INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 83–94, 2003

Prosodic Phrasing: Machine and Human Evaluation

M. CÉU VIANA
Centro de Linguística da Universidade de Lisboa (CLUL), Av. Prof. Gama Pinto, no. 2, Lisboa, Portugal
mcv@clul.ul.pt

LUÍS C. OLIVEIRA
Instituto de Engenharia de Sistemas e Conputadores—Investigão e Desenvolvimento em Lisboa (INESC ID/IST), Rua Alves Redol no. 9, 1000-029 Lisboa, Portugal
lco@l2f.inesc-id.pt

ANA I. MATA
Centro de Linguística da Universidade de Lisboa (CLUL), Av. Prof. Gama Pinto, no. 2, Lisboa, Portugal
aim@clul.ul.pt

Abstract. This paper describes a set of experiments aiming at the construction and evaluation of a new phrasing module for European Portuguese text-to-speech synthesis, using classification and regression trees learned from hand-labelled texts. Using the assessment criteria of matching boundary predictions against the corresponding labelled ones, the best solution achieves an overall performance of 91.9%, with 86.3% of correctly assigned breaks and 4.3% of false insertions. Although in absolute terms such scores may be considered surprisingly good given the size of the training set, the total number of exact matches at the sentence level is much lower (22%). This suggested a more formal experiment to test the acceptability of the predicted phrasing in the judgement of human evaluators. As the model was not trained on a labelled speech corpus but on hand-labelled texts, the reference phrasing needed also to be assessed. The evaluation experiment involved 90 participants who were asked to grade both the automatic and the reference phrasings, and also to express their opinion on where the breaks should be placed. As expected, the results showed a large variability among the subjects in their acceptance of a specific sentence partition, and criteria had to be defined to summarise the data from the different evaluators. With the adopted criteria, the performance of the automatic assignment procedure at the sentence level is better rated by human evaluators than by simple matching with the reference corpus (78% vs. 22%, respectively).

Keywords: prosodic phrasing, speech synthesis, evaluation

1. Introduction

Linguistic theory posits a hierarchy of nested prosodic phrasing levels above the word, which are domains for sandhi rules and manifest themselves more or less directly in the speech signal in terms of $F0$, duration patterns, location of pauses, etc. (Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988; Selkirk, 1984, 1986; Nespor and Vogel, 1986; Ladd, 1996). Although the number and designation of phrasing levels differ from author to author and different hypotheses have been presented concerning the number of levels needed to account for tonal as well as durational patterns, there is common agreement that prosodic structures are not fully congruent with syntactic ones. They are generally flatter and cannot be predicted using syntactic information only: semantics and discourse structure, as well as rhythmic constraints, play an important role (Gee and Grosjean, 1983).

Previous attempts have been made to incorporate such types of information in TTS systems using more or less elaborated parsing strategies (Bachenko and Fitzpatrick, 1990). However, many rule-based or statistically based TTS systems have achieved satisfactory results with much simpler phrasing algorithms (Silverman, 1987; Wang and Hirschberg, 1992; Hirschberg and Prieto, 1996; Taylor and Black, 1998).

In order to build a phrasing module for a new version of the DIXI system (Oliveira et al., 1991) in the Festival framework, we closely followed Taylor and Black (1998). In line with Wang and Hirschberg (1992) and Hirschberg and Prieto (1996), classification and regression tree (CART) techniques (Breiman et al., 1984) were adopted and all the experiments were performed on hand-annotated text, instead of hand-labelled recorded speech. As shown in Hirschberg and Prieto (1996), text-based methods considerably speed up the process of building new phrasing modules or updating existing ones, and the resulting decision trees may reach equivalent or even slightly higher cross-validation scores.

Models based on self-learning procedures have some well-known advantages over rule-based ones. They can be easily retrained and tested as more and more annotated speech material becomes available or their quality is improved. They may be also used as an efficient method to determine which variables are linguistically meaningful and what is their relative weight. They are thus particularly useful in the case of languages like European Portuguese, for which large annotated speech corpora are still under construction. Moreover, there is not enough knowledge about the most relevant features for annotating different prosodic events or about the way they are interrelated with each other. In this sense, CART techniques appear as the most natural choice among the available self-learning procedures, since the resulting trees are easier to read and can be manually modified.

Phrasing models are often evaluated by counting the number of times the predicted value for every word boundary matches the corresponding labelled value in a test corpus. Although this type of evaluation is crucial for model optimisation, the resultant performance scores may be misleading, since some sentences may be uttered with more than one acceptable pattern. Therefore, a non-perfect match does not necessarily mean that the prediction is wrong, and other assessment methods are needed for a more realistic evaluation. A possible solution consists in having the predicted phrasing for a written test corpus validated by linguists trained in the assignment of prosodic structure (Hirschberg and Prieto, 1996). The evaluation experiment presented in this paper uses a similar strategy, assuming, however, that non-expert native speakers are able to rate subjectively the predicted phrasings, as well as to indicate where they would place breaks when reading the text aloud. For a small set of sentences, the results from both experts and non-experts were compared with hand-labelled versions of the same sentences produced by two professional speakers.

This paper is organised as follows: Section 2 describes the reference corpus and the annotation procedure. Section 3 presents the experiments aiming at the construction of the phrasing model using CART techniques. After obtaining a reasonable performance with the automatic phrasing system, an evaluation tool was developed and a test was carried out to verify the acceptability of both the reference and the predicted phrasing in the judgement of human evaluators. The evaluation procedure, as well as the test results and the above mentioned comparison are described in Section 4. A brief summary and future work directions are presented in Section 5.

2. Data and Methods

In all the experiments described below, a selection of 35 written texts was used, covering a variety of genres: excerpts from school textbooks, novels, press articles and interviews, letters, cooking recipes, traditional rhymes and popular jokes. These were collected in the scope of another CLUL project to exemplify the heterogeneity of language uses for teaching purposes. This heterogeneity is a desirable feature for the development of a TTS system ideally designed to deal with unrestricted text. All the texts were read by two professional speakers (a male and a female) and recorded with the same protocol used for other speech corpora collected by the group. Manual annotation, however, was restricted to a small subset.

Two of the authors, both European Portuguese (EP) native speakers and linguists experienced with the prosodic annotation of speech corpora, each hand-labelled half of the selected written materials for prosodic boundaries, and checked the other half. Three types of junctures were considered, which roughly correspond to ToBI (Silverman et al., 1992; Beckman and Elam, 1997; Beckman and Hirschberg, 1994) levels

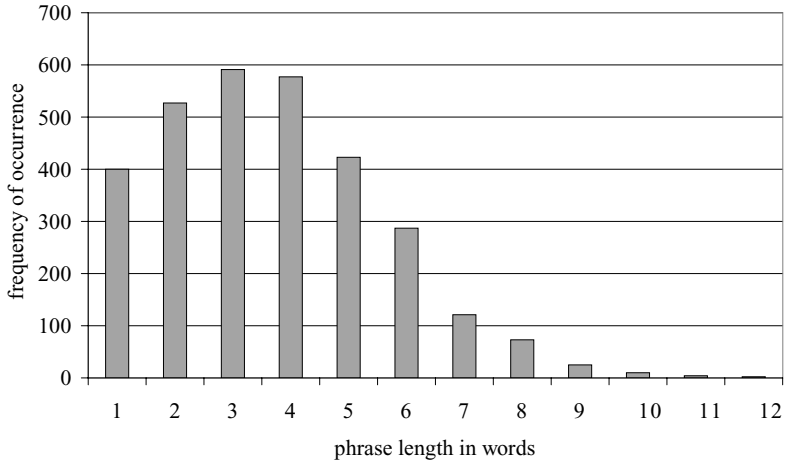

Figure 1. Distribution of phrase lengths in the reference data.

1, 3 and 4 (no break, minor or intermediate break, and major or intonational break, respectively). Breaks were assigned paying attention to punctuation and in accordance with what the annotators believed to be appropriate intonational boundary locations in a slow but fluent and smooth oral reading. Since the decisions concerning the presence or absence of a break at a given word boundary may be treated independently of the strength of that boundary, levels 3 and 4 were collapsed into a single category.

The resultant reference corpus has 11167 word boundaries, 27% of which were marked as breaks by the annotators. This parsing gave a phrase length distribution plotted in Fig. 1 with an average length of 3.8 words per phrase. Part-of-speech (POS) tags were also annotated, as well as other types of information (sentence length in words and distance measures from the boundary to previous and following punctuation marks). POS tags were obtained with Palavroso, an EP morphological analyser developed by INESC, whose output has been checked for ambiguities and manually corrected. The distance measures were automatically extracted from the raw text.

The Edinburgh Speech Tools were used for statistical modelling. The resulting CARTs can be easily integrated into the Festival framework in which the new version of the DIXI system (Oliveira et al., 1991) is currently being developed.

As most of the phrasing methods considered in the following experiments were also tested for English by Taylor and Black (1998), similar performance measures were adopted for the sake of comparison:

Correct Breaks: $CB = \frac{B-M}{B} \times 100\%$
Correct Junctures: $CJ = \frac{N-M-I}{N} \times 100\%$
False Insertions: $FI = \frac{I}{N} \times 100\%$
Missing Break: $MB = \frac{M}{N} \times 100\%$

where N is the total number of word boundaries in the test corpus, B is the total number of boundaries with a break in the test corpus, I is the number of times a break is predicted but there is no break in the test corpus and M is the number of times no break is predicted but there is a break in the test corpus. Note that while the CB score only gives credit to breaks correctly predicted, the CJ score accounts both for breaks and non-breaks correctly predicted and is, thus, sensitive to the ratio between breaks and non-breaks in a text.

Given the limited size of the reference corpus, it was randomly divided to allow for five-fold cross-validation estimates using 80% of the data for training and the remaining 20% for testing.

3. Experiments and Results

In accordance with the Festival recommendations, the experiments described in this section account for a progression in complexity which is useful in the development of TTS systems for new languages. They can be used at different stages and in accordance with the availability of the necessary linguistic resources. Moreover, the results can be compared to those reported by Taylor and Black (1998), who applied the

same suite of experiments to English. The performance measures given for all the experiments presented in this section correspond to five-fold cross-validated scores obtained by matching the system predictions with the hand-labelled annotations described in Section 2.

3.1. Experiment 1: Punctuation Only (PO)

Punctuation marks, provided they are correctly assigned, are an important source of prosodic information, indicating namely how a sentence must be phrased. In our reference corpus, the punctuation marks account for more than half of the total number of breaks. A system using just punctuation would have an average performance of 61.1% correct breaks and a total of 89.4% correctly classified junctures. There would be no false insertions and the 10.6% error would correspond to a failure in predicting a break contemplated in the reference corpus and not coincident with a punctuation mark. These results are comparable to the ones reported for English by Taylor and Black (1998) (correct breaks: 54.3%; correct junctures: 90.8%; false insertions: 0.9%). As often pointed out, this type of method may result in really bad performance when sentences are relatively long and have little or no punctuation.

3.2. Experiment 2: Punctuation Plus Content/Function Word Distinction (PCF)

Another method, independently proposed for English by Silverman (1987) and for French by Sorin et al. (1987), is to insert a break not only on punctuation marks but also after any content word followed by a function word.

This method may be regarded as an attempt to parse sentences into phonological phrases (ϕ domains), as proposed by Nespor and Vogel (1986), but dispensing syntactic information and based only on a list of function words. As ϕ boundaries may coincide with those of association domains for pitch accents and boundary tones (Gussenhoven, 1988), this algorithm often achieves good results.

The rate of correctly predicted breaks increased from 61.1% to 85.1%, with a concomitant reduction in the rate of failure in predicting a break from 10.6% to 4.0%. However, as the number of false insertions was also increased from 0% to 16.8%, the total number of junctures correctly predicted decreased from 89.4% to 79.1%.

For many sentences, the word sequences between any two breaks may effectively correspond to those obtained by the application of the ϕ construction and the ϕ restructuring rules. The first of these rules joins into a same n-ary branching ϕ a lexical head and whatever is on the non-recursive side of the head within its maximal projection, until another head is reached. As non-branching ϕ's are avoided, the second rule (optional) may join the first complement of a head on its recursive side to the ϕ containing the head.

However, as there is insufficient part-of-speech information, the sentences are often parsed in unacceptable ways. The verb, for instance, is separated from its first complement every time this complement begins with a function word and it is always grouped with a preceding noun phrase, ending in a noun or an adjective. If in some cases such groupings may be considered acceptable or even good, most of the time they correspond both to rhythmic and syntactically ill-formed structures. Complex noun phrases can also be split into two or more intonational constituents. This is clearly unacceptable, especially when the break is located before a short prepositional phrase which is a complement selected by a noun or an adjective.

As suggested by Viana (1987) and partially implemented in the first version of the DIXI system for European Portuguese (Oliveira et al., 1991), some of these errors could be avoided by imposing further constraints on degree of constituent branching and length, but for a real improvement, the correct location of the verb is mandatory.

3.3. Experiment 3: Punctuation Plus Part-of-Speech Information (PPOS)

Most linguistic algorithms for the construction of ϕ domains and/or intonational domains rely on a more or less elaborate syntactic parsing. In the absence of reliable syntactic information, many phrasing algorithms for speech synthesis purposes explore combinatory restrictions on POS-tags sequences.

In order to consider part-of-speech information, two questions must be answered: how many different tags have to be considered? and how many words must be included in the analysis window?

In a first step, the original set of 260 tags produced by Palavroso was reduced to 42 by removing all nominal and verbal inflexion marks. Several exploratory experiments were conducted to gain insight into the behaviour of these variables. The window size was varied from

3 to 5 words and two other sets of tags were used, one with 36 categories and another with 11. Best results were obtained with 36 categories and the longest window.

In these exploratory experiments, different distance measures in terms of number of words from the boundary to previous and following punctuation marks were also taken into account. The results showed that the final tree hardly took these measures into consideration. The major decision factor was the location of the punctuation marks and the POS tags of the words.

In order to optimise the tag set, a greedy-type algorithm was used, discarding distance measures. The initial 42 different labels were first reduced to 41 by merging the first two tags into one. A CART was then trained and tested on the resulting data and the performance was recorded. We then repeated the procedure for all other possible combinations of two labels in the original set. The combination producing the best result in terms of correctly placed breaks was selected for the next step of the algorithm. The procedure terminated when the tag set reached a size of 4 labels.

In Fig. 2 the best score is plotted for each tag set size. Given these results, a tag set of 12 labels was selected (see Table 1).

Table 2 summarises the best results obtained in the experiments described in this section. As this table

Table 1. The selected tag set of 12 labels.

Tag	Description
adj	adjective
np	proper noun
adv	adverb
ncard	cardinal number
advm	adverb of manner
PrepC	Preposition (contractions included)
conj	conjunction
v	verb
det	determiner
vpp	verb past participle
nc	common noun
misc	other tags

Table 2. Best results (% correct) obtained for the different experiments.

Model	CB	CJ	FI	MB
PO	61.1	89.4	0.0	10.6
PCF	85.1	79.1	16.8	4.0
PPOS	81.6	92.4	2.6	5.0
greedy	86.3	91.9	4.3	3.8

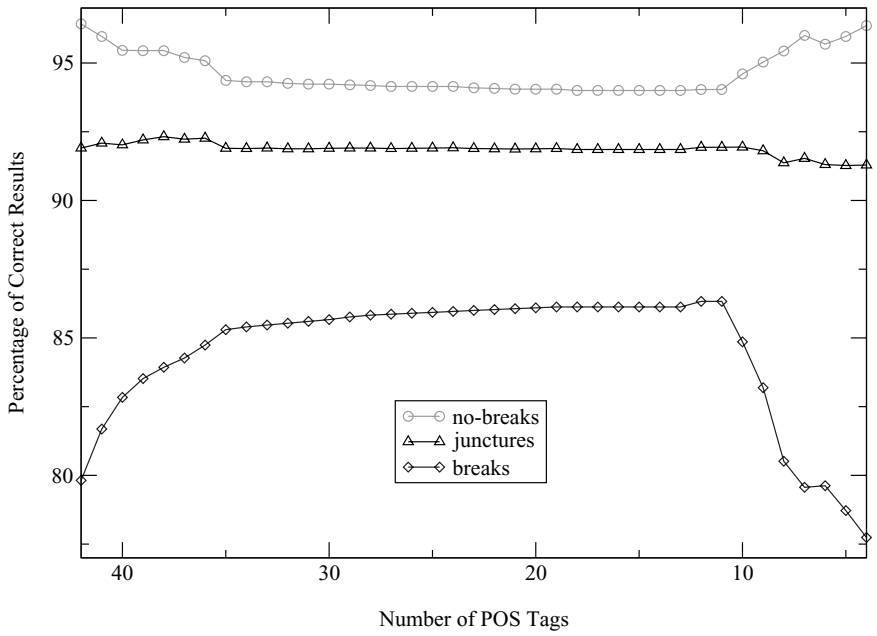

Figure 2. Best rate of correctly placed breaks and the corresponding percentage of correct junctures and no-breaks for each tag set size.

shows, the number of correct junctures obtained with this greedy method and a tag set of 12 categories is very similar to the one obtained with the PPOS method with a tag set of 36 categories and a window size of 5 words (3 to the left and 2 to the right of the boundary). The greedy method, however, allow for an improvement of the CB score and a more equilibrated distribution of false insertions and missing breaks. Both the PPOS and the greedy results obtained compare well with those obtained for English by Taylor and Black (1998) (CB: 79.3%; CJ: 91.6%; FI: 5.6% and MB: 2.8%).

4. Evaluation

So far, the performance of the system for automatic phrasing has been evaluated by matching the predicted value for every word boundary with the corresponding values in the reference test set. However, when a similar evaluation is performed at the sentence level, the results can be deceptive: a perfect match with the reference phrasing pattern is only observed for about 25% of the test sentences.

Although there is common agreement that some sentences may be uttered with different phrasing patterns which are equally acceptable to a human listener, this result is rather problematic since an entire sentence may be also rejected because of a single error. For a more realistic evaluation of the automatic system, an additional assessment method had to be used. For that purpose, an evaluation procedure was developed in order to assess the ability of the automatic system to locate prosodic breaks and also to get some idea of how a group of native speakers evaluates possible partitions of a sentence. Given the expected variability among the subjects in their acceptance of a specific phrasing, the evaluation procedure was designed in such a way that a considerable number of evaluators could easily be recruited. The test was carried out over the Internet and limited to about half an hour per subject.

4.1. Evaluation Tool

The evaluation tool developed for this test requires that the evaluator has Internet access and a web browser such as Netscape or Internet Explorer. The tool runs on an HTTP server and uses the Common Gateway Interface (CGI) to generate forms for the evaluator to fill.

The evaluators were recruited by e-mail announcing the URL address of the test. Snowball recruiting was also attempted by asking the evaluators to spread the address of the test.

The test design had three main objectives:

1. To evaluate the acceptability of the prosodic breaks assigned by the automatic system.
2. To obtain the opinion of the evaluators about the reference phrasing used for training the automatic system.
3. To assess the variability of the evaluators in the task of segmenting a sentence into a limited number of prosodic phrases.

The two first objectives could be reached with the same test, simply by asking the evaluators to rate the sentence phrasing. The written text was marked with the locations of the prosodic breaks, for example:

```
Na Madeira/ haverá chuva/ passando a
  aguaceiros.
(In Madeira/ it will rain/ followed by
  heavy showers.)
```

For the rating, a scale of three values was chosen and it was presented to the evaluators in the following form:

G: Good, I could read it this way.
A: Acceptable, I would not read it this way but it could be a possible reading.
U: Unacceptable, it does not seem to be a natural reading of the sentence.

To carry through the third objective, the sentence was presented to the evaluator with buttons between the words. The evaluator was asked to place breaks in what he/she considered to be appropriate locations.

A preliminary informal test showed that the evaluators had different sensitivity to the break level: some subjects marked only the major breaks, while others produced a large number of phrases. To solve this problem, we tried to force the evaluators to mark a number of breaks within a specified range, between the number of breaks of the automatic and reference phrasing. As this solution was too restrictive (in some cases both phrasings had the same number of breaks), the subjects were allowed to place one break less than the previously calculated minimum.

4.2. Test Sentence Selection and Allocation

To take account of the heterogeneity of the texts, we decided to select 90 sentences for evaluation among

the 819 constituting the full corpus. We chose one in every five sentences, but restricted to be longer than 7 words and removed the excess ones randomly. The test sentences had a maximum of 65 words and an average length of 19 words.

The reference phrasing of this test set had between 2 and 16 prosodic breaks, including the final one, which gave an average of 4 phrases per sentence. After this selection, the CART was retrained on the remaining materials using the set of 12 tags resulting from the greedy algorithm, in such a manner that the test material was not seen during the training procedure.

In this preliminary test, the evaluation of each sentence took one minute on average, which limited the number of sentences per evaluator to 30. Since we needed at least 90 sentences in the test, a strategy had to be devised to allocate a set of sentences to each evaluator.

The sentences were first randomly split into 9 sets of 10 sentences each (set0 to set8). To accomplish the three intended tasks, 3 versions of each sentence were produced: one with the break marks produced by the automatic system (a), another one with the reference phrasing (r) and a third one with just the sentence text to be marked by the evaluator (m). These 270 sentences were randomly ordered to prevent the evaluator from identifying the automatic and reference phrasings, and divided into 9 tasks of 30 sentences each, in accordance with Table 3.

4.3. Subject Enrolment

The test was carried out between March 27th and April 3rd, 2001. The evaluators were asked to participate through e-mail messages sent to researchers of our laboratories (INESC-ID and CLUL), to professors and students of our universities (IST and FLUL) as well as to some personal contacts.

Each evaluator had to select a user name and was then asked for his/her full name. After the identification procedure, one of the nine tasks was assigned to the evaluator. The assignment mechanism was designed to distribute the tasks evenly among the evaluators. It was possible for the evaluator to interrupt the test at any moment and continue later, by giving the user name.

Table 3. Each evaluator's task included 3 sets of 10 sentences for the evaluation of the automatic phrasing (a), the reference phrasing (r) and for marking prosodic boundaries (m).

task0	task1	task2	task3	task4	task5	task6	task7	task8
set0r	set1r	set2r	set3r	set4r	set5r	set6r	set7r	set8r
set1a	set2a	set0a	set4a	set5a	set3a	set7a	set8a	set6a
set2m	set0m	set1m	set5m	set3m	set4m	set8m	set6m	set7m

The 30 sentences of the evaluator set were consecutively presented on separate web pages, as soon as a reply form was submitted for the previous page. After submission, the answer could not be changed. The evaluator was asked to grade the phrasing of 20 sentences (10 with the reference phrasing and the other 10 with the automatic phrasing) as Good, Acceptable or Unacceptable (cf. above). For the 10 remaining sentences, the evaluator had to mark the location of the prosodic breaks that he/she would introduce in a slow but fluent reading.

After April 3rd, the evaluators could visit the URL of the test in order to compare their answers with those given by other evaluators. The goal was to show them that there was no *right* answer and, hence, the importance of a large number of participants. We hope that this will increase their willingness to be involved in future tests.

A total of 105 evaluator registrations was received, of which 91 corresponded to complete tests. Eight tasks were completed by 10 different evaluators, and one task by 11. The result of the last evaluator of this task was discarded to have the same number of evaluators for all tasks.

4.4. Evaluation at the Sentence Level

The results of the automatic phrasing performance, previously presented, accounted for the number of boundaries correctly or incorrectly located. The selected evaluation method allows us to study the performance of the system at the sentence level. The performance results computed this way are more demanding for longer sentences because a single phrase boundary in an unacceptable location is enough to make the sentence as a whole unacceptable, even if all the remaining boundaries are acceptable. This is clearly shown by the fact that in the 90 sentences of this test only 20 have the same phrasing as the reference: with this criterion we would have only 22% correct phrasing. However, the fact that the automatic phrasing differs from reference phrasing does not mean that it is unacceptable.

4.4.1. Evaluators' Variability. As expected, the subjectivity of the phrasing evaluation produced large

Table 4. The evaluation results of the 20 sentences for which the automatic phrasing matched the reference.

Sentence		Evaluation (%)			
id	Length	Agree	Good	Accept.	Unaccept.
8	11	90	85	10	5
18	11	90	40	40	20
43	19	0	55	40	5
53	30	0	30	40	30
103	23	40	40	45	15
118	8	70	80	30	5
133	13	50	85	10	5
173	8	90	85	5	10
323	23	10	55	30	15
378	17	10	40	45	15
383	31	10	30	35	35
418	9	80	95	5	0
438	12	80	90	10	0
443	11	90	55	25	20
458	16	80	80	10	10
473	9	90	80	20	0
488	9	80	75	25	0
538	11	100	80	20	0
543	16	60	100	0	0
548	34	0	20	35	45

differences in the evaluators' judgements. For the 20 sentences with identical automatic and reference phrasing, we have for each sentence the judgement of 20 evaluators (10 evaluating the automatically generated phrasing and 10 evaluating the identical reference phrasing) and the results are presented in Table 4. For instance, for sentence 443, 11 evaluators considered the phrasing good, while 4 found it unacceptable but 9 of the 10 evaluators repeated exactly that phrasing when asked to insert breaks in 443. It is also noticeable that the longer sentences have a larger unacceptability ratio.

Although an important variability in the judgements for relatively short utterances may also be observed, there is a clear relation between the evaluators' agreement values and the utterance length. The number of possible break locations and combinations tends to grow with the length of the sentence. Figure 3 shows the number of phrasing patterns assigned by the evaluators for a given sentence as a function of the number of words. The boxes contain the middle half of the data divided into two quartiles and the vertical dashed lines represent the extreme values. In the special case of the 5 sentences with 12 words, the number of different patterns assigned by the evaluators were: two sentences with two patterns, two sentences with three patterns and one sentence with six patterns. This final data point was considered an outlier and was plotted by itself. For sentences longer than 32 words, the 10 evaluators made 6 or more different phrasing patterns.

Table 5. Percentage of evaluators that agreed on the same phrasing and the grading of that phrasing.

Phrasing agreement (%)	Number of sent.	Average evaluation (%)		
		Good	Accept.	Unaccept.
100	2	80	20	0
90	6	71	19	10
80	5	80	17	3
70	7	67	23	10
60	7	73	15	12
50	9	68	17	15
40	10	54	30	16
30	7	38	38	23
20	7	43	26	31
10	17	36	39	25

An interesting result of the test is to compare the phrasing patterns assigned by the test subjects with the automatic and reference phrasing. This comparison can be made for all the sentences of the test set for which some of the patterns assigned by the evaluators match the automatic or reference phrasing of the sentence. Table 5 shows the percentage of the evaluators that assigned the same phrasing pattern, the number of sentences for which it occurred and the average classification of those patterns. For example, there were 2 sentences for which all the evaluators assigned the same phrasing pattern that matched the automatic or reference phrasing. On the other end, there were 17 sentences for which only one evaluator matched the reference or automatic phrasing, and those phrasings were considered unacceptable by, on average, 25% of the evaluators. Even when 9 of the 10 evaluators agreed on a phrasing, 10% found it unacceptable but rejection rates tend to increase as the degree of agreement for the segmentation performed by the evaluators decreases. This regularity is particularly interesting since the segmentation and grading tasks were never performed by the same group of subjects. In a certain way, it legitimates the methodology of the test and contributes to a better definition of the performance criteria to adopt.

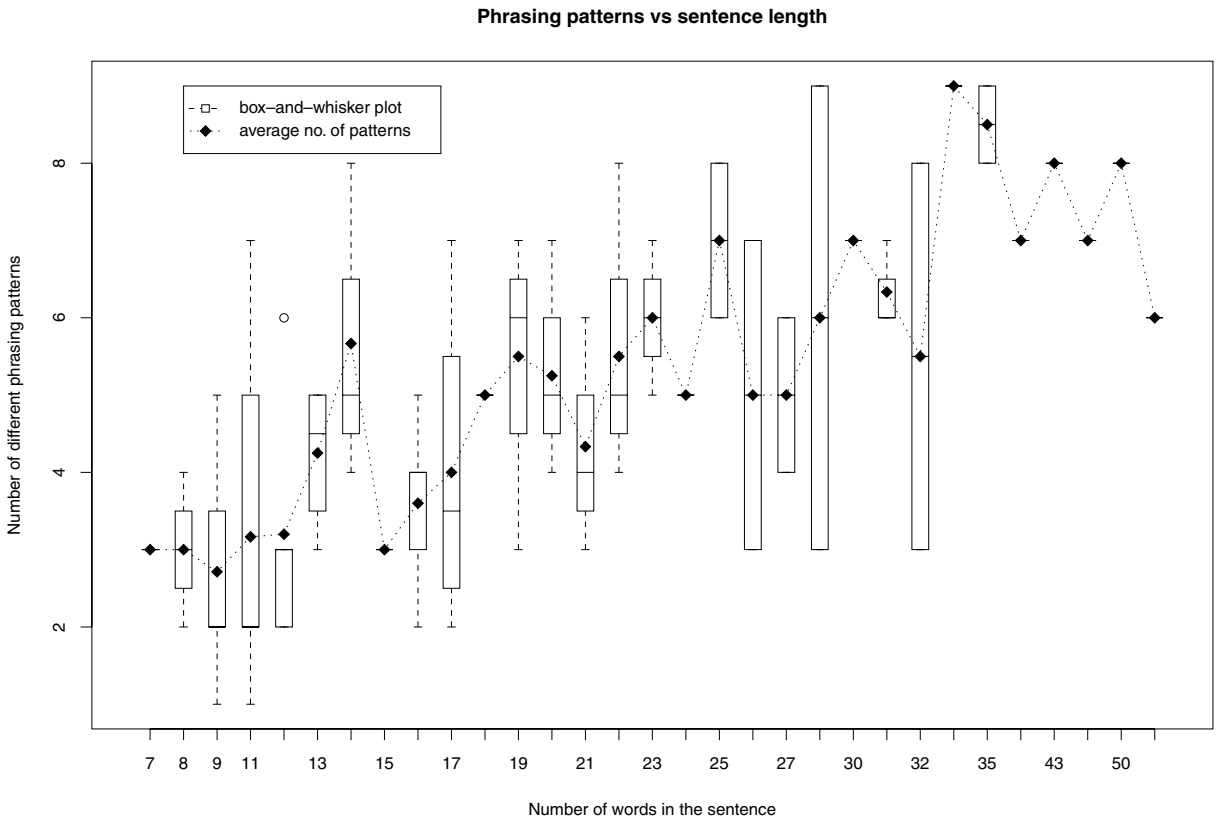

Figure 3. Number of phrasing patterns assigned by the evaluators for each sentence length.

The sum of the second column of Table 5 gives the total number of patterns assigned by the evaluators that were also evaluated. Four of those 77 patterns received the worst grading, for which 5 of the 10 evaluators found the phrasing unacceptable. In two of these sentences 3 evaluators agreed on that phrasing. In accordance with these observations, it was decided that a phrasing could only be rated as unacceptable if more than half of the evaluators considered it as such. Accordingly, it could only be considered a good phrasing if it was classified as such by more than 50% of the evaluators. With this criterion, 41 of the 77 patterns would be rated as Good.

4.4.2. Automatic Phrasing Results. Using the criteria that we have just defined, the evaluators found that the automatic phrasing of 20 of the 90 (22%) sentences was unacceptable. These include the two longest sentences and a sentence that the evaluators considered unacceptable even in the reference phrasing. Of the remaining sentences, 40 (44%) were considered to have an acceptable phrasing and 30 (33%) a good one. Note, however, that for 40 sentences (44%) there was at least one evaluator who performed the same phrasing of the sentence as the automatic procedure did.

4.4.3. Reference Phrasing Results. Applying the same criteria to the judgement made by the evaluators on the reference phrasing, the phrasing of 6 (7%) sentences was considered unacceptable, 31 (34%) good and 53 (59%) acceptable. In this case, the number of sentences for which at least one evaluator reproduced the same phrasing went up to 50 (56%).

4.5. Break Level Results

According to Huang et al. (2001): “There are many reasonable places to pause in a long sentence, but few where it is critical not to pause”. The results presented above show, in fact, that the errors in the assignment of prosodic breaks cannot be found only by matching the breaks in the reference phrasing, there are surely other

Table 6. Number and type of boundary errors in sentences using the performance measures.

No. of errors	Reference			CART		
	No. of sent.	Insert.	Delet.	No. of sent.	Insert.	Delet.
0	72	0	0	48	0	0
1	16	12	4	31	14	17
2	2	3	1	8	6	10
3	0	0	0	3	6	3
Total	90	15	5	90	26	30

acceptable locations. They also show, however, that it may not be enough just to determine the breaks that were assigned to places where it is critical not to pause. There are other places where it is almost mandatory to pause, since the phrasing can become unnatural if a break is missing at that location.

Given these assumptions, the evaluation was made using three performance measures:

1. **correct break:** at least one evaluator placed a break at that same location;
2. **false insertion:** none of the evaluators placed a break at that same location;
3. **missing break:** there should be a break at that location because more than 2/3 of the evaluators agreed on breaking there.

The evaluation results showed that in the 1715 word boundaries of the 90 sentences, the automatic system inserted 389 breaks (22.7%) giving an average phrase length of 4.4 words. The reference phrasing located 448 breaks (26.1%) with an average length of 3.8 words per phrase, while the evaluators introduced on average 370 breaks (21.6%) with an average phrase length of 4.6 words.

Table 6 shows the number of sentences without false insertions and deletions and with 1 to 3 errors due to badly located or missing breaks.

Of the 389 breaks assigned by the automatic system, 26 (6.7%) were considered false insertions because none of the evaluators placed a phrase break at those locations. On the other hand, the system failed to place a break in 30 locations where more than 2/3 of the evaluators agreed. Considering all possible break locations the system assigned 1.5% incorrect breaks and failed to introduce 1.7%.

Performing a similar analysis on the reference phrasing, 15 (3.3%) breaks were considered wrong and 5 breaks were missing. Considering all possible locations for the breaks, the evaluators did not agree on 0.9% of the assigned breaks and would have added 0.3% more breaks.

Of the 20 sentences considered unacceptable by the evaluators, 8 had missing breaks, 6 had incorrectly assigned breaks and the remaining 6 had both missing and incorrectly assigned breaks.

To confirm the adequacy of the performance measures further, we applied the same evaluation procedure to the phrasings produced by the two professional speakers for the 20 sentences analysed above, all of them rated as acceptable or good by the evaluators.

For that set of sentences, with identical automatic and reference phrasing, we have for each sentence the phrasings proposed by 10 evaluators and those produced by the two professional speakers. A comparison between the different global patterns shows that the reference and the CART ones match those of the two professional speakers in 95% of the cases. This corresponds to 90% agreement with the female speaker (IB) and to 70% with the male speaker (LG), who only agree between themselves for 65% of the sentences. On average, the evaluators have an agreement rate of 58% with the reference, 55% with IB and 44% with LG.

As Table 7 shows, for the CART-reference as well as for speaker IB, only one sentence could be rejected by the evaluation at the break level. All three have a missing break in sentence 548, precisely the one that also has 45% of unacceptable judgements.

For the other professional speaker (LG), 20% of the sentences would be classified as unacceptable. This speaker sometimes displaces prosodic boundaries to pause before the constituents he wants to emphasise, as often occurs in spontaneous speech. His reading style does not conform, thus, with the most common reading one.

One of the more interesting aspects of the evaluation procedure described in this paper is the information it is possible to get when crossing the results of the different evaluation tasks. This analysis allows for the identification of the linguistic contexts where the presence/absence of a break is critical. Some of the rejected or badly rated sentences contained restrictive relative clauses preceded by a break or adverbs that were erroneously grouped with a preceding verb. Without more detailed semantic and syntactic information, it is difficult to foresee how to avoid this type of errors. Only a small improvement concerning adverb association may

Table 7. Number and type of boundary errors in sentences using the performance measures for the 20 sentences produced by the 2 professional speakers.

	Reference/CART			IB			LG		
No. of errors	No. of sent.	Insert.	Delet.	No. of sent.	Insert.	Delet.	No. of sent.	Insert.	Delet.
0	19	0	0	19	0	0	16	0	0
1	1	0	1	1	0	1	2	1	1
2	0	0	0	0	0	0	2	3	1
Total	20	0	1	20	0	1	20	4	2

be eventually obtained by increasing the tag set to allow for finer distinctions within this class. However, most of the unfavoured phrasing patterns show discrepancies in the placement of breaks before coordinated constituents and prepositional phrases. As there is a clear relation between the acceptance (or rejection) of breaks at these locations and rhythmic constraints, many potential errors could be avoided if distance measures were taken into consideration.

5. Summary and Conclusions

We have described a set of experiments for building and evaluating a new phrasing module for European Portuguese based on hand-annotated text and using CART techniques.

Results confirm the efficiency of this procedure for acquiring phrasing rules for a new language and for testing the relative weight of different variables. They compare well with results obtained for English by Taylor and Black (1998) who also used information directly obtained from text but on a much larger data set.

To validate our results an evaluation tool has been developed and a test performed using human evaluators who rated the sentences and also assigned prosodic breaks themselves. To deal with the large variability among the evaluators, criteria had to be defined to summarise the data from the different evaluators and tasks. Although the reference phrasing was not rated 100% acceptable, the corpus may be considered adequate since 98.8% of the breaks and 93% of complete sentences were accepted by the evaluators. The CART results are also promising, since only 3.2% of break errors were observed. However the evaluators only accepted 78% of the predicted phrasing patterns at the sentence level. This is nevertheless a much better rate than the 22% initially obtained by a simple match with the reference phrasing patterns. We believe this is a more realistic evaluation of the system performance.

When a majority of opinions is required to validate a judgement, regularities in the evaluators behaviour become evident and a cross-task analysis possible, showing the major sources of errors and pointing to some directions for future work. In the near future, we plan to increase the size of the reference corpus to get a better coverage of some common structures which are not well represented. We will also try to use nested CART trees in order to account for rhythmic constraints and distinguish between breaking levels.

Acknowledgments

The research reported here has been conducted within the Dixi+ project, supported by the Portuguese Foundation for Science and Technology (FCT) Praxis XXI program. This work has benefited from several suggestions and comments of Isabel Trancoso to whom we are particularly grateful. We would also like to thank Isabel Mascarenhas who hand-corrected most of the part-of-speech tagging, as well as all those who freely gave some of their time to participate in the evaluation experiment.

We are grateful to the anonymous reviewers of this work for their careful reading and for many thoughtful suggestions and comments.

References

Bachenko, J. and Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, *16*(3):155–170.

Beckman, M.E. and Elam, G.A. (1997). *Guidelines for ToBI Labeling. Guidelines Version 3.0.* Cleveland, OH: Ohio State University Research Foundation.

Beckman, M.E. and Hirschberg, J. (1994). *The ToBI Annotation Conventions. Appendix A.* Cleveland, OH: Ohio State University Research Foundation.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth and Brooks.

Gee, J.P. and Grosjean, F. (1983). Performance structure: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*:411–458.

Gussenhoven, C. (1988). Intonational phrasing and the prosodic hierarchy. In W.U. Dressler, H.C. Luschutzky, O.E. Pfeiffer, and R. Rennison (Eds.), *Phonologica 1988*. Cambridge University Press, pp. 89–99.

Hirschberg, J. and Prieto, P. (1996). Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, *18*:281–290.

Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliff, NJ: Prentice Hall.

Ladd, D.R. (1996). *Intonational Phonology*. Cambridge, UK: Cambridge University Press.

Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht, The Netherlands: Foris Publications.

Oliveira, L.C., Viana, M.C., and Trancoso, I.M. (1991). DIXI—Portuguese text-to-speech system. *Proc. of the European Conference on Speech Technology*. Genoa, Italy, pp. 1239–1242.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Pierrehumbert, J. and Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Selkirk, E. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Selkirk, E. (1986). On derived domains in sentence prosody. In C.J. Ewen and J.M. Anderson (Eds.), *Phonology Yearbook 3*. London: Cambridge University Press, pp. 371–405.

Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*, PhD thesis, Cambridge University, Cambridge, UK.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Whightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling english prosody. *Proceedings of International Conference on Spoken Language Processing, ICSLP'92*. Banff, Canada, pp. 867–870.

Sorin, C., Larreur, D., and Llorca, R. (1987). Rhythm-based prosodic parser for text-to-speech system in French. *Proceedings of the 11th International Congress of Phonetic Sciences*. Tallinn, Estonia, USSR, pp. 125–128.

Taylor, P. and Black, A. (1998). Assigning phrase breaks from part-of-speech sequences. *Computer Speeech and Language*, *12*(2):99–117.

Viana, M.C. (1987). *Para a Síntese da Entoação em Português*, PhD thesis, CLUL-INIC, Lisbon, Portugal.

Wang, M.Q. and Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, *6*:175–196.