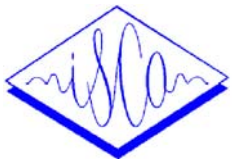

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Automatic Determination of Phrase Breaks for Argentine Spanish

Humberto M. Torres and Jorge A. Gurlekian
Laboratorio de Investigaciones Sensoriales, CONICET
Universidad de Buenos Aires, Argentina
jag@fmed.uba.ar

Abstract

This work evaluates the efficiency of different word classes -part of speech-, normalized vs. non normalized counting for syllable and word occurrences, to predict non orthographic breaks of an Argentine Spanish database, designed for the development of the prosody component for a Text To Speech system. Within a set of 741 sentences, regression trees were trained and tested with two different proportions of data. The results show an error range of 8 to 15% whose minimum value is related to a reduced amount of morphologic categories, and a normalized counting of syllables and words.

1. Introduction

Text to speech systems require the determination of prosody components which should be based on the linguistic information carried out just by the input text. In order to predict phrase breaks that define an intonative group, two indicators are available: one is the presence of punctuation marks and the other, is the insertion of non orthographic breaks between words, produced acoustically by pauses and tonal movements by a standard speaker. Some agreements between syntactic and prosodic structures usually found in sentences allow the use of statistical methods of inference, such as classification and regression trees[3]. Searching for the automatic determination of breaks, recent works have demonstrated the power of binary classification trees (CARTs) on the determination of phrase breaks sentences using data driven approaches [7]. Hirschberg and Prieto [6], and Aguero and Bonafonte [1], both used the same input features for Spanish databases, which included tonal accent information. Also they hand labelled breaks for likely prosodic boundaries in their text.

In this paper, two main differences are considered: first, we are not using any syntactic information nor tonal accent information, and second, the database was recorded by professional announcers who produced all kind of acceptable phrase and tonal variations in Argentine Spanish.

This paper is organized as follows: the database is described in Section 2 with the definition of different word classes. Input features are summarized in section 3. Results are described in section 4. Sections 5 and 6 show the discussion and a brief conclusion respectively.

2. Database

The corpus SECYT [5] consists of 741 declarative sentences extracted mainly from Argentine newspaperswith a majority of two intonational phrases.

Table I. *Number of intonation phrases per sentence (observed and relative frequency).*

Number of phrases	N.	%
1	183	14.27
2	797	62.16
3 (or more)	294	22.93
Total	**1282**	**100**

The sentences contain 97% of all Spanish syllables; including all the possible occurrences of accent patterns.

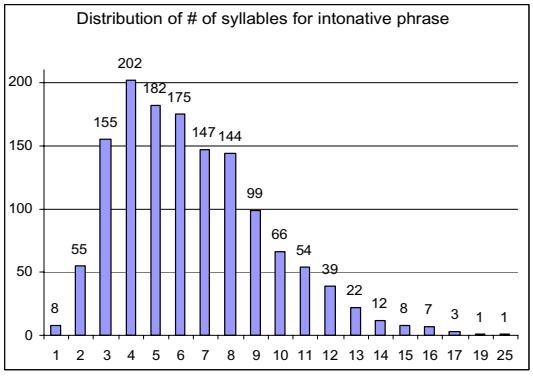

Figure 1. *Shows the distribution of syllables for intonational phrase in the database.*

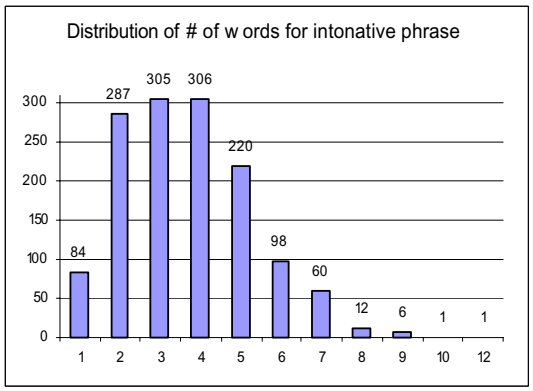

Figure 2. *Shows the distribution of words for intonational phrase in the database.*

Two professional announcers native from Buenos Aires read the sentences. Recording took place in a sound-proof chamber. All recordings were performed at 16 kHz and 16 bit. The instructions to the speakers were to read the sentences with all kind of acceptable phrase and tonal variations.

10.21437/SpeechProsody.2004-128

2.1 Labeling

Each sound file was manually labeled twice, by four speech therapists, with additional musical training. We decided to select people with musical training for their ability in distinguishing tonal and rhythmic variations. First, all the files were phonetically transcribed in order to add precision to the labeling, since all the other marks are then aligned with it. The second stage involves an orthographic transcription, after which break indexes (0 to 4) are added to indicate the degree of perceived juncture between words and phrases. As in ToBI, break indexes 3 and 4 are associated with major units, i.e. intermediate and intonation phrases. We followed ToBI criteria for assigning those indexes. Although we agree with Beckman et al. [2], that there is still no conclusive evidence to postulate two levels of phrasing in Spanish, we considered it practical to mark a distinction between two degrees of perceived disjuncture, since the database will be used to model prosody, and different levels of disjuncture would produce a more natural speech. Finally, each word was hand labeled with the following list of POS categories.

2.2 Labels

Two sets of POS categories were defined as detailed in Table II. One with 19 items and the second with 8 items.

Table II. *Part of Speech tags(POS). First set at left (19 items). Second set at right (8 items).*

1	C	Coordinative Conjunction	**1**
2	A	Preposition or Subordinate Conjunction	**2**
3	E	Determiner	
4	D	Adjective	**3**
5	S	Singular Noun	**4**
6	U	Plural Noun	
7	T	Singular Proper Noun	
8	M	Pronoun	**5**
9	V	Adverb	**6**
10	X	Foreign word	**-**
11	N	Cardinal number	**7**
12	J	Interjection	**-**
13	I	Infinitive verb	**8**
14	H	Simple Past verb	
15	P	Present Participle verb	
16	B	Past participle verb	
17	R	Present verb	
18	F	Future verb	
19	O	Transitive Verb	

3. Input Features

Input features for the prediction tree are the POS categories in a 5 word window, total number of syllables, total number of words, number of syllables and words from each word to the next punctuation mark.
Syllables included fusion of syllables at word boundaries.

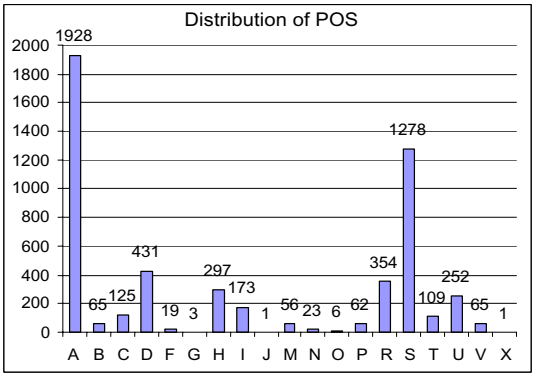

Figure 3. *Shows the distribution of total POS categories in the database.*

4. Results

Experiments were designed to confirm the efficiency of CARTS on the determination of non orthographic breaks, and to determine the relative contribution of input features that feed the trees. The results obtained by the trees trained to predict phrase breaks are shown in tables III to VIII. The scores correspond to the tree applied to data.

The following codes were used in the tables:

NB: means no phrase breaks, **B**: presence of a phrase break.
NBp and **Bp** are the correspondent predicted values.

The different parameters are:

Number of POS: **20, 8,**
Test/Training proportion: **30/70, 10/90,**
Syllable/word counting normalized to total number: **NonNorm, Norm,**
Type of ToBI Break 3 and 4, only 4: **3-4, 4.,**
Prunning: **Pru.**

Table III. *19POS,TT30/70, NonNorm, B3-4*

Test	NB	B	#Err	Err %
NB	900	139	139	14.39
B	88	451	88	

Table IV. *8POS, 30/70, NonNorm, B3-4*

Test	NB	B	#Err	Err %
NB	888	151	151	15.78
B	98	441	98	

Table V. *19POS,TT 30/70, Norm, B4*

Test	NB	B	#Err	Err %
NB	1052	82	82	12.86
B	121	323	121	

Table VI. *8POS, 30/70, Norm, B4*

Test	NB	B	#Err	Err %
NB	1068	66	66	11.53
B	116	328	116	

TableVII. *8POS, 10/90, NonNorm, B4*

Test	NB	B	#Err	Err %
NB	1052	82	82	9.84
B	121	323	121	

Table VIII. *8POS, 10/90, nonNorm, B4, Pru*

Test	NB	Bp	Err	err%
NB	366	13	13	8.18
B	30	116	30	

Five different sentences selection, using the features described in Table VI (70% for training and 30% for validation) resulted in similar results, where errors ranged from 10.03% to 12.27%. Then a random sentence selection was used.

Kappa statistic [4] that relates the overall score 91.82% of the best tree (Table VIII) and the proportion of non breaks (0.7219) in the data was calculated. The value of 0.705 was obtained, which indicates a good score for the performance of the tree.

By observation of the resulting tree, the first and decisive branches are signaled by the POS and the total number of syllables.

5. Discussion

The results indicate a better performance of the reduced number of POS categories. See table V and VI. Observing the terminal node in the resulting tree, when all POS categories are used, a subset of labels determine one category (B or NB). We can hypothesize that when we reduce the number of labels, we are simplifying the regression tree task. Besides, we are increasing data for each of the remaining labels, which in turn reinforce the training of the CART.

Normalization to the total number of syllables-words resulted in a clear improvement of the general performance. Compare table VI and VII. This is in accordance with, the longer the sentence the more phrase groups should be proportionally expected.

Our results also show that a particular selection of training and validation data sets also influence the system performance.

We choose a random selection for this purpose, expecting to have a minimum influence on final results.

Besides, two proportions of number of elements were chosen: 70%-30% and 90%-10%, and as expected the latter rendered better results. The outcome is that the system has more data to learn, the performance will be better. Nevertheless, the results obtained seem to be acceptable compared to other works and new information seems to be necessary, more than additional data.

The general error reduction observed when only breaks type 4 were evaluated indicates that breaks type 3 are more subtle labeled and more information cold be needed to solve it. At this point, the addition of syntactic information, such as flag indicating that some POS combinations constitute a prepositional phrase reduced the error rate from 8.18% to 5.18%. For this reason we are now labeling three syntactic layers: 1. type of sentence, 2. subject/predicate, 3. Noun, verbal, adverbial and adjective phrases, whose contribution will be presented shortly.

Regarding tonal accents we have considered the general approach to modelate prosodic markers independently, which can be useful and less sensitive to accumulative errors.

6. Conclusions

In this work we have confirmed the efficiency ofCARTs to indicate phrase breaks in Argentine Spanish. This information will allow the determination of intonative groups of the sentence, which is the first step to include prosody information in a TTS system

Number and classes of POS categories are relevant t

General performance, as the total number of syllables, which also influences the normalization factor when counting syllables and words.

Improvements associated with the training-test proportions confirm the expected results observed in other contributions.

Overall performance results obtained in these experiments are slightly inferior compared similar published works on Spanish [1] [6], but they result promissory as they could be easily implemented and improved by adding syntactic information.

7. Acknowledgments

The authors are grateful to Dr. Laura Colantoni who helped to define the POS tags and labeled the entire database.

We also recognize the support of the National Research Council of Scientific and Technical Research of Argentina , (grant CONICET PIP99).

References

[1] Aguero, P.D.; Bonafonte, A., 2003. Phrase break prediction: a comparative study. *XIX Congress of the Spanish Society for natural language processing*, Alcalá de Henares, Spain.

[2] Beckman, M.; Campos, M.D.; McGory, J.T.; Morgan, T.A., 2002. Intonation across Spanish, in the tones and break indexes framework. *Probus*, 14, 9-36.

[3] Breiman, L.; Friedman, J.H.; Olsen, R.A.; Stone, C.J., 1984. *Classification and Regression Trees*. Chapman&Halle.

[4] Carletta, J.C.,1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249-254.

[5] Gurlekian, J; Rodríguez, H.; Colantoni, L; Torres, H.M., 2001. Development of a prosodic database for an Argentine Spanish TTS system. Proceedings of the IRCS Workshop on Linguistic Databases. 11-13 December 2001. University of Pennsylvania, Philadelphia, USA.

[6] Hirchsberg, J.; Prieto, P., 1996. Training intonational Phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, Vol. 18, 281-290.

[7] Navas, E.; Hernáez, I.; Ezeiza, N., 2002. Assigning Phrase Breaks using CART's in Basque TTS. Proceedings of the 1st Int. Conf. on Speech Prosody, Aix-en-Provence, 527-531.