Phrase break prediction: a comparative study.

Pablo Daniel Agüero
Universitat Politècnica de Catalunya
Jordi Girona 1-3 Barcelona. Spain.
pdaguero@gps.tsc.upc.es

Antonio Bonafonte
Universitat Politècnica de Catalunya
Jordi Girona 1-3 Barcelona. Spain.
antonio.bonafonte@upc.es

Resumen: Este artículo presenta un estudio comparado de varios métodos de predicción de pausas, usando el mismo corpus etiquetado. Algunos métodos propuestos por otras publicaciones sobre el tema son probados, combinando algunas técnicas previas para aprovechar sus principales ventajas. Un nuevo método es propuesto modelando explícitamente la función densidad de probabilidad de la distancia entre pausas. Los resultados muestran que las técnicas basadas en datos ofrecen muy buenos resultados.
Palabras clave: predicción de pausas, conversión de texto a voz, basado en datos

Abstract: This article presents a comparative study of several methods of phrase break prediction, using the same labelled corpora. Some previous methods proposed in the literature are tested, mixing techniques to take advantage of their benefits. It is proposed an approach based on explicitly modelling the probability density function of the distance between breaks. The results have shown that data-driven techniques provide very good results.
Keywords: phrase break prediction, text-to-speech, data-driven

1. Introduction.

Phrasing is one of the key topics in the linguistic part of text-to-speech technologies. Phrasing consist on breaking long sentences into smaller prosodic phrases. Boundaries are acoustically characterized by a pause, a tonal change, and a lengthening of the final syllable. Punctuation is quite correlated with prosody. In many cases the main function of punctuation is more related to syntaxis, without acoustic correlates. The following sentence (taken from *Harry Potter and the Sorcerer's Stone*) shows that some phrase breaks appear without punctuation:

```
Y, por último, <PB>observadores
de pájaros de todas partes, <PB>han
informado que hoy las lechuzas de la
nación <PB>han tenido una conducta
poco habitual. <PB>
```

Some speakers may prefer to add a break after *y*, or *conducta*, or do not include the break after *han*. From a text-to-speech perspective, any of these options are acceptable. But it is not correct to limit breaks to punctuation marks, or to include breaks in incorrect positions, as after *por*.

Phrase breaks have strong influence on naturalness, intelligibility and interpretation of a sentence. The presence or absence of them can produce a change in the meaning of a sentence.

In general, there are two approaches to solve problems of natural language processing. The knowledge based approach consists in incorporating information inside the system produced by human experts. On the other hand, the data-driven approach uses labelled corpora to induce automatically information, in the form of rules, decision trees or statistical information, to mention some ways of representing the acquired knowledge. It requires less experience and human resources, and the results may be similar to the other approach, with the advantage that it facilitates the migration between languages.

Several data-driven approaches have been proposed in the literature.

Hirschberg and Prieto (1996) proposed to train a decision tree to place phrase boundaries using the following features: a 4-word POS window (POS: part of speech, morphological category of the word); 2-word window for pitch accents; the total number of words and syllables in the utterance; the distance of the word from beginning and end of the sentence in words, syllables, and stressed syllables; distance from the last punctuation in words; whether the word is at the end, within, or at the beginning of an NP (Noun phrase), and if within an NP, its size and the distance of the word from the start of the NP.

P. Koehn, S. Abney, J. Hirschberg, and M. Collins (2000) have proposed a modification to the previous system adding syntactic fea-

tures, reporting a significant improvement.

These two methods place boundaries taking into account local information, and they do not use the location of previous boundaries on the decision.

E. Navas, I. Hernaez, N. Ezeiza (2002) have proposed an interesting method based on CART for assigning phrase breaks in Basque language, using syntactic and morphological information.

A. Black and P. Taylor (1997) have proposed a different system based in Bayes Decision Rule. They proposed to maximize the expression

$$J(C_{1,n}) = argmax_{j_{1,n}} P(j_{1,n}|C_{1,n})$$

where $J(C_{1,n})$ is the sequence of n junctures. These junctures can be breaks or not breaks. C_i is the context information of the juncture, which considers two previous POS tags and the following to the position of the phrase boundary.

$P(j_{1,n}|C_{1,n})$ is calculated as

$$P(j_{1,n}|C_{1,n}) = \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-l} \cdots j_{i-1})$$

where $P(j_i|C_i)$ is the probability of a juncture according to the adjacent tags, $P(j_i)$ is the probability of each juncture (break or non-break), and $P(j_i|j_{i-l} \cdots j_{i-1})$ is the n-gram of the juncture probability according to the previous l junctures.

X. Sun and T. H. Applebaum (2000) extended the approach of Black and Taylor, but they estimate the probabilities $P(j_i|C_i)$ using binary decision trees.

Marín, R. , L. Aguilar y D. Casacuberta (1996) use the stress group concept to place phrase breaks. The authors assume that in Spanish there are not phrase breaks inside a stress group. If this hypothesis is true, it can be used in all the other methods as context information in place of POS tags, reducing the search space, and possible errors. A stress group is a sequence of words belonging to non-stress POS classes (determiner, possessive adjectives, prepositions, conjunctions and non-stress pronouns) ending with a word belonging to a content class (noun, adjective, stress pronouns, verb and adverb). The stress groups are labelled considering the POS tag of the head and the POS tag of the content word.

Other methods were previously proposed by E. Sanders, P. Taylor (1995). These methods make some strong assumptions to simplify the problem which do not always apply.

In summary, there are several data-driven methodologies that achieve good results. However, most of the experiments have been done in English and different corpus have been used for evaluation, which turns difficult to make a fair comparison. On the other hand, some of the good ideas of the methods can be used in the other methods, for example, Hirschberg and Prieto (1996) features can be used with Bayes Decision Rule approach. Furthermore, the POS tag features can be replaced by stress groups.

In this paper a new method is proposed using a probability density function to model the probability of a phrase break in a given position. The level building algorithm is applied to solve the search problem.

2. Experimental framework.

2.1. POS tagger.

In this paper we have used an extension of the PAROLE POS tagset. For example, it is ignored gender and number in nouns, and person and some information about tense on verbs. The total number of tags is 52.

We have used the POS tagger of our TTS system. The statistical tagger has an estimated accuracy of 94,54 %. The LEXESP corpus was used to train the POS tagger. The corpus was produced in the project of the same name, carried out by the Psychology Department of the University of Oviedo and developed by the Computational Linguistics Group of the University of Barcelona and Language Processing Group of the Catalonia University of Technology.

2.2. Corpus for phrase break prediction.

To evaluate the different approaches we have produced a corpus introducing break labels in a written text, following the methodology proposed by Hirschberg and Prieto (1996).

The corpus for phrase break prediction consists in 63000 words. Two types of breaks have been labelled, associated to the break index 2, 3 and 4 of the ToBI break index tier. However, in this work we will only consider breaks (B) and not breaks ($\neg B$).

The number of breaks in the corpora is 8505, and the number of breaks inside stress groups is 657, which represents an omission of 7,73 % of breaks. Those breaks were manually checked, and most errors are caused by POS tagger mistakes. The other pauses may be ignored, without altering naturalness. The distributions of breaks in the corpora can be seen on table 1, were B represents the number of phrase breaks, $\neg B$ represents the number of non phrase breaks, P represents that the position is after punctuation marks and $\neg P$ represents that the position is after non punctuation marks. The phrase breaks after the full stop are not going to be considered in this paper, because no prediction is needed.

The corpus is divided in 70 % for training and 30 % for testing purposes.

	P	$\neg P$
B	3696	4809
$\neg B$	3005	58043

Cuadro 1: Distribution of breaks.

2.3. Probability estimation.

The probabilities of n-grams are calculated using the concept of x-grams (A. Bonafonte (1996)). X-gram is an extension of n-grams. In this extension, the memory of the model (n) is not fixed a priori.

CART are trained using wagon, which is part of the Edinburgh Speech Tools Library.

3. Phrase break prediction.

3.1. CART based phrase break prediction.

This method consists of learning decision trees to place phrase boundaries. The feature set is the proposed in Hirschberg and Prieto (1996) publication, but we do not include the information of Noun Phrase, because syntactic information is not available in our TTS system. Table 2 shows the results of applying this method, where B indicates error percentage in breaks, $\neg B$ indicates error percentage in non-breaks, and *Total* indicates the total error percentage. Two additional columns are included, P and $\neg P$, to analyze the errors after a punctuation mark or after a non punctuation mark. The same presentation will be used in the rest of the paper.

	Global	$\neg P$	P
B	31.52 %	31.65 %	23.57 %
$\neg B$	2.44 %	2.69 %	0.44 %
Total	6.01 %	6.55 %	0.93 %

Cuadro 2: Results of the method CART-PT

3.2. CART based phrase break prediction using stress groups.

This method extends the previous one, but POS tags have been replaced by stress groups.

In Spanish it can be assumed that there are not phrase breaks inside a stress group. The use of this information improves the system, because it avoids placing phrase breaks inside a stress group.

The results are shown in table 3.

The results are worse. A possible reason is that the number of stress group tags (206) is larger than the number of POS tags (50). As a consequence, the training set has not the necessary information to let the tree to generalize. The errors of the assumption of stress group are not considered.

	Global	$\neg P$	P
B	31.83 %	48.04 %	15.93 %
$\neg B$	4.71 %	6.06 %	3.4 %
Total	6.29 %	14.69 %	5.95 %

Cuadro 3: Results of the method CART-SG

3.3. Bayes Decision Rule phrase break prediction.

A. Black and P. Taylor (1997) proposed to solve the problem using Bayes Decision Rule. It should improve the performance of the system, because the phrase break decision takes into account past decisions, instead of performing local decisions.

The goal of this method is to maximize the probability

$$J(C_{1,n}) = argmax_{j_{1,n}} P(j_{1,n}|C_{1,n})$$

where $J(C_{1,n})$ is the sequence of n junctures. These junctures can be breaks or not breaks. C_i is the context information of the juncture, which considers two previous POS tags and the following to the position of the phrase boundary.

The previous expression can be written as

$$J(C_{1,n}) = argmax_{j_{1,n}} \frac{P(j_{1,n}, C_{1,n})}{P(C_{1,n})}$$

where $P(j_{1,n}, C_{1,n})$ can be decomposed as

$$P(j_{1,n}, C_{1,n}) =$$

$$= \prod_{i=1}^{n} P(C_i|j_{1,i}, C_{1,i-1}) P(j_i|j_{1,i-1}, C_{1,i-1})$$

If we make the assumptions that

$$P(C_i|j_{1,i}, C_{1,i-1}) = P(C_i|j_i)$$

$$P(j_i|j_{1,i-1}, C_{1,i-1}) = P(j_i|j_{i-k,i-1})$$

We obtain

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} P(C_i|j_i) P(j_i|j_{i-k,i-1})$$

If we use the equality

$$P(C_i|j_i) = \frac{P(C_i)P(j_i|C_i)}{P(j_i)}$$

We finally obtain

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} \frac{P(C_i)P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1})$$

As a consequence, we maximize the following expression

$$J(C_{1,n}) = argmax_{j_{1,n}} \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1})$$

where $P(j_i|C_i)$ is the probability of a juncture according to the context C_i, $P(j_i)$ is the probability of each juncture (B or $\neg B$), and $P(j_i|j_{i-l} \cdots j_{i-1})$ is the n-gram of the juncture probability according to the previous k junctures. The results of this method are shown in table 4.

	Global	$\neg P$	P
B	23.93 %	38.03 %	5.57 %
$\neg B$	4.72 %	4.08 %	17.1 %
Total	7.07 %	6.68 %	10.75 %

Cuadro 4: Results of the method BDR-BT

If C_i is changed to the group of characteristics of Hirschberg and Prieto (1996) , we get a global improvement that can be seen on table 5. However, the number of phrase breaks after non punctuation marks increases up to 44.39 %, which is not a good percentage.

	Global	$\neg P$	P
B	27.45 %	44.39 %	5.41 %
$\neg B$	3.00 %	2.27 %	17.13 %
Total	5.99 %	5.5 %	10.67 %

Cuadro 5: Results of the method BDR-PT

The variation of the accuracy according to the number n of the n-grams is shown on figure 1. By increasing the length of the history of n-grams, phrase break error percentage improves, but non phrase break error percentage is worse. As a consequence, there is a compromise in the selection of n.

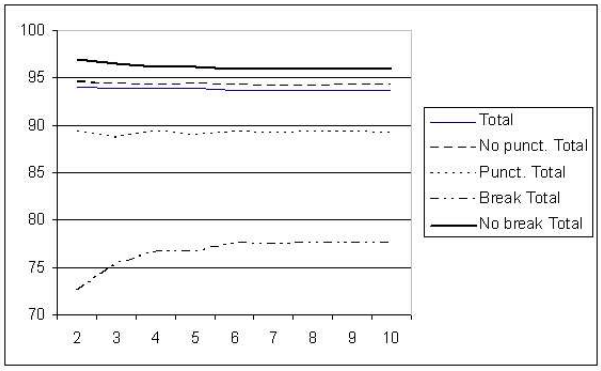

Figura 1: Variation of the accuracy according to the number n of the n-grams.

Another configuration has been considered where POS tags are replaced by stress group labels, giving the results of table 6.

The variation of the accuracy according to the number n of the n-grams taken is shown in figure 2. The same conclusions of figure 1 apply here.

	Global	$\neg P$	P
B	27.05 %	44.66 %	4.89 %
$\neg B$	3.26 %	4.33 %	42.67 %
Total	6.24 %	9.55 %	14.63 %

Cuadro 6: Results of the method BDR-SG

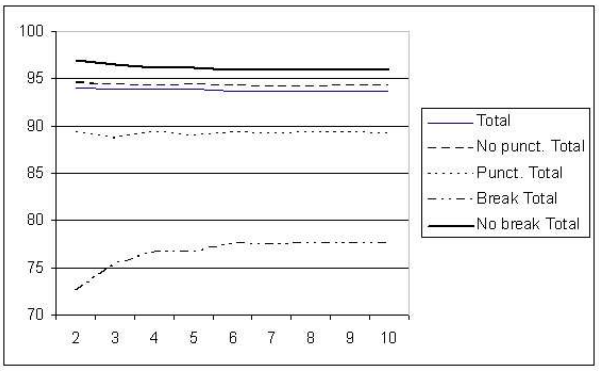

Figura 2: Variation of the accuracy according to the number n of the n-grams.

3.4. Phrase break prediction using probability density function of distance between phrase breaks.

This algorithm uses a probability distribution of the distance between phrase boundaries (figure 3), the probability distribution of the appearance of n consecutive non phrase boundaries (figure 4), and the probability of the appearance of a juncture in a given context, estimated using the decision tree of the CART-SG method.

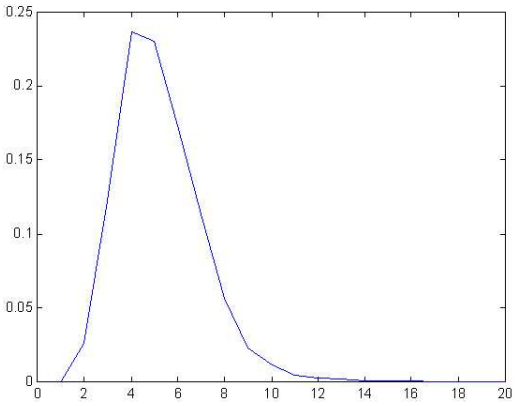

Figura 3: Probability density function of the distance of phrase boundaries. ($P_{pb}(d)$)

The method uses the level building algorithm in order to find the optimum number of breaks and their optimum place.

In each iteration, the algorithm finds the probability of the appearance of a break in each position, taking into account the optimum break position of the breaks estimated in the previous iteration. The steps of the algorithm are:

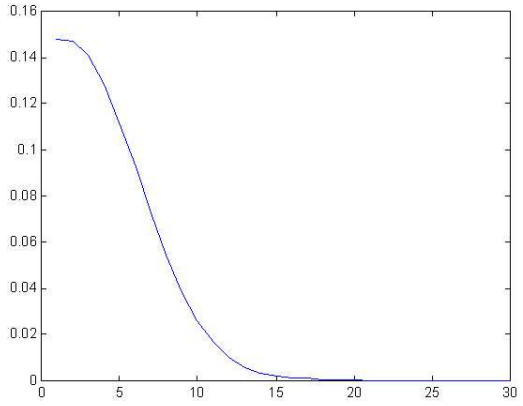

Figura 4: Probability density function of n consecutive non phrase boundaries. ($P_{npb}(d)$)

1. For each position i in the sentence (only the places in the boundary of a stress group are considered) find the probability of placing a break, considering that no other breaks are put before.

2. For each position i in the sentence (only the places in the boundary of a stress group are considered) find the probability of placing a break, considering that the previous break is at position j, with $j < i$.

3. Repeat iteratively step 2, until the number of breaks is equal to the number of stress groups.

4. Then a backtracking is performed from the iteration where the probability of placing a break after the last stress group is highest.

The algorithm maximizes the expression

$$max\left[\prod_{i=1}^{n} P(j_i|C_i)P(j_i|j_{1,i-1})\right]$$

where $P(j_i|j_{1,i-1})$ is approximated in case that $j_i = B$ by $P_{pb}(d)$, being d the distance between consecutive phrase breaks, and in case $j_i = \neg B$ by $P_{npb}(d)$, being d the distance between consecutive non phrase breaks

The results of applying this method are shown on table 7.

The same method can be applied replacing stress groups by POS tags. The results are shown on table 8.

In both cases, the phrase break error percentage decreases, but the number of non phrase break error percentage increases. It is not desirable, because it will affect naturalness. However, it is necessary to analyze the phrase breaks to reveal the number of

serious errors. In the literature the methods have shown better results when the errors are checked, because some of them are not real errors, and the total accuracy is higher.

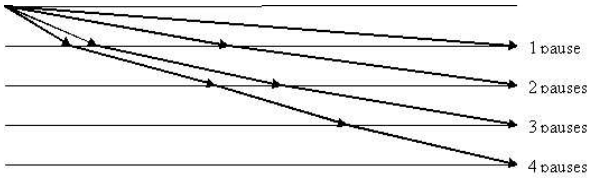

Figura 5: Representation of the work done in each iteration

	Global	$\neg P$	P
B	17.15 %	28.88 %	2.39 %
$\neg B$	7.26 %	11.2 %	57.33 %
Total	8.51 %	13.5 %	16.55 %

Cuadro 7: Results of the method LB-SG

	Global	$\neg P$	P
B	17.57 %	28.07 %	3.92 %
$\neg B$	6.37 %	5.64 %	20.5 %
Total	7.75 %	7.36 %	16.36 %

Cuadro 8: Results of the method LB-PT

4. Conclusions.

In this work we have evaluated seven methods to predict phrase breaks using a data-driven approach. Some of the methods have followed the proposal of Marín, R. , L. Aguilar y D. Casacuberta (1996) of using the tagged stress group as the basic unit. However these methods yield worse results. We believe that a reason is that the number of possible tags associated to stress groups is larger than the number of POS tags, making difficult to have a robust estimation of probabilities. Furthermore, errors in the POS tagger can imply errors in the boundary prediction.

Tables 9, 10, 11 and 12 summarize the results of all methods.

Method CART-PT has the highest accuracy placing non phrase breaks, but the accuracy placing phrase breaks is low.

Methods BDR-BT, BDR-PT and BDR-SG improve the accuracy of phrase breaks, but the accuracy of non phrase breaks is lower.

The methods LB-PT and LB-SG have shown to be the best, taking into consideration the global F-measure, and punctuation and non-punctuation boundary F-measure. This method has the advantage that allows to choose the number of phrase boundaries, which will help in a system that varies the speed of talking. The problem is that this method has the highest non phrase break error percentage, which can cause naturalness problems.

It is necessary to make an evaluation of the errors, to analyze how many of them are serious. In the literature, the analysis of errors have shown that most of them are not serious, and they are due to multiple possible versions of phrase break tags. With this consideration, the accuracy of a system would result higher.

	B	$\neg B$
CART-PT	68.48	97.56
CART-SG	68.17	95.29
BDR-BT	76.07	95.28
BDR-PT	72.55	97.00
BDR-SG	72.95	96.74
LB-SG	82.85	92.74
LB-PT	82.43	93.63

Cuadro 9: Summary of total accuracy

	Precision	Recall
CART-PT	68.35	96.21
CART-SG	51.96	89.56
BDR-BT	61.97	93.82
BDR-PT	55.61	96.08
BDR-SG	55.34	92.74
LB-SG	71.12	86.39
LB-PT	71.93	92.73

Cuadro 10: Summary of precision and recall of each method after non punctuation marks

5. Acknowledgements.

This work has been (partially) sponsored by the Spanish Government under grant TIC2002-04447-C02.

Pablo Daniel Agüero has a scholarship of Generalitat de Catalunya, in the framework of the promotion of the International Graduate School of Catalonia (DOGC 3721-17/09/2002).

	Precision	Recall
CART-PT	76.43	99.43
CART-SG	84.07	96.11
BDR-BT	94.43	84.67
BDR-PT	94.59	84.67
BDR-SG	95.11	69.03
LB-SG	97.61	63.00
LB-PT	96.08	82.42

Cuadro 11: Summary of precision and recall of each method after punctuation marks

	F global	F $\neg P$	F P
CART-PT	80.13	80.04	79.92
CART-SG	78.86	66.81	65.76
BDR-BT	84.15	74.75	74.64
BDR-PT	82.65	70.43	70.45
BDR-SG	82.80	70.13	69.32
LB-SG	87.16	80.20	78.02
LB-PT	87.32	81.05	81.02

Cuadro 12: Summary of F-measure for each method

References

P. Prieto, J. Hirschberg. 1996. Training Intonational Phrasing Rules Automatically for English and Spanish text-to-speech. *Speech Communication 18, ps. 281-290..*

P. Koehn, S. Abney, J. Hirschberg, and M. Collins. 2000. Improving Intonational Phrasing with Syntactic Information. *Proceedings of ICASSP-00, Istanbul..*

A. Black and P. Taylor. 1997. Assigning Phrase Breaks from Part-of-Speech Sequences. *Proceedings of Eurospeech'97, Rhodes, pp. 995-998.*

X. Sun and T. H. Applebaum. 2001. Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model. *Proc. of 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, Vol 1, pp. 537-540..*

A. Bonafonte. 1996. Language Modeling Using x-grams. *Proc. of ICSLP-96, Philadelphia, October 1996. .*

Marín, R. , L. Aguilar y D. Casacuberta. 1996. El Grupo Acentual Categorizado como Unidad de Análisis Sintáctico-Prosódico. *XII Congreso de Lenguajes Naturales y Lenguajes Formales, La Seu D'Urgell, 23-27 de septiembre de 1996. .*

E. Navas, I. Hernaez,N. Ezeiza. 2002. Assigning Phrase Breaks using CART's in Basque TTS. *Proc. of the 1st Int. Conf. on Speech Prosody, Aix-en-Provence, pp. 527-531, 2002. .*

E. Sanders, P. Taylor. 1995. Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis. *Proceedings EUROSPEECH 1995, Madrid, Spain.*