that humans segment utterances into vocalic and consonantal intervals, compute statistics such as %V, ΔC and ΔV, and associate distinct rhythm types with the different clusters of values. But could we really predict which pairs of languages can and which cannot be discriminated on the basis of rhythm? At first glance one might be tempted to say that the (%V, ΔC) chart predicts that discrimination should take place between rhythm classes, as previously hypothesized. However, specific predictions crucially depend on how much overlap there is between the different languages, and on how large the sets of sentences are (the larger, the shorter the confidence intervals for each language). And as we will see, the discrimination task that is used in a particular experiment can also play a role. This will lead us to model the discrimination tasks that are used with adults and with infants, and run simulations as faithful as possible to the real experiments.

4.1 Adults

4.1.1 Language discrimination results

Contrarily to infants, adults can be expected to use a much broader range of cues to perform the discrimination task: possibly rhythm, but also intonation, phonetics and phonotactics, recognition of known words, and more generally, any knowledge or experience related to the target languages and to language in general. In order to assess adults' ability to discriminate languages on the basis of rhythm alone, it is thus crucial to prevent subjects from using any other cues.
There are practically no studies on adults that have fulfilled this condition. A few studies have tried to degrade stimuli in order to isolate prosodic cues: Bond and Fokes (1991) superimposed noise onto speech to diminish non-prosodic information. Others have managed to isolate intonation by producing a tone following the fundamental frequency of utterances (de Pijper, 1983; Maidment, 1976; 1983; Willems, 1982). Ohala and Gilbert (1979) added rhythm to intonation by modulating the tone with the envelope of utterances. Den Os (1988),

Dehaene-Lambertz (1995) and Nazzi (1997) used low-pass filtered speech. Finally, Ramus and Mehler (1999) used speech resynthesis and manipulated both the phonemes used in the synthesis and F0.
Among all these studies, only den Os (1988) and Ramus and Mehler (1999) have used stimuli that were as close as one can get to pure rhythm. Den Os monotonized the utterances (F0 = 100 Hz) by means of LPC synthesis and then low-pass filtered them at 180 Hz. Ramus and Mehler resynthesized sentences where consonants were all replaced by /s/, vowels by /a/, and F0 was made constant at 230 Hz. However, whereas Ramus and Mehler did not disclose the target languages and tried to make it impossible for the subjects to use any other cue than rhythm, den Os tried to cue them in various ways: subjects were native speakers of one of the target languages (Dutch), and the auditory stimuli were also presented in written form, thus making the evaluation of the correspondence between rhythm of the stimuli and their transcription possible. Therefore we cannot consider that den Os' experiments assess discrimination on the basis of rhythm alone.
Thus the only relevant results for our present purpose are those of Ramus and Mehler (1999), showing that French subjects can discriminate English and Japanese sentences on the basis of rhythm only, without any external cues.

4.1.2 Modeling the task

The procedure used by Ramus & Mehler consisted in training subjects for L1/L2 categorization on 20 sentences uttered by 2 speakers per language, and then having them generalize the categorization on 20 new sentences uttered by 2 new speakers.

This procedure is formally analogous to a logistic regression. Given a numerical predictor variable V and a binary categorical variable L over a number of points, this statistical procedure finds the cut-off value of V that best accounts for the two values of L. The procedure can be applied to one half of the existing data, amounting to the training phase of the behavioral experiment, and the cut-off value thus determined can be used to predict the values of L on the other half of the data, amounting to the test phase of the experiment.

Straightforwardly, we take language, restricted to the English/Japanese pair, as the categorical variable, and %V as the numerical variable. %V was chosen rather than ΔC because it has lesser variance and therefore can be expected to yield cleaner results. 10 sentences of each language, uttered by 2 speakers per language, will be used as training set, and the 10 remaining sentences per language, uttered by other speakers, will be used as test set. The simulation will thus include the same number of sentences as the behavioral experiment. In addition, sentences used in the experiment and the simulation are drawn from the same corpus, uttered by the same speakers, and there is even some overlap between the two.

4.1.3 Results

4.1.3.1 English/Japanese

We found that the regression coefficients calculated on the 20 training sentences could successfully classify 19 of them, and 18 of the 20 test sentences (90% hit rate in the test phase). To assess any asymmetry between the training and the test sets, we redid the regression after exchanging the two sets, and we obtained a 95% hit rate in the test phase (chance is 50%). This analysis shows that it is possible to extract enough regularities (in terms of %V) from 20 English and Japanese sentences to be able to subsequently classify new

sentences uttered by new speakers, thus simulating the performance of subjects in (Ramus & Mehler, 1999).
What's more, due to the overlap between the sentences of the experiment and those of the simulation (26 sentences out of 40), it is possible to compare the subjects' and the simulation's performance sentence by sentence. Of the three sentences that were misclassified in the simulation, two were used in the experiment. It appears that these are the very two sentences that were most misclassified by subjects too, yielding respectively 38% and 43% correct classification, meaning that subjects classified them more often as Japanese than as English. This similarity between the experiment and the simulation is striking, but it rests on two sentences only.
Figure 4 shows subjects' average classification scores for the 26 common sentences against %V. This figure first shows the almost perfect separation of English and Japanese sentences along the %V dimension, thus explaining the high level of classification in the simulation. More interestingly, it appears that the lower the value of %V for a sentence, the better it is classified as English by subjects, as shown by a linear regression ($R=.87$, $p<0.001$). Thus, at least for English, it seems that %V is a good indicator of subjects' classification. Such a correlation is not apparent for Japanese. Note, however, that Japanese sentences present a much lesser variance in %V, and tend to be well-classified as a whole, except for one sentence. And it happens that this sentence has indeed the lowest %V among Japanese sentences.
This correspondence between %V and subjects' classification scores is, we think, primary evidence for the psychological plausibility of the proposed model: Subjects' results are as if they actually computed %V, and based their English/Japanese decision at a value of approximately $\%V=0.46$. On English sentences, the data moreover show an interpretable distance effect, sentences having a %V further from the decision threshold being easier to classify. Why wouldn't Japanese sentences show such an effect? Apart from the lesser variance which reduces the probability of observing the effect, we may conjecture that variation in %V among English sentences probably reflects differences in syllable complexity, whereas among Japanese sentences it may reflect differences in vowels' lengths. Since vowel length is not contrastive in French, the French subjects in (Ramus & Mehler, 1999) may have been less sensitive to these differences.

4.1.3.2 Other pairs of languages

Obviously, the classification scores presented in the previous section are much higher in the simulation than those obtained in the experiment (68% in the test phase). This is imputable to the fact that the logistic regression finds the best possible categorization function for the training set, whereas human subjects have trouble doing that, even after 3 training sessions (mean classification score after first training session: 62.5%).
This discrepancy may let one think that the logistic regression doesn't quite simulate the performance of human subjects, and would predict the discrimination of many more pairs than are actually discriminated by subjects. Here, we lack behavioral results, but in order to get a more global picture, and for the purpose of predicting future discrimination experiments on adult subjects, we also did the simulation for all other pairs of languages presented in this paper.

For all the pairs of languages, the predictor variable was %V, and the regression was performed twice, exchanging the training set and the test set in-between to avoid asymmetries. The classification score reported is thus the average of the two scores obtained in the test phase.

As can be seen in table 2, it is not the case that we predict discrimination of all the pairs of languages; very high scores are found only in the cases where Japanese is put against another language. Rather, the pattern of scores seems to follow closely the rhythm classes[6], that is, discrimination scores are always higher between classes (60% or more) than within class (less than 60%), with only one exception, that of Dutch/Spanish (between class, 57.5 %). We know of no behavioral result or prediction regarding this pair, and until we have more data, we will regard it as the exception that proves the rule.

These simulations provide quantitative predictions regarding the proportion of sentences an adult subject could be expected to classify correctly in a language discrimination task, assuming the subject has a measure of speech rhythm equivalent to %V. We can only hope to have more behavioral results in the future to compare with these predictions.

4.2 Infants

4.2.1 Language discrimination results

There are numerous reports of language discrimination experiments in infants available in the literature, using various language pairs and types of stimuli, and involving subjects of different ages and linguistic backgrounds. Table 2 presents results obtained with newborns only, because undesirable factors intervene at older ages. Indeed, after 2 months, infants seem to discriminate only between native and foreign language, due to early focusing on their native language (Christophe & Morton, 1998; Mehler et al., 1988). Moreover, there is evidence that after that age infants can perform the discrimination using other cues than rhythm, presumably intonation, and phonetics or phonotactics (Christophe & Morton, 1998; Guasti, Nespor, Christophe, & van Ooyen, in press; Nazzi & Jusczyk, submitted). Of all the results obtained with newborns, only one won't be considered here, French/Russian, since we don't have Russian in our corpus.

It should be noted that not one experiment on infants has demonstrated discrimination based on rhythm only, since the stimuli used always preserved some other type of information as well. The hypothesis that newborns base their discrimination on speech rhythm thus relies only on the pattern of discriminations found across the different pairs of languages, which is consistent with the standard rhythm classes. Success of our simulations to predict this very pattern would thus confirm the feasibility, and therefore the plausibility, of rhythm-based discrimination.

4.2.2 The discrimination task

Discrimination studies in infants report two kinds of behavior: first, recognition and/or preference for maternal language, and second, discrimination of unfamiliar languages. The first supposes that infants are already familiar with their maternal language, i.e. that they have formed a representation of what utterances in this language sound like. It is with this representation that utterances from an unfamiliar language are compared. Discrimination of unfamiliar languages does not suppose familiarization prior to the experiment. It requires, however, familiarization with one language during the experiment, as is done in habituation/dishabituation procedures. Infants then exhibit spontaneous recovery of the behavioral measure (dishabituation) when the language changes in a perceptible way.

Thus, in both cases, discrimination behavior involves forming a representation of one language, and comparing utterances from the new language with this representation. Discrimination occurs when the new utterances do not match the earlier representation. However, a crucial point of all these experiments is that infants are never told *which* language

is their native language, nor *when* a change in language occurs. That is, utterances in a new language are spontaneously discriminated from those that were previously represented, without any external sign that these utterances are new and should be compared with the earlier ones. For this reason, neither standard comparisons between sets of data nor procedures involving supervised training (like the logistic regression) can adequately model the task of the infant, since they would a priori presuppose two categories, when infants discover by themselves that there are (or not) two categories.
Here, we will try and model as closely as possible the task of the infant as it occurs in non-nutritive sucking discrimination experiments such as those of Nazzi, Bertoncini and Mehler (1998). For this purpose we will model infants' representation of sentences' rhythm, their representation of a language, and their arousal in response to sentences. We will then simulate experiments, by simulating subjects divided in an experimental and a control group, a habituation and a test phase, and sentences drawn in a random order.

4.2.3 A model of the task

Below are the main features and assumptions of the model:

- An experiment unfolds in a number of steps, each consisting of the presentation of one sentence. Here, we note steps with the index n.
- the rhythm of each sentence S_n heard by the infant at step n is represented as its $\%V_n$ value;
- in the course of an experiment, the infant forms a prototypical representation P_n of all sentences heard . Here this prototype will be taken as the average %V of all the sentences heard so far[7]: $P_n = \frac{1}{n}\sum_{i=1}^{n} \%V_i$;
- the infant has a level of arousal A_n, which is modulated by stimulation and novelty in the environment. In the experiment, all things being equal, arousal is dependent on the novelty of the sentences heard. For the simulation, we take as arousal level the distance between the last sentence heard and the former prototype. That is, at step n, $A_n = \left|\%V_n - P_{n-1}\right|$
- We further assume that there is a causal and positive correlation between arousal and sucking rates observed in experiments, that is, a rise in arousal causes a rise in sucking rate. Given this assumption, we won't model the link between arousal and sucking rate, and will assess the subject's behavior directly through arousal.

The simulation of a discrimination experiment for a given pair of languages involves:

- The simulation of 40 subjects, divided into two groups: experimental (language and speaker change), and control (same language, speaker change). The order of appearance of languages and speakers is counterbalanced across subjects. Subjects belonging to the same counterbalancing subgroup differ only with respect to the order of the sentences heard within a phase (individual variability is not modeled).
- For each subject:
  - in the habituation phase, 10 sentences uttered by 2 speakers in the assigned language are presented in a random order. P_n and A_n are calculated at each step.
  - Automatic switch to the test phase after 10 habituation sentences.
  - in the test phase, 10 new sentences uttered by 2 new speakers in the assigned language are presented in a random order. P_n and A_n are calculated at each step.
- Comparison of arousal pattern between the experimental and control groups.

There are important differences between the proposed simulations and the real experiments that deserve to be discussed:

- in the experiments, switch to the test phase follows reaching a certain habituation criterion, namely, a significant decrease in the sucking rates. This is to ensure 1) that the switch occurs at a comparable stage of every infant's sucking pattern, 2) that infants have the possibility to increase their sucking again after the switch, and thus to show dishabituation. Here, these conditions serve no purpose, because they address only the link between arousal and the sucking behavior, which we don't model. In the simulations, after presentation of the 10 habituation sentences in a given language, all the subjects have reached the same state: their prototype P_{10} is just the average of the 10 sentences' %V values, and it will not be significantly modified by presenting the same sentences again until a habituation criterion is met, as is done in the real experiments.
- in most experiments, the stimuli used consisted in longer samples of speech in each language than we have here. In Nazzi, Bertoncini and Mehler (1998), for instance, 40 sentences per language were used, when here we have only 20. But in a forthcoming study, discrimination between Dutch and Japanese was also shown in newborns using only 20 sentences per language (Ramus et al, in preparation), suggesting that 20 sentences are enough for babies to reliably represent and discriminate two languages. If anything, using only 20 sentences rather than 40 should reduce the probability of observing a significant discrimination in the simulation, since more sentences would lead to more accurate prototypes at the end of habituation.

4.2.4 Results

Simulations were run on all 26 pairs of languages studied in this paper. Discrimination was assessed by testing whether arousal was higher in the experimental group than in the control group during the test phase[8]. The dependent variable was the average arousal level over the 10 test sentences $\left(\frac{1}{10}\sum_{n=11}^{20} An\right)$ and the factor was group. We used a non-parametric Mann-Whitney test because we have no hypothesis on the distribution of arousal levels.

Significance levels of this test for all the simulations are presented in Table 4. As presented, these levels are directly comparable to each other, since the tests were computed on the same type of data and with the same number of subjects (40).

Four language pairs (marked by asterisks in the table) present a peculiar arousal pattern, in that the control group has a higher average arousal in the test phase than the experimental group. This even reaches significance in the case of Catalan/Italian. This should not, however, be interpreted as predicting a discrimination, in the contrary. The four pairs concern syllable-timed languages that are very close to each other. The arousal pattern shows that the average differences between these languages are so small that they can even be smaller than speaker differences within the same language (recall that the control group switches from 2 speakers to 2 other speakers of the same language). And these speaker differences are significantly greater than the language differences in the case of Catalan/Italian. This means that the four corresponding p values in table 4 are slightly misleading: the Mann-Whitney test performed being two-tail, they represent the probability of accepting the null hypothesis (experimental = control), but discrimination is predicted only when the alternative hypothesis (experimental > control) is accepted (a one-tail test), for which the p values must be very close to 1 in the four cases.

In order to better visualize the data, we present arousal curves for three representative pairs of languages. The figures show mean arousal values at each step of the simulation for the

experimental and control groups. Note that arousal is not defined at step 1 ($A_1 = |\%V_1 - P_0|$ and P_0 is not defined), and therefore is not shown on the charts. Switch from the habituation to the test phase occurs between steps 10 and 11. Figure 5 shows arousal curves for the English/Japanese simulation, Figure 6 for English/Spanish, and Figure 7 for Spanish/Italian, illustrating respectively a large, a moderate, and a null discrimination effects.
It appears that the simulations can successfully predict the results of all the behavioral data available (shown in boldface in table 4). Moreover, they are highly consistent with the rhythm class hypothesis. Only 2 pairs of languages do not conform to this pattern: Polish/French and Spanish/Dutch, for which no discrimination is predicted by the simulation. It is interesting to note that in simulations of adult experiments, a relatively low classification score was already predicted for Spanish/Dutch, though not for Polish/French. Although no existing behavioral result is in contradiction with these predictions, it seems to us that they are not completely compatible with the otherwise high coherence of our data. Only future research, consisting of more measurements on other samples of these languages, and of the corresponding discrimination experiments, will tell us whether these predictions reflect an idiosyncrasy of our present corpus or a more profound difference between the concerned languages.

4.2.5 Groups of languages

Nazzi, Bertoncini and Mehler (1998) have also tested discrimination between groups of languages (Table 3), and we have tried to simulate this experiment as well. We followed the design of the experiment as closely as possible. As in the real experiment, the number of speakers was reduced to 2 per language. This was done in order not to increase speaker variability, by keeping only 2 speakers per phase, as in the previous experiments. Subjects in the experimental group switch from English+Dutch to Spanish+Italian, whereas subjects in the control group switch either from Spanish+Dutch to Italian+English, or from Spanish+English to Italian+Dutch, with order of the groups of languages counterbalanced across subjects. Within a phase, sentences were drawn at random from the assigned set, irrespective of their language. As in the previous simulations, there were twice as few sentences as in the real experiment, that is 10 sentences uttered by 2 speakers per language. Note that in this experiment there is no control group *stricto sensu*, in the sense of a group having the same habituation phase as the experimental group. Indeed, subjects in the control group are presented a different combination of languages than subjects in the experimental group. Thus, there is no guarantee that the prototypes P10 will be the same for both groups at the end of habituation, and assessing discrimination through comparison of average arousal in the test phase only is not appropriate. Here, we use as dependent variable the difference in average arousal between the 9 sentences following the switch and the 9 sentences preceding it: $\frac{1}{9}\left(\sum_{n=11}^{19} A_n - \sum_{n=2}^{10} A_n\right)$ (recall that A1 is not defined).
32 subjects were simulated, the same number as in the experiment. There was a main effect of group ($p = 0.01$), showing that arousal increases significantly more when switching from one rhythm class to the other, than when switching between incoherent pairs of languages. Here again, our simulations give the same result as the experiments.

4.3 Discussion

With the exception of the French/Russian pair, which we could not simulate here, the overall pattern of success and failure to discriminate languages evident in Table 3 has been entirely simulated and predicted, on the basis of quite a simple model. This model assumes that subjects can compute the vowel/consonant temporal ratio %V and that their categorization of

sentences and languages is based on %V. In the one case where the direct comparison of categorization results for individual sentences was possible (English/Japanese in adults), subjects' scores were found to be highly consistent with predictions based on %V, comforting the psychological plausibility of this model.

The generality of the agreement between the behavioral data and the simulations is still limited in several respects:

1. by the set of languages used in the behavioral experiments on the one hand, and in the simulations on the other hand. For sure, the agreement holds only for the pairs of languages studied both in experiments and in simulations, and the next behavioral result that comes out could well disconfirm the predictions of the simulation. It is for this reason that we have provided predictions for all pairs of languages present in our corpus, not only those for which a behavioral result is already available. These predictions await further language discrimination studies, be they in adults or in newborns.
2. by the potentially infinite number of variables that can in principle be derived from the durations of vocalic and consonantal intervals. In the present paper we have derived 3 variables, and shown that one of them led to the right predictions. What about the other ones? If the pattern of behavioral results was to change or to be extended, wouldn't it always be possible to derive a variable that could fit the new pattern?

In this respect, it is reassuring that the variable used in the simulation was the most straightforward to compute from the durations, %V, and not some sophisticated ad-hoc variable. ΔC and ΔV also follow directly from the nature of the data. More importantly, all three variables are interpretable from the phonological point of view, in the sense that they are directly linked to the phonological properties supposedly responsible for speech rhythm (see section 3.5). But could ΔC and ΔV have predicted the same results as %V? As we have explained, we chose %V on the basis of 1) consistency with the rhythm classes and their phonological properties, 2) lesser variance than ΔC. From Figure 1 we can guess that ΔC would have predicted the same pattern of results, but simulations might have been less sensitive. As regards ΔV, Figure 2 suggests that this variable makes different predictions. Most notably, it suggests that Polish might be discriminable from English and Dutch. We checked this by running again both the logistic regression and the arousal pattern simulation on the English/Polish and Dutch/Polish pairs using variable ΔV. The logistic regression gave respectively 85% and 87.5% classification scores, and the arousal pattern predicted both discriminations at $p < 0.001$.

Unfortunately these pairs of languages have never been experimentally tested, so it remains an open question whether ΔV can contribute to modeling subjects' behavior, that is whether the most appropriate model should be based on %V alone, ΔV alone, or both. In the latter case, the respective weighting of the variables would be an additional parameter to adjust.

5 General discussion

Phonetic science has attempted to capture the intuitive notion that spoken languages have characteristic underlying rhythmic patterns. Researchers have proposed classifications of languages according to their rhythm type. However, although they have measured many properties the speech signal, they have failed to identify reliable acoustic characteristics of language classes. Measurements estimating the periodicity of inter-stress intervals have not helped to capture these intuitive categories, and efforts to classify languages on the basis of acoustic measures have eventually been abandoned. In this paper, we have presented measurements of the speech signal that appear to support the idea that the standard rhythm classes are meaningful categories, that not only appeal to intuitions about rhythm, but also

reflect actual properties of the speech signal in different languages. Moreover, our measurements are able to account for infant discrimination behaviors, and thus provide a better understanding as to how a coarse segmentation of speech could lead infants to classify languages as they appear to do.

What can we conclude from the data reported? Taken alone, the fact that the proportion of vocalic intervals (%V) and the variability of consonantal intervals (ΔC) in eight languages are congruent with the notion of rhythm classes does not demonstrate that spoken languages pattern into just a few such classes. At this point, we are agnostic about whether all languages can be sorted into a few stable and definite rhythmic classes. We studied only eight languages, and they were selected from those used by linguists to propose the existence of the three standard classes. Hence, it is imperative to study more languages. It is entirely conceivable that the groupings already established may dissolve when more languages are added. Indeed, by adding other languages the spaces between the three categories may become occupied by intermediary languages yielding a much more homogeneous distribution. This continuous distribution would be a challenge to the notion that languages cluster into classes and would show that it was the scarcity of data points that was suggestive of clusters rather than the way languages actually pattern. Alternatively, adding more languages to this study may uncover additional classes. This possibility is consistent with typological work by Levelt and van de Vijver (1998), who have proposed 5 classes of increasing syllable markedness (= syllable complexity). Three of these classes seem to correspond to the standard rhythm classes (Marked I, III and IV in their typology). One class (Marked II) is postulated for languages whose syllable complexity would be intermediate between syllable-timed and mora-timed languages. One more class (Unmarked) is postulated beyond Japanese (this class would consist of strictly CV languages). Since languages of the Unmarked and Marked II types are not part of our corpus, we cannot assess the relevance of these two additional classes, but it is a fact that on Figure 1 for instance, there seems to be space for a distinct class between Catalan and Japanese, and of course there's also space for another class beyond Japanese. Using another rationale, Auer (1993) has also proposed five rhythm classes, which seem to only partially overlap with those mentioned above. Given all these considerations, we believe that the notion of three distinct and exclusive rhythm classes has not yet been definitively proven, but rather is the best description of the current evidence.

There are additional reasons that incline us to pursue this line of research. First, it seems that well-organized motor sequences require precise and predictable timing (Lashley, 1951). Language is a very special motor behavior but there is every reason to expect it to have a rhythmical organization comparable to that found in other motor skills such as walking or typing. Why should a language like English have such a very different periodic organization from, say, Japanese? Could it be that language has to conform to a basic rhythm that can be modulated by the adjustment of a few settings? It is too early to answer this question. But at least there are reasons to think that the temporal organization of language, like that of every other rhythmic activity, should not be arbitrary. And as for virtually every linguistic property showing variation across languages, we may expect rhythmic organization to take a finite number of values.

Moreover, if putative rhythmic classes existed they would comfort theorists who postulate that phonological bootstrapping is an essential ingredient of language acquisition (Christophe & Dupoux, 1996; Gleitman & Wanner, 1982; Mehler et al., 1996; Morgan, 1986; Pinker, 1984). Indeed, if all languages could be sorted into a few rhythm classes, the likelihood that the properties underlying the classes might be cues that allow for the setting of grammatical parameters would be increased. Correlations between rhythm type and some syntactic parameters have been proposed in the past. For instance, syllable-timed languages have been

associated with Complement/Head word order and prefixing, while stress-timed languages have been associated with Head/Complement and suffixing (Donegan & Stampe, 1983). Even though such correlations would be of enormous help in language acquisition, they unfortunately do not seem to hold beyond the languages studied by the authors (see Auer, 1993 for a more thorough critique, and Nespor, Guasti, & Christophe, 1996 for more plausible cues to word order). More plausibly, rhythm type could help acquiring other phonological properties that are less evident in the surface[9]. A number of properties have been proposed to be more or less connected with rhythm: vowel reduction, quantity contrasts, gemination, the presence of tones, vowel harmony, the role of word accent and of course syllable structure (Dauer, 1987; Donegan & Stampe, 1983; see Auer, 1993 for a survey). Given the current state of knowledge, we believe that only syllable structure can be considered associated with rhythm reliably enough to make it a guiding element in acquisition.

Precursory formulations of this notion proposed that speakers of Japanese represent speech in moras, speakers of French in syllables, and speakers of English in feet, and that for every given language rhythm allows the infant to set the correct representation unit (Cutler, Mehler, Norris, & Segui, 1983; Cutler, Mehler, Norris, & Segui, 1986; Cutler & Otake, 1994; Mehler et al., 1996). This should give way, we think, to the more general notion that each language has principles governing the structure that its syllables may take. We hypothesize that an early determination of rhythm type may allow the child to set some of these principles[10]. This hypothesis can be formulated more explicitly using the formalism of the current linguistic theories.

Within the Principles & Parameters theory (Chomsky, 1981), syllable structure is described by the values taken by binary parameters such as Complex Nucleus, Obligatory Onset, Complex Onset, Coda, Complex Coda (Blevins, 1995). The available evidence suggests that 1) mora-timed languages have (- Complex Onset) and (- Complex Coda), 2) both stress-timed and syllable-timed languages have (+ Complex Onset) and (+ Coda), 3) stress-timed languages have (+ Complex Coda). Thus, a child in presence of a mora-timed language can set two parameters at once: Complex Onset and Complex Coda. Similarly, in presence of a syllable-timed language, he/she can set both Complex Onset and Coda. Finally, in presence of a stress-timed language, the child can set Complex Onset, Coda and Complex Coda. In addition to the deterministic setting of one or two parameters, it is possible that rhythm type may impose constraints on the possible combinations of parameters.

Within Optimality Theory (Prince & Smolensky, 1993), syllable structure is described by the ordering of structural constraints like Onset, No-Coda, *Complex-Onset (* stands for No), *Complex-Coda, and faithfulness constraints like Fill and Parse. Syllable complexity (markedness), is reflected in the level of the faithfulness constraints with respect to the structural constraints (Levelt & van de Vijver, 1998). Each level of the faithfulness constraints correspond to a class of languages sharing the same markedness of syllable structure, hence the 5 classes of languages mentioned above. Thus, regardless of whether there are actually 3 or 5 such classes, it appears that knowing the type of rhythm can enable to set the faithfulness constraints right at the appropriate level.

Although the acquisition scenarios described above remain speculative, they provide a set of hypotheses that can be tested by studying in greater detail syllable processing in infants. Thus, in addition to clarifying a few questions in phonetic theory, and providing a basis for the interpretation of language discrimination capacities in infants, we hope that the present work can help uncovering some of the mechanisms through which infants acquire the phonology and possibly other properties of their native language.

Acknowledgements

This work was supported by the Délégation Générale pour l'Armement and the Human Frontiers Science Program. MN thanks the University of Amsterdam for allowing her a sabbatical leave during which the study was done. We thank Christophe Pallier for extensive advice on the analysis of the data, Emmanuel Dupoux and Anne Christophe for their suggestions, and Sharon Peperkamp and Susana Franck for comments on an earlier version of this paper.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.

Allen, G. D. (1975). Speech rhythm: its relation to performance and articulatory timing. *Journal of Phonetics, 3*, 75-86.

Auer, P. (1993). *Is a rhythm-based typology possible? A study of the role of prosody in phonological typology* (KontRI Working Paper 21): Universität Hamburg.

Bahrick, L. E., & Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant Behavior and Development, 11*, 277-296.

Bertinetto, P. (1989). Reflections on the dichotomy "stress" vs. "syllable-timing". *Revue de Phonétique Appliquée, 91-93*, 99-130.

Bertinetto, P. M. (1981). *Strutture prosodiche dell' italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici*. Firenze: Accademia della Crusca.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*(1), 21-33.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech, 38*, 311-329.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant perception. *Infant Behavior and Development, 4*, 247-260.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology, 29*, 711-721.

Blevins, J. (1995). The syllable in phonological theory. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 206-244). Cambridge: Blackwell.

Bolinger, D. (1965). Pitch accent and sentence rhythm, *Forms of English: Accent, morpheme, order* . Cambridge, MA: Harvard University Press.

Bond, Z. S., & Fokes, J. (1991). *Identifying foreign languages.* Paper presented at the XIIth International Congress of Phonetic Sciences.

Borzone de Manrique, A. M., & Signorini, A. (1983). Segmental durations and the rhythm in Spanish. *Journal of Phonetics, 11*, 117-128.

Bosch, L., & Sebastián-Gallés, N. (1997). Native language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition, 65*, 33-69.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Christophe, A., & Dupoux, E. (1996). Bootstrapping lexical acquisition: the role of prosodic structure. *The Linguistic Review, 13*, 383-412.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science, 1*(2), 215-219.

Cooper, G., & Meyer, L. B. (1960). *The rhythmic structure of music*. Chicago: University of Chicago Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language specific comprehension strategy. *Nature, 304*, 159-160.

Cutler, A., Mehler, J., Norris, D., & Segui, 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. G., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology, 24*, 381--410.

Cutler, A., & Otake, T. (1994). Mora or phoneme - further evidence for language-specific listening. *Journal of Memory and Language, 33*, 824-844.

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

Dauer, R. (1987). *Phonetic and phonological components of language rhythm.* Paper presented at the XIth International Congress of Phonetic Sciences, Tallinn.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51-62.

de Pijper, J. R. (1983). *Modelling British English intonation*. Dordrecht - Holland: Foris.

Dehaene-Lambertz, G. (1995). *Capacités linguistiques précoces et leurs bases cérébrales.* Unpublished doctoral dissertation, EHESS, Paris.

Dehaene-Lambertz, G., & Houston, D. (1998). Faster orientation latencies toward native language in two-month old infants. *Language and Speech, 41*(1), 21-43.

den Os, E. (1988). *Rhythm and tempo of Dutch and Italian.* Unpublished doctoral dissertation, Rijksuniversiteit, Utrecht.

Donegan, P. J., & Stampe, D. (1983). *Rhythm and the holistic organization of language structure.* Paper presented at the CLS Parasession on The Interplay of Phonology, Morphology and Syntax.

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception, 10*(3), 343-378.

Gleitman, L., & Wanner, E. (1982). The state of the state of the art. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3-48). Cambridge UK: Cambridge University Press.

Guasti, M. T., Nespor, M., Christophe, A., & van Ooyen, B. (in press). Pre-lexical setting of the head-complement parameter through prosody. In J. Weissenborn & B. Höhle (Eds.), *Signal to Syntax II* .

Hayes, B., & Puppel, S. (1985). On the rhythm rule in Polish. In H. van der Hulst & N. Smith (Eds.), *Advances in nonlinear phonology* (pp. 59-81). Dordrecht: Foris.

Ladefoged, P. (1975). *A course in phonetics*. New York: Harcourt Brace Jovanovich.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112-136). New York: Wiley.

Lea, W. A. (1974). *Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding* (Univac Report PX10791). St Paul, Minnesota: Sperry Univac.

Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Perception, 1*, 229-252.

Levelt, C., & van de Vijver, R. (1998, 11-12/06/1998). *Syllable types in cross-linguistic and developmental grammars.* Paper presented at the Third Biannual Utrecht Phonology Workshop, Utrecht.

Lloyd James, A. (1940). *Speech signals in telephony*. London.

Maidment, J. A. (1976). Voice fundamental frequency characteristics as language differentiators. *Speech and hearing: Work in progress, University College London*, 74-93.

Maidment, J. A. (1983). Language recognition and prosody: further evidence. *Speech, hearing and language: Work in progress, University College London, 1*, 133-141.

Mascaró, J. (1976). *Catalan Phonology and the phonological cycle.* Unpublished Doctoral, MIT, Cambridge.

Mehler, J., & Christophe, A. (1995). Maturation and learning of language during the first year of life. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 943-954): Bradford Books / MIT Press.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan & K. Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition* (pp. 101-116). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*, 495-500.

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

Nazzi, T. (1997). *Du rythme dans l'acquisition et le traitement de la parole.* Unpublished doctoral dissertation, Ecole des Hautes Etudes en Sciences Sociales, Paris.

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

Nazzi, T., & Jusczyk, P. W. (submitted). Discriminating languages from the same rhythmic class: Data from 5-month-old English-learners. .

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

Nespor, M., Guasti, M. T., & Christophe, A. (1996). Selecting word order: The rhythmic activation principle. In U. Kleinhenz (Ed.), *Interfaces in Phonology* (pp. 1-26). Berlin: Akademie Verlag.

O'Connor, J. D. (1965). *The perception of time intervals* (Progress Report 2): Phonetics Laboratory, University College London.

Ohala, J. J., & Gilbert, J. B. (1979). Listeners' ability to identify languages by their prosody. In P. Léon & M. Rossi (Eds.), *Problèmes de prosodie* (Vol. II, pp. 123-131). Ottawa: Didier.

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

Pinker, S. (1984). *Language learnability and language development*. Cambridge: Harvard University Press.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora-timing in Japanese. *Journal of the Acoustical Society of America, 81*(5), 1574-1585.

Prince, A., & Smolensky, P. (1993). *Optimality theory: constraint interaction in generative grammar* (TR-2). New Brunswick: Rutgers University.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America, 105*(1), 512-521.

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (Ed.), *Linguistic controversies* . London: Edward Arnold.

Rubach, J., & Booij, G. E. (1985). A grid theory of stress in Polish. *Lingua, 66*, 281-319.

Shen, Y., & Peterson, G. G. (1962). Isochronism in English. *University of Buffalo Studies in Linguistics - Occasional papers, 9*, 1-36.

Steever, S. B. (1987). Tamil and the Dravidian languages. In D. Comrie (Ed.), *The world's major languages* (pp. 725-746). Oxford: Oxford University Press.

van Ooyen, B., Bertoncini, J., Sansavini, A., & Mehler, J. (1997). Do weak syllables count for newborns? *Journal of the Acoustical Society of America, 102*(6), 3735-3741.

Wenk, B., & Wiolland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*, 193-216.

Wheeler, M. (1979). *Phonology of Catalan*. Oxford: Blackwell.

Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht - Holland: Foris.

Footnotes

[1] The pairs of languages tested were respectively French/Russian and English/Italian, English/Spanish, French/English, English/Japanese. Bosch and Sebastián-Gallés (1997) also showed that Spanish/Catalan discrimination was possible in four-month-olds, but rhythm was not thought to be the critical cue.

[2] This relies on the assumption that the syllable inventory of a language always starts with the most simple syllables (with the exception of V which is not legal in all languages). We are not aware of any language that would have complex syllables without having the more simple ones, and this is indeed excluded by phonological theories (see for example Blevins, 1995).

[3] We are aware, of course, that the consonant/vowel distinction may vary across languages, and that a universal consonant/vowel segmentation may not be without problems. We assume that our hypothesis should ultimately be formulated in more general terms, e.g., in terms of highs and lows in a universal sonority curve. We consider, however, that for a first-order evaluation of our approach, and given the languages we consider here, such problems are not crucial.

[4] We thank Laura Bosch and Núria Sebastián-Gallés at the University of Barcelona for recording the Catalan and Spanish material for us.

[5] %V was calculated as the total duration of vocalic intervals in the sentence divided by the total duration of the sentence (*100). Note that %C is isomorphic to %V, and thus needs not be taken into consideration. ΔV and ΔC are the standard deviations of the means of vocalic and consonantal intervals by sentence. In Table 1 these were averaged by language.

[6] It should be noted that the scores given for Catalan/Italian and Catalan/Spanish (respectively 35 and 37.5%) do not reflect discrimination through mislabeling. They rather suggest that these pairs are so close that there may be more important rhythmic differences between some speakers than between the languages.

[7]A more realistic model could implement a limited memory, storing, say, the last 10 sentences. Here, as the number of sentences is low anyway (10 in each phase), this would hardly make a difference.

[8]As we have explained in the preceding section, both groups have attained the same average prototype at the end of habituation. It is thus not necessary to take into account the arousal level at the end of habituation, through a subtraction or a covariance analysis, as is done when analyzing sucking experiments. Such a procedure could only add more noise to the analysis in the present case.

[9]Such a process would probably not be called phonological bootstrapping by those who coined the term, who meant bootstrapping of syntax through phonology.

[10]It should be noted that syllable structure is not necessarily transparent in the surface: before the infant can actually parse speech into syllables, syllable boundaries are evident only at prosodic phrase boundaries, which gives only partial and dissociated evidence as to which onsets, nuclei and codas are allowed.

Table 1.

Total number of measurements, proportion of vocalic intervals (%V), standard deviation of vocalic intervals over a sentence (ΔV), standard deviation of consonantal intervals over a sentence (ΔC), averaged by language, and their respective standard deviations. ΔV, ΔC and their respective standard deviations are shown multiplied by 100 for ease of reading.

Languages	*Vocalic intervals*	*Consonantal intervals*	*%V (StDev)*	*ΔV (StDev) (* 100)*	*ΔC (StDev) (* 100)*
English	307	320	40.1 (5.4)	4.64 (1.25)	5.35 (1.63)
Polish	334	333	41.0 (3.4)	2.51 (0.67)	5.14 (1.18)
Dutch	320	329	42.3 (4.2)	4.23 (0.93)	5.33 (1.5)
French	328	330	43.6 (4.5)	3.78 (1.21)	4.39 (0.74)
Spanish	320	317	43.8 (4)	3.32 (1)	4.74 (0.85)
Italian	326	317	45.2 (3.9)	4.00 (1.05)	4.81 (0.89)
Catalan	332	329	45.6 (5.4)	3.68 (1.44)	4.52 (0.86)
Japanese	336	334	53.1 (3.4)	4.02 (0.58)	3.56 (0.74)

Table 2

Simulation of adult discrimination experiments for the 26 pairs of languages. Scores are classification percentages on the test sentences obtained from logistic regressions on the training sentences. %V is the predictor variable. Chance is 50%.

	English	*Dutch*	*Polish*	*French*	*Italian*	*Catalan*	*Spanish*
Dutch	57.5						
Polish	50	57.5					
French	60	60	65				
Italian	65	62.5	65	55			
Catalan	65	62.5	65	57.5	35		
Spanish	62.5	57.5	62.5	50	50	37.5	
Japanese	92.5	92.5	95*	90	90	87.5	95*

* In these cases one of the two regressions failed to converge, meaning that the solution of the regression was not unique. This happens when the predictor variable allows to completely separate the sentences of the two languages (100% classification on the training set). Only the classification percentage of the regression that converged is reported in the table.

Table 3

Language discrimination results in 2 - 5 day-old infants.

Language Pair	*Discrimination*	*Stimuli*	*Reference*
French/Russian	Yes	Normal and filtered[a]	(Mehler et al., 1988)
English/Italian	Yes	Normal	(Mehler et al., 1988; see reanalysis by Mehler & Christophe, 1995)
English/Spanish	Yes	Normal	(Moon et al., 1993)
English/Japanese	Yes	Filtered[a]	(Nazzi et al., 1998)
English/Dutch	No	Filtered[a]	(Nazzi et al., 1998)
Dutch/Japanese	Yes	Resynthesized[b]	Ramus, in preparation
Spanish/Catalan	No	Resynthesized[b]	Ramus, in preparation
English+Dutch vs. Spanish+Italian	Yes	Filtered[a]	(Nazzi et al., 1998)
English+Spanish vs. Dutch+Italian or English+Italian vs. Dutch+Spanish	No	Filtered[a]	(Nazzi et al., 1998)

[a] Stimuli were low-pass filtered at 400 Hz.

[b] Stimuli were resynthesized in such a manner as to preserve only broad phonotactics and prosody (see Ramus & Mehler, 1999).

Table 4.

Simulation of infant discrimination experiments for the 26 pairs of languages. Statistical significance is shown for Mann-Whitney tests of the group factor over 40 subjects. Pairs for which behavioral data is available are shown in boldface.

	English	*Dutch*	*Polish*	*French*	*Italian*	*Catalan*	*Spanish*
Dutch	**p=0.18**						
Polish	p=1	p=0.84					
French	p<0.001	p=0.02	p=0.18				
Italian	**p<0.001**	p=0.006	p=0.02	p=0.68*			
Catalan	p<0.001	p<0.01	p=0.007	p=0.51	p=0.04*		
Spanish	**p=0.006**	p=0.21	p=0.04	p=0.97*	p=1	**p=0.68***	
Japanese	**p<0.001**	**p<0.001**	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001

* For these pairs of languages, the control group was above the experimental group. See text.

Figure Captions

Figure 1. Distribution of languages over the (%V, ΔC) plane. Error bars represent +/-1 standard error.
Figure 2. Distribution of languages over the (%V, ΔV) plane. Error bars represent +/-1 standard error.
Figure 3. Distribution of languages over the (ΔV, ΔC) plane. Error bars represent +/-1 standard error.
Figure 4. English/Japanese discrimination in adults. Average classification scores for individual sentences across subjects, plotted against their respective V% values.
Figure 5. Simulated arousal pattern for English/Japanese discrimination. 20 subjects per group. Error bars represent +/-1 standard error.
Figure 6. Simulated arousal pattern for English/Spanish discrimination. 20 subjects per group. Error bars represent +/-1 standard error.
Figure 7. Simulated arousal pattern for Spanish/Italian discrimination. 20 subjects per group. Error bars represent +/-1 standard error.

Figure 1

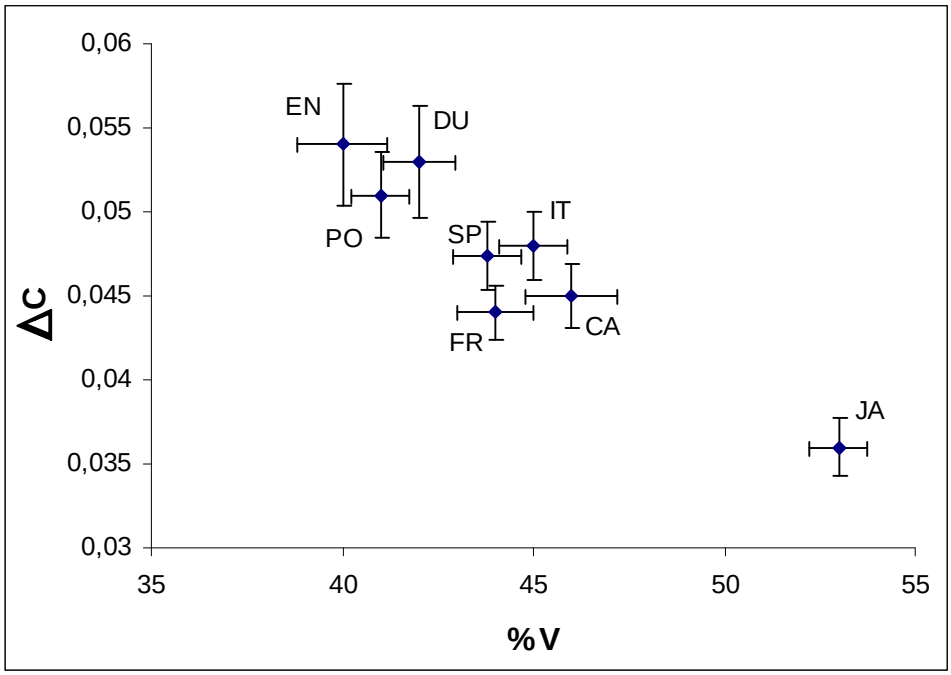

0,06
0,055
0,05
0,045
0,04
0,035
0,03
ΔC
35
40
45
50
55
%V
EN
DU
PO
SP
IT
FR
CA
JA

Figure 2

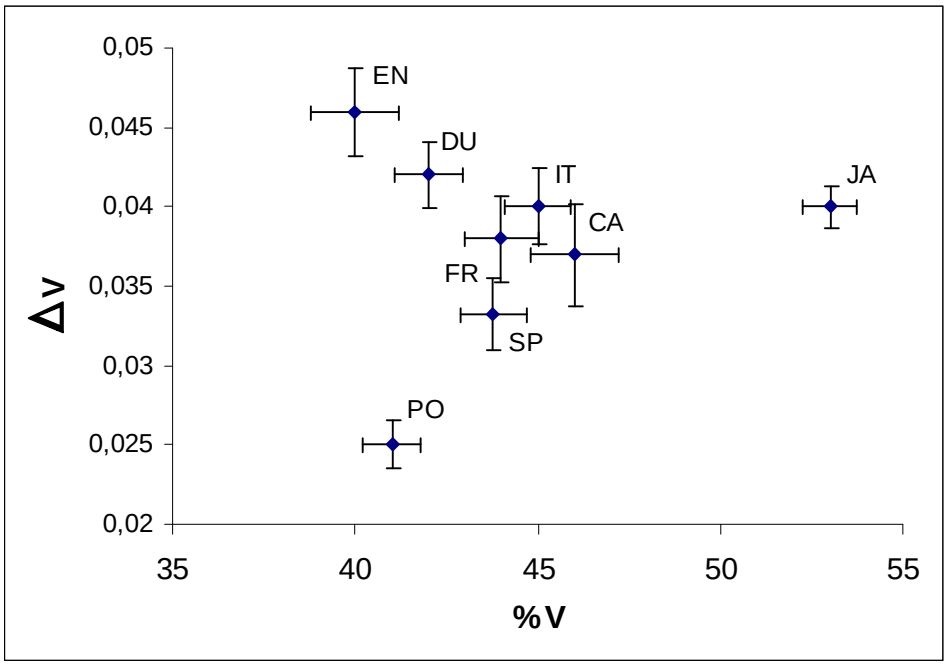

0,05
0,045
0,04
0,035
0,03
0,025
0,02
Δv
35
40
45
50
55
%V
EN
DU
IT
JA
CA
FR
SP
PO

Figure 3

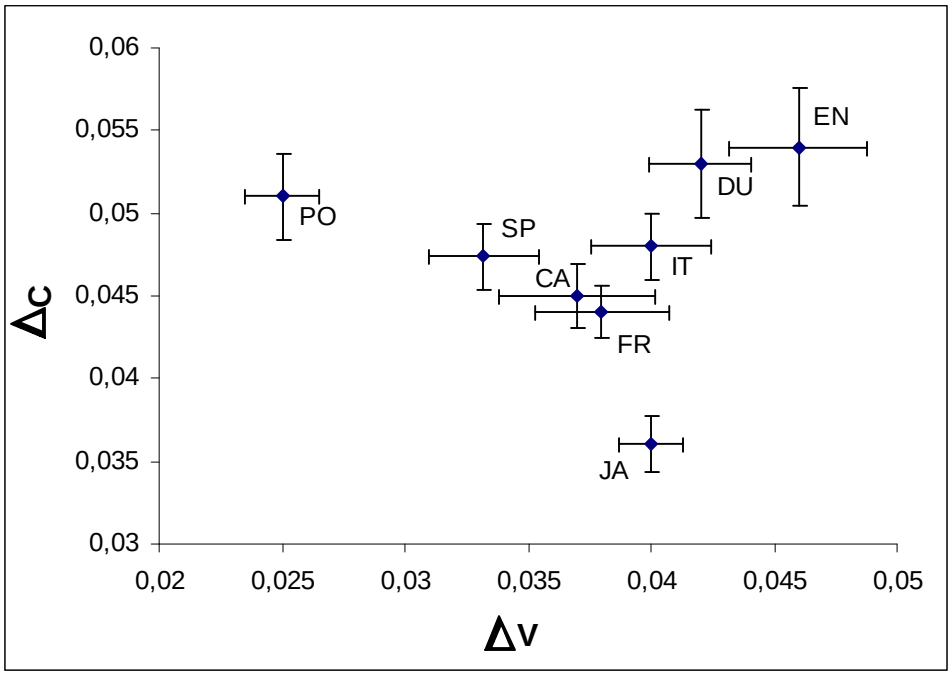

0,06
0,055
0,05
0,045
0,04
0,035
0,03
ΔC
0,02
0,025
0,03
0,035
0,04
0,045
0,05
ΔV
PO
SP
CA
FR
IT
DU
EN
JA

Figure 4

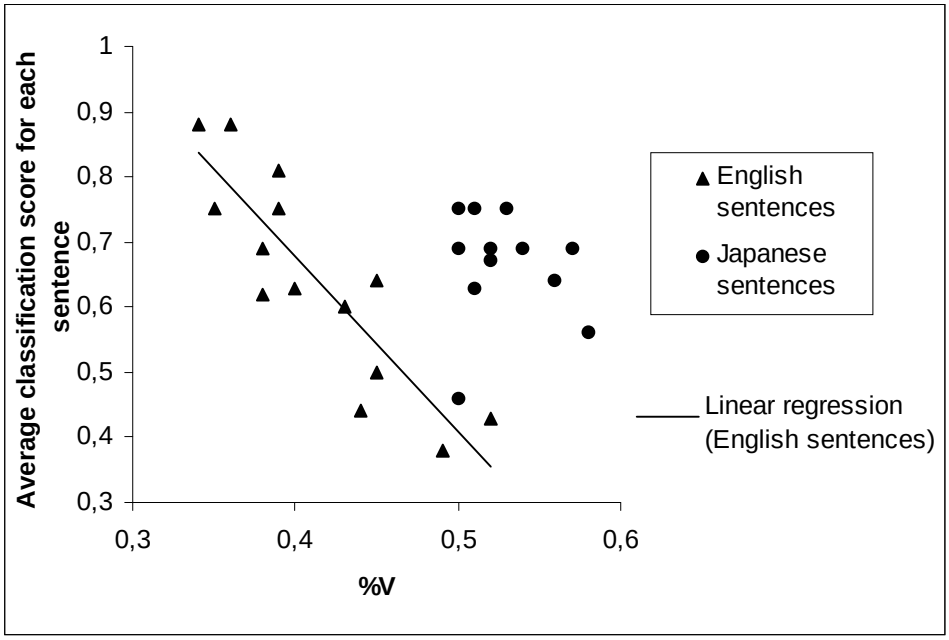

Average classification score for each sentence
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,3
0,4
0,5
0,6
%V
English sentences
Japanese sentences
Linear regression (English sentences)

Figure 5

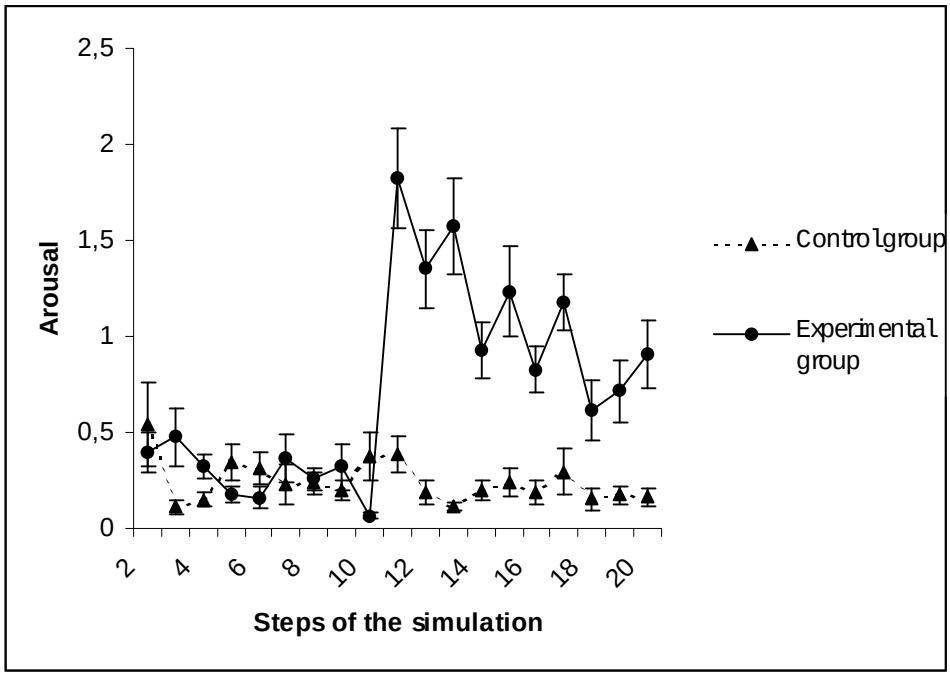

2,5
2
1,5
1
0,5
0
Arousal
2
4
6
8
10
12
14
16
18
20
Steps of the simulation
Controlgroup
Experimental group

Figure 6

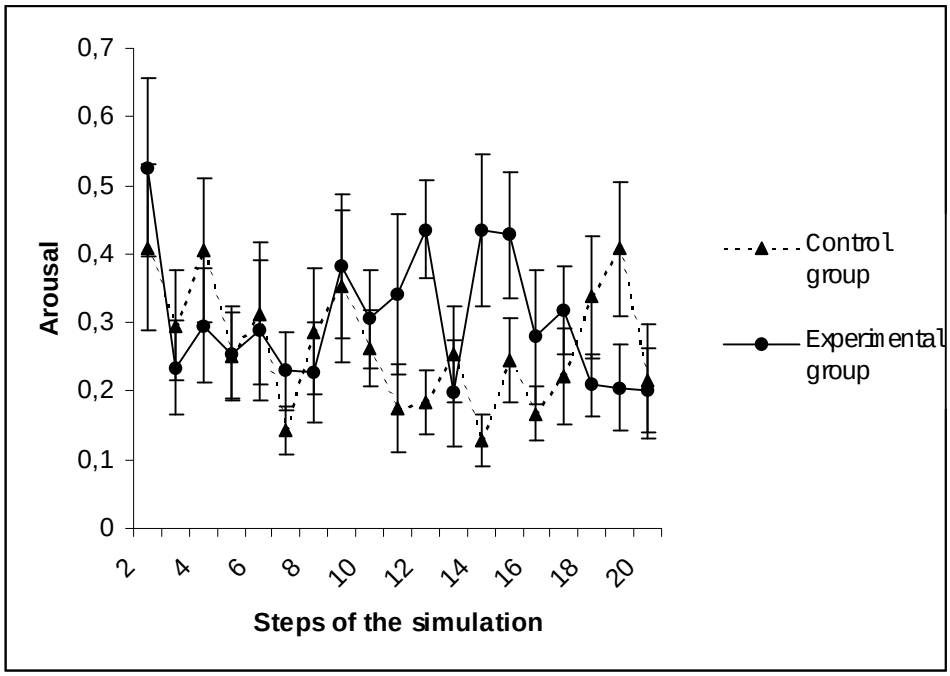
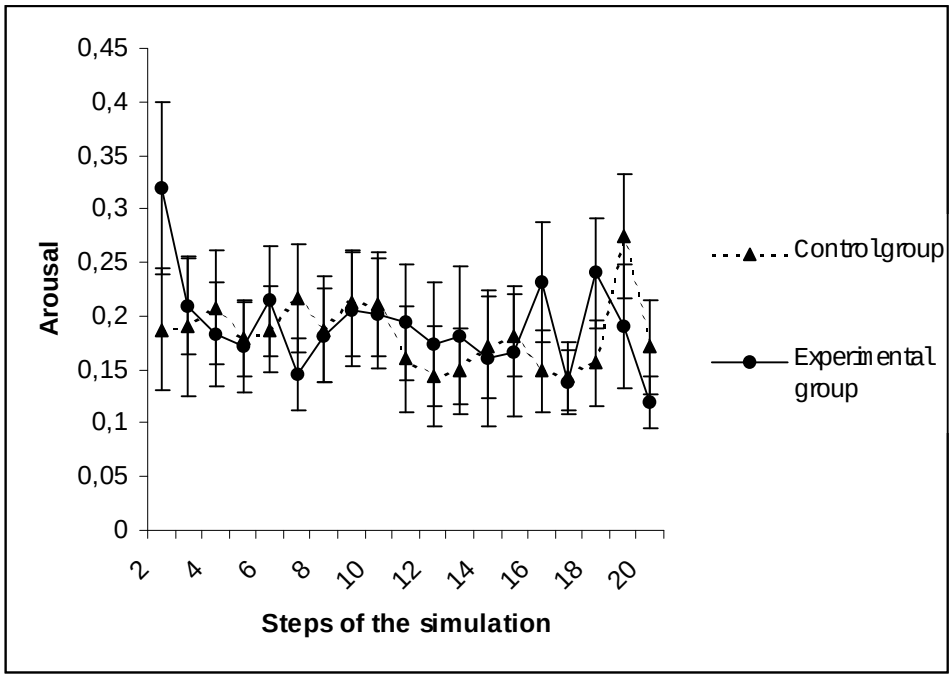

0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Arousal
2
4
6
8
10
12
14
16
18
20
Steps of the simulation
Control group
Experimental group

Figure 7


0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
Arousal
2
4
6
8
10
12
14
16
18
20
Steps of the simulation
Controlgroup
Experimental group