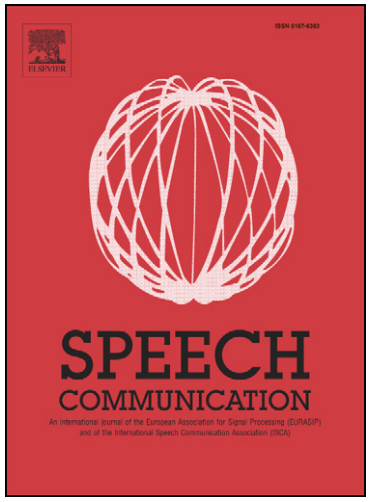
Evaluation of automatic break insertion for an agglutinative and inflected language

Eva Navas, Inmaculada Hernáez, Iñaki Sainz



Evaluation of automatic break insertion for an agglutinative and inflected lan guage

Eva Navas, Inmaculada Hernáez, Iñaki Sainz

PII: S0167-6393(08)00026-5
DOI: 10.1016/j.specom.2008.03.003
Reference: SPECOM 1692

To appear in: *Speech Communication*

Received Date: 13 June 2007
Revised Date: 14 November 2007
Accepted Date: 10 March 2008

Please cite this article as: Navas, E., Hernáez, I., Sainz, I., Evaluation of automatic break insertion for an agglutinative and inflected language, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.03.003

Evaluation of automatic break insertion for an agglutinative and inflected language

Eva Navas[1], Inmaculada Hernáez, Iñaki Sainz
Departamento de Electrónica y Telecomunicaciones, University of the Basque Country, Alda. Urquijo s/n, 48013 Bilbao, Spain

Abstract

This paper presents the evaluation of automatic break insertion for standard Basque. Basque is an agglutinative and inflected language and POS features, widely used for other languages, are not enough to accurately predict the insertion of breaks in the text. Other morpho-syntactic features, like grammatical case and information about syntagms have also been taken into account. With a textual corpus specially gathered for this study where the sentence internal punctuation marks have been removed, CARTs have been used to predict break locations. After applying parameter selection to the whole morpho-syntactic feature set, the best features were employed to build two CARTs, one that gives the same importance to deletion and insertion errors, T1, and another one, T2, that tries to minimise insertion errors. The objective evaluation of the break insertion algorithms gives a κ statistic of 0.518 and an F of 0.757 for T1 tree. The algorithms have also been subjectively evaluated and although T1 had better objective measures, the number of serious errors made by this tree is larger than the number of serious errors made by T2.

Keywords: break prediction, prosodic phrasing, speech synthesis

1 Introduction

TTS systems must produce a natural sounding speech signal. Correctly placed breaks contribute greatly to this naturalness: they have influence in the grapheme to phoneme transcription due to the occurrence of external sandhi effects (such as the liaison in French), in the process of assigning durations to phones and in the generation of the intonation curve. Breaks give a rhythmic structure to the speech signal and in some cases they can even change the meaning of a sentence (Hirschberg and Prieto, 1994). However, even in error free texts, only part of the breaks is indicated in the text. Usually, punctuation marks divide the text and indicate how the elements are distributed. Although it is common to make a break when finding some punctuation marks such as full stops, exclamation marks, question marks, colon, semicolon and suspension points, other punctuation marks are not always realised with a break. This is the case of commas, parentheses, quotation marks and hyphens. In spite of having some logical indications of where to place the breaks, the reader usually places them where he/she thinks it is more convenient, often in places where no punctuation mark is present. This is the reason why a TTS system needs an algorithm to insert breaks in suitable places.

The task of automatically placing breaks in texts is undoubtedly a very difficult one. Due to the great influence in the quality of the produced synthetic speech that correct break positions have, usually TTS systems allow the use of the XML <break> tag to insert precise pauses. This is the case of the systems from AT&T Natural Voices™ [2], Festival[3], IBM[4] and VERBIO Speech Technologies[5]. <break> is an element defined in most of markup languages for speech synthesis, like Java Speech API Markup Language Specification (JSML), Speech Synthesis Markup Language (SSML), Virtual Human Markup Language (VHML) and Voice XML.

A particular framework where a break insertion algorithm is critical is a Speech to Speech translation (S2ST) task. The goal of the S2ST research is to enable interpersonal communication via natural spoken language for people who do not share a common language. A S2ST system makes use of very different technologies: automatic speech recognition (ASR) to transcribe the source speech, machine translation (MT) to convert the text from the source language to the target language and text to speech (TTS) conversion to produce the final speech signal in the target language. The TTS system must deal with text produced by the translation system, which in turn, uses the output text from the recognition system. Errors made by the two systems will reach the input of the TTS system which will have to cope with them. One important aspect prone to errors in this context is punctuation. Recognition and translation of punctuation is a difficult task.

[1] Corresponding author. Tel: +34 94 601 7306; fax: +34 94 601 4259.
E-mail addresses: eva.navas@ehu.es, inma.hernaez@ehu.es, inaki.sainz@ehu.es
[2] http://www.research.att.com/~ttsweb/tts/
[3] http://www.cstr.ed.ac.uk/projects/festival/
[4] http://www-306.ibm.com/software/pervasive/voice_server/
[5] http://www.verbio.com/webverbio2/html/productes.php?lang=en#

Speakers divide their utterances into prosodic phrases separated by weak or strong prosodic boundaries. These boundaries can be perceptually classified according to the ToBI break indices (Silverman et al., 1992). Prosodic phrase break prediction can be decomposed into two tasks: prediction of break location and prediction of its acoustic realization. This paper deals with the former one, i.e. the building of an automatic model for the prediction of the positions of breaks.

Early models of break insertion used rules that searched for the best place to insert a break taking into account the word context (Bachenko and Fitzpatrick, 1990). Liberman and Church (1991) decide the insertion of a break using the punctuation marks and the presence of a function word following a content word. This simple algorithm has been widely used in TTS systems, like in the MITalk TTS system (Allen et al., 1987) and the multilingual one from Telefónica (Castejón et al., 1994). These are simple models that have good performance in most of the cases. But they are not suitable for agglutinative and inflected languages, where very few function words are used.

The goal of this study is to build a good break prediction algorithm for standard Basque that could be used in a S2ST application using available tools for automatic morpho-syntactic analysis and to check the validity of morphological and syntactic information for this task in Basque. Since Basque is an agglutinative and inflected language morpho-syntactic features not commonly used for break location prediction, like grammatical case and information about the grouping of words into syntagms, will be used.

The following sections are organised as follows: Section 2 presents the particularities of Basque related to this work. Section 3 describes the main characteristics of the corpus gathered for this study and presents the shallow analysis of the breaks in relation to the morpho-syntactic labels. Section 4 presents the statistical analysis carried out over the different predicting features considered and the predicting models built. It also describes the parameters used for the objective evaluation of the system performance. The objective and subjective evaluation of the developed algorithms is described in section 5. Finally, section 6 presents the discussion of the results.

2 Main features of Basque

Basque language is official in the Autonomous Basque Community (Spanish area), so it is possible to study in Basque in all levels of education, although graduate level is limited to a small number of curricula, and post-graduate studies are almost non-existent. The language has a very high dialectal fragmentation with 7 main dialects and more than 50 varieties according to modern commonly accepted assumptions. A standardisation process of written Basque started on 1975, unifying and regulating the lexicon, the grammar, the syntax, the morphology and every other aspect of the written language. This process is still under way, particularly concerning lexica.

Basque is not an Indo-European language and differs considerably in grammar from the languages spoken in surroundings regions. It is an inflectional language in which grammatical relations between components within a clause are represented by suffixes. This is a distinguishing feature since the morphological information of the words is richer than in surrounding languages. Basque is an agglutinative language; that is, most words are formed by joining morphemes together. In particular, the affixes corresponding to the determination, number and grammatical case are taken in this order and independently of each other. One of the main characteristics of Basque is its declension system with numerous cases (more than 20), which differentiates it from the languages of the nearby countries. The inflections of determination, number and case appear only after the last element in the noun phrase. It is an extremely inflected language, inflecting both nouns and verbs. For an extensive review on Basque grammar, see Hualde and Ortiz de Urbina (2003).

As a minority language, language technology development for Basque differs in several aspects from the development for widely used languages. On the one hand, the size of the speakers' community is small. Consequently it has to face up to the scarcity of the resources and tools that could make possible this development at a reasonable and competitive rate. On the other hand, there are language-specific problems, related to language typology. It is not always possible to use or to adopt the language technologies developed for other languages. This is especially relevant in rule-based approaches, but also in corpus-based approaches, because truly efficient exploitation of corpus demands annotation, and this process is in most cases based on rule-based procedures, like morphological and syntactic analysis. For example, romance languages like Galician, Catalan or Occitan can take advantage of NLP developments for French or Spanish, but these developments are not so applicable to some languages, for example Basque. This applicability (or portability) depends largely on language similarity. As stated before, Basque is an agglutinative language, with a rich flexional morphology, and this requires specific procedures for language analysis and generation. For a review of resources and tools available for Basque see Díaz de Ilarraza et al. (2003).

3 Corpus

Since there was no suitable corpus in standard Basque for the study of the locations of breaks, a new corpus had to be collected. A textual database is considered appropriate to study break insertion because it is easy to collect and it provides with all the information needed. We have gathered a corpus introducing break labels in a written text, following the methodology proposed by Hirschberg and Prieto (1996). Texts were collected from one electronic newspaper and one electronic magazine, as well as from other Internet pages. The coverage of this corpus was analyzed using a large Basque Corpus of 1 100 000 words as a reference for the Basque language. With the corpus selected for this study the 64.4% of the large Basque Corpus is covered.

Our study makes the assumption that a break is always made when finding the following punctuation marks associated with a sentence boundary: full stops, exclamation marks, question marks, colon, semicolon and suspension points. In the framework of a S2ST system these breaks will be inserted in the TTS input text according to the pauses detected by a voice activity detector (VAD) working on the source speech. The correlation between pause durations and sentence boundary marks was studied for broadcast news data (Gotoh and Renals, 2000) and it was observed that the longer the pause duration, the greater the chance of a sentence boundary existing. There have been several studies that have been able to detect sentence boundaries from speech and recognized text with success (Hakkani-Tür et al., 1999; Kim and Woodland, 2001).

However, for an speech recognition system it is very difficult to place the sentence internal punctuation marks, like commas, quotation marks and parentheses, even taking into account acoustic information from the source speech (Chen, 1999; Christensen et al., 2001). In the context of S2ST systems, these sentence internal punctuation marks will not be indicated or at least will not be reliable. So for our purposes, punctuation marks not indicating end of utterance were removed from the corpus, and the prediction of breaks will be made without this information.

In this work, both sentence internal and sentence boundary breaks will be predicted. Sentence internal breaks have to be predicted because they can be placed at points where no punctuation mark gives an indication, and because in our work all the information about internal punctuation marks has been removed as stated above. Sentence boundaries must be predicted because not all punctuation marks are related with a sentence boundary: full stops can be found in abbreviations, as decimal separator in numbers and in WEB addresses, question marks can be found mid-sentence…

3.1 Annotating the breaks

The breaks were manually labelled by a Basque linguist, who hand labelled the corpus for likely prosodic boundaries. He was instructed to insert a break symbol wherever he thought a prosodic boundary was necessary when reading aloud the text. Only one level of break was considered in this study, hence a word juncture is considered to be either a break or a non break. As an example of the sentences from the corpus and the break annotation made, the two following sentences are presented, together with their closest translation to English[6] (The sign | has been used to indicate the annotation of a break by the labeller):

- Sentence 1:

Errusiar honek | bost urte eman zituen | egunean hiru litro coca-cola edaten. |
(Errusiar) (honek) (bost) (urte) (eman zituen) (egunean) (hiru) (litro) (coca-cola) (edaten)
(Russian person) (this) (five) (years) (spent) (a day) (three) (litres) (coca-cola) (drinking)
This Russian person spent five years drinking three litres of coca-cola a day.

- Sentence 2:

Askok esperoko ez duten arren | Garcia da Euskal Herriko biztanleen artean | gehien errepikatzen den abizena. |
(Askok) (esperoko) (ez) (duten) (arren) (Garcia) (da) (Euskal) (Herriko) (biztanleen) (artean) (gehien) (errepikatzen) (den) (abizena)
(A lot of people) (expect) (no) (would) (although) (Garcia) (is) (Basque) (Country) (inhabitants) (among) (more) (repeated) (that is) (surname)
Although not a lot of people would expect it Garcia is the surname that is more repeated among the inhabitants of the Basque Country.

The main features of the labelled corpus are presented in table 1. As this table shows there is only 17.5% function words in this corpus. This small proportion implies that a simple algorithm based in

[6] The first line shows the sentence in Basque. The second and third lines group the words in Basque and English in order to have a one-to-one correspondence between each group in the two languages. The forth line reorders the words in English to obtain a correct English sentence. The breaks are not marked in the English version, due to the differences in word ordering between both languages.

content/function word discrimination cannot be used. In the corpus, only 9.4% of the annotated breaks followed a function word, and only 5.5% of the breaks were preceded by a function word. Therefore, most of the breaks occur between content words.

The number of syllables between breaks was also analyzed and compared with the distribution of the number of syllables in the sentences. These distributions are presented in figure 1. The mean number of syllables between breaks is 13.8, while the mean number of syllables in the sentences is 33.1.

3.2 Labelling the corpus

Traditionally part of speech (POS) information has proven to be very useful for the prediction of breaks in several languages like English (Black and Taylor, 1997), French (Boula de Mareüil and d'Alessandro, 1998), Thai (Tesprasit et al., 2003), Spanish (Bonafonte and Agüero, 2004), German (Apel et al., 2004), Dutch (Marsi et al., 2003), Greek (Maragoudakis et al., 2003), Mandarin (Zheng et al., 2004) and Korean (Kim et al., 2006). This information has also been used in standard Basque (Navas et al., 2002). But, since Basque is a highly inflected language, other grammatical information that to our knowledge has never been used for the prediction of breaks, like grammatical case, can also be taken into account.

The corpus used in this study was morphologically and syntactically annotated, using some tools developed by the IXA research group[7]. For the POS tagging, the morphological analyzer for Basque MORFEUS (Ezeiza et al., 1998) has been used. This program provides the lemma of each word and a set of main tags, corresponding to the POS or category of the word. Most of them have two more levels of subcategorization, according to grammatical considerations: the subcategory and the grammatical case. The accuracy of MORFEUS is 90% when category, subcategory and case are taken into account (Ezeiza et al., 1998). The number of different tags supplied by MORFEUS was too large and detailed for our purposes. Besides, giving non significant features to a classifier introduces noise that may degrade its performance (Read and Cox, 2004). So, the full tag set was reduced, especially at the level of subcategory. The tag sets used for the category, subcategory and case, together with the corresponding description and an example in Basque with the closest translation to English are listed in tables 2, 3 and 4. These tables also give the percentage of words with a specific label that were followed by a hand annotated break in the corpus.

Table 2 shows that category PUNT, corresponding to punctuation marks, is followed by a break in the 99.2% of the cases. Taking into account that sentence internal punctuation marks (like commas) have been removed from the corpus, a 100% has to be expected. Actually, the cases with PUNT category not followed by a break are due to dots following abbreviations that have been misinterpreted as punctuation marks by the automatic labelling system. Category LAB (abbreviations) is always followed by a break in the corpus. In fact, only three abbreviations appear in the corpus and they happen to be followed by a break. This datum is not statistically meaningful, but most abbreviations are placed after the noun (like mister, department, street) and would fulfil the same conditions. In this table we can also see that the probability of having a break after categories SIG (acronyms) and PRT (particles) is very small. Finally categories BST (others) and IOR (pronouns) are almost never found before a break.

Concerning subcategory labels, shown in table 3, the following tags are never followed by a break in the corpus: ELK (reciprocal pronouns), FAK (factitive verbs), IGB (indeterminate pronouns) and ZIU (security adverbs). On the other hand, the subcategory LOK (connectives), a subcategorization of category LOT (conjunctions) is followed by a break in more than 70% of cases.

When applying this same analysis to grammatical case, we can see in table 4 that both genitive cases GEL (for places) and GEN (for people) are almost never followed by a break. Case DESK (descriptive) is never followed by a break while case DES (benefactive) is followed by a break in more than 70% of cases.

The tag NONE was added to the subcategory and case sets, for the categories that have no subcategory or grammatical case. In the same way the tag MISSING was used when the word should have subcategory or grammatical case, but the automatic labelling system did not assign any tag.

Syntactic information has also been used in the prediction of breaks (Ingulfsen et al., 2005; Koehn et al., 2000; Ostendorf and Veilleux, 1994). Although there is not a direct mapping between syntactic and prosodic structures (Frazier et al., 2004), there is undoubtedly a relation between them. Syntactic information for standard Basque was obtained with an automatic labeller developed also by the IXA research group (Aduriz et al., 2004). This tool provides us with the syntactic function of each word and also groups the words into syntagms. Breaks will likely occur between syntagms, so information related with the grouping of words might be very valuable for the break insertion algorithm. The performance of

[7] IXA research group of the University of the Basque Country (http://ixa.si.ehu.es/)

this tool is 83% precision rate (correctly selected groups / number of groups returned) and 81.4% recall rate (correctly selected groups / actual groups in the sentence).

The set of syntactic labels provided by this tool was reduced in the same way as POS, subcategory and case tag sets. The values considered in this study are listed in table 5. As table 5 shows, label MP (subordinate sentence) is always followed by a hand annotated break, while the probability of making a break after JADNAG_MP_KM (subordinate main verb as noun), JADNAG_MP_IZLG (subordinate main verb as adjectival complement), IZLG (adjectival complement) and PJ (coordinating conjunction) is very small.

Table 6 shows the labels used for the grouping of words. In this table we can see that words in the beginning or inside the syntagms have very small probability of having a break after them. In fact, the cases where this happens are due to labelling errors.

For all types of information obtained from the morphological and syntactic labellers, tags within a window of five words have been considered. This window size has been chosen taking into account that the average length of each syntagm is 2.17 words, therefore to include (in average) all the words in the current word syntagm the following and preceding two words must be considered. This window has been centred in the word preceding the break to predict.

4 Experiments

The prediction of break locations is a classification problem, where each inter-word space or juncture is a potential phrase break and the classifier task is to decide whether the juncture is a phrase break. Many techniques have been applied to this problem, such as Classification and Regression Trees (CART) (Apel et al., 2004; Koehn et al., 2000; Sangho and Yung-Hwan, 1999; Wang and Hirschberg, 1992), Memory Based Learning (MLB) (Busser et al., 2001; Zhao et al., 2002), Transformational Rule Based Learning (Fordyce and Ostendorf, 1998), Hidden Markov Models (HMM) (Bell et al., 2006; Schmid and Atterer, 2004; Taylor and Black, 1998), N-grams combined with CARTs (Sun and Applebaum, 2001), Finite State Transducers (FST) (Veilleux et al., 1990), Support Vector Machines (SVM) (Zhao et al., 2002), Bayesian induction (Zervas et al., 2003), Maximum Entropy Models (ME) (Li et al., 2004) and Logistic Generalized Linear Models (Logistic GLM) (Li et al., 2006).

In this work CART (Breiman et al., 1984) has been used as classification technique. The trees have been built using 10-fold cross validation (Stone, 1974) and taking into account the priors in the corpus (20.7% breaks vs. 79.3% non breaks).

4.1 Error measures

The evaluation of the performance of the break insertion algorithms is not trivial. Several efficiency measures, like percentage of junctures correct, percentage of breaks correctly predicted and ratio of false insertions to the total number of junctures, have been traditionally proposed for the objective evaluation of the break insertion algorithms (Taylor and Black, 1998). To calculate all these measures a juncture wrongly classified is considered an error. In our classification problem there are two types of possible errors: insertions (I), when the algorithm makes a break that was not in the reference database and deletions (D), when the algorithm does not make a break that was indicated in the reference database.

The overall score achieved by the tree is calculated as the number of junctures correctly classified divided by the total number of inter-word spaces. It is calculated according to equation 1.

$$S(\%) = \frac{N - D - I}{N} * 100 \quad (1)$$

where
- N stands for total number of junctures
- D stands for total number of deletions
- I stands for total number of insertions

This datum has to be interpreted carefully, because it depends on the proportion of breaks in the original corpus. In our test database 81.05% of the inter-word spaces were labelled as non-breaks, therefore an algorithm that does not place any break at all would achieve an overall score of 81%. To avoid this problem another measure to compute the score of the algorithms, the kappa statistic, has been proposed. This statistic was first suggested for linguistic classification tasks by Carletta (1996) and has since been used by others to avoid the dependency of the score on the proportion of non-breaks in the text (Navas et al., 2002; Sanders, 1995; Torres and Gurlekian, 2004).

The kappa statistic is calculated as indicated by equation 2.

$$\kappa = \frac{S - \frac{NB}{N}}{1 - \frac{NB}{N}} \quad (2)$$

where S stands for the overall score
NB stands for the total number of non-breaks in the data
N stands for the total number of junctures

In expression (2), the overall score achieved by the tree is compared with the probability of having a non-break label in the data, eliminating the dependency on the structure of the data. If the algorithm does not insert any break, the value of the kappa statistic will be 0. If the method predicts every inter-word space correctly, κ equals 1. Values lower than 0 indicate that the breaks placed by the algorithm are in the wrong places, so it would be better not to use it.

Other appropriate measures for this problem are precision (P) and recall (R). They are widely used in the field of information retrieval to evaluate the search performance (Salton, 1972). These values and their weighted harmonic mean, known as balanced F-score (van Rijsbergen, 1979), have also been applied to evaluate the performance of break prediction algorithms (Busser et al., 2001; Li et al., 2004; Read and Cox, 2007; Sun and Applebaum, 2001). In this context, precision evaluates the proportion of breaks correctly identified to all the predicted breaks and recall measures the proportion of breaks correctly identified, out of all manually annotated breaks in the corpus.

The F-score is also known as the F1 measure, because recall and precision are evenly weighted. It is calculated according to equation 3.

$$F = \frac{2P.R}{(P+R)} \quad (3)$$

where P stands for precision and R for recall. If no breaks are inserted P and F are undefined and R is zero. If the algorithm inserts breaks, but places them in incorrect places, P and R are zero and F is undefined.

4.2 Feature selection

Choosing which features to keep and which to discard is not a trivial question in prediction problems. The full set of features used to predict the location of breaks may not be optimal. Certain attributes might be noisy and could compromise the accuracy of the prediction. The challenge is to remove the noisy attributes from the database, without losing the usefulness of the data. The feature selection methods are often performed according to the nature of the data, and therefore they are not generally applicable to all kinds of data analyses (Blum and Langley, 1997; Langley, 1994).

Another alternative for parameter selection is the use of methods that rely on identifying potentially discriminating attributes by using some statistical measures of the information content of attributes, such as their correlation coefficients. The motivation behind these methods is that two well-correlated attributes contain much the same information, and therefore one of them could be eliminated from the analysis. In doing so, we would reduce some noise in the data, thus hopefully improving the prediction accuracy.

In our database all the features are categorical. To study the correlation among them a Pearson's Chi-Square test (Siegel and Castellan, 1988) was applied. With this purpose the necessary contingency tables were obtained and the Contingency Coefficient C was calculated according to equation 4:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \quad (4)$$

where N stands for the number of observations and χ^2 is Chi square value. This C coefficient was then corrected to adjust its variation to the range 0-1, obtaining the Sakoda's coefficient C^* (Agresti, 1996). The value of Sakoda's coefficient measures the association between two categorical variables. A value of 0 for this coefficient means that both categorical features are independent. It is calculated according to equation 5.

$$C^* = \frac{C}{\sqrt{\frac{k-1}{k}}} \quad (5)$$

where k is the minimum between r (number of rows of the corresponding contingency table) and c (number of columns of the corresponding contingency table).

Sakoda's coefficient between all the predictor variables and break class has been calculated to know the theoretical power of prediction of each of them. Likewise Sakoda's coefficient between each pair of predicting features has also been obtained, to estimate the degree of correlation between them. The results of this study are presented in table 7. According to this analysis the variable more related to break is syntactic label (with a C^* of 0.368), followed by POS (C^*=0.362) and subcategory (C^*=0.350). The

predicting features that are more related are POS and subcategory, and the more independent ones case and syntagm.

To know which kind of information has actually more influence in the prediction of the placing of breaks, the predicting information was combined in different trees and their performance was evaluated using the test corpus. A set of trees trained with only one type of feature was built. Their results are shown in table 8. Among them the best tree is the one that uses syntactic labels with a κ of 0.498 and an F of 0.713. This is the feature that shows the bigger correlation with break class according to Sakoda's contingency coefficient calculated.

Then all the possible trees using two types of features were built. In this case the tree using syntactic and POS labels is the one with better results among all the possible combinations. In accordance with the statistical analysis, POS is the feature with the second best correlation with break class.

When training the set of trees that use three types on information, the next relevant feature turns out to be grammatical case. According to Sakoda's contingency coefficient, the next relevant feature should be subcategory. But looking to the coefficient between subcategory and POS (0.613) and subcategory and syntactic information (0.600), we can see that subcategory is correlated with both of them. This may be the reason why this is not the next best feature to consider in the prediction of breaks using the tree. Although it is correlated with break class, it does not add enough information to the system.

Information about syntagms is the next significant type of information to take into account in the prediction of breaks. Adding the subcategory makes the results worse. The best results for each number of parameters used are presented in figure 2.

Taking these results into account the break prediction trees have been built without subcategory information.

5 Results

For the prediction experiments, the corpus is divided in 75% for training and 25% for testing purposes. Two different prediction trees have been built. The first one, T1, gives the same importance to both types of predicting errors (insertions and deletions). In the second one, T2, insertion errors are considered more harmful and therefore, these insertion errors have been given a weight of 1.33. This coefficient has been chosen by testing a few different weights on specific test examples.

5.1 Objective evaluation

To allow an objective comparison of the trees' performance, we consider a fictitious break insertion system that offers the minimum necessary performance. This minimum performance would be the correct placement of all the breaks corresponding to sentence boundaries. This fictitious system is used as a baseline to evaluate the performance of the break insertion algorithms developed. With this baseline system, no intra sentence break would be marked; it would only place breaks associated with the punctuation marks corresponding to sentences boundaries. The results of this baseline system for our test database are shown in table 9, together with the ones achieved by the two trained trees.

Baseline system has a precision of 100%, because all the breaks it indicates are correct. However, its recall is very low, because it misses a lot of breaks (all the sentence-internal breaks). Both trees built have worse precision but better recall than the baseline system, achieving a value of F above 70%. Objective results for T1 are slightly better (F 75.7%, κ 0.518) than the ones obtained for T2 (F 72.5%, κ 0.511). T1 has better recall (it places more breaks from the breaks to be placed) but worse precision (from the breaks it places, fewer are correct) than T2.

5.2 Subjective evaluation

Measuring the efficiency of a break insertion algorithm by comparing its results with the text labelling has the disadvantage that the subjective importance of the error is not considered. Moreover, all the errors have the same influence in the final measure, although some of them are more serious than others. To have an estimation of the subjective performance of the algorithms a subjective evaluation was carried out. For a perceptual evaluation either a TTS system or a real speaker must be used to read aloud the text. However, we thought that this would introduce more variability in the evaluation results. Since there are many different ways to realize the breaks, the results would have a strong dependence on the acoustic realization of the breaks.

30 sentences from the test corpus were randomly selected for the evaluation. The shortest one has 9 words and the longest one 29. 15 people took part in the subjective evaluation. All of them are fluent in standard Basque.

Evaluators were asked to insert the appropriate breaks in the sentences. They had to annotate the text introducing the likely prosodic boundaries they considered missing. The occurrence of pauses is strongly

speaker dependent (Zellner, 1994) and each evaluator will insert the breaks he/she considers appropriate. Their labelling was compared with the reference labelling and the prediction made by the two trained trees. After this comparison the breaks were classified into these three classes:

- Correct breaks: breaks labelled by the algorithm which have also been labelled by any of the evaluators.
- Omitted breaks: breaks labelled by any of the evaluators, but not predicted by the algorithms.
- Wrong breaks: breaks predicted by the algorithms that were not labelled by any of the evaluators.

The following data are used as a measure of the performance:

- The absolute number of correct, omitted and wrong breaks for each algorithm. The larger the quantity of correct breaks and the smaller the quantity of the omitted and wrong breaks are the better.
- Mean score of the breaks, i.e., the mean number of evaluators that selected the breaks. This quantity has sense only for correct and omitted breaks. In the case of correct breaks, the bigger this mean score is the better, because this means that more evaluators agree with the breaks inserted by the algorithm. For omitted breaks, the smaller the mean score the better the result is. In this case, small mean scores indicate that few evaluators selected these breaks, therefore they are not very important.
- Number of breaks selected by at least 80% of the evaluators. This measure is calculated for correct and omitted breaks and gives an idea of the number of breaks that are important for the majority of the evaluators.

Results of each of the algorithms are presented in table 10.

As expected, the reference labeller achieves the best scores. His number of correct breaks is 79, selected by 68% of the evaluators. Among these 79 breaks, 39 have been selected by at least 80% of the evaluators. Therefore, the evaluators strongly agree on these 39 breaks. T1 and T2 have fewer correct breaks, 62 and 45 respectively, but the correct breaks they insert are the very important ones (with scores of 0.71 and 0.74). Considering omitted breaks, the reference labeller also has the best results. He has omitted only 52 breaks of the ones indicated by the evaluators, with less than a third of the evaluators having selected these 52 breaks on average. T2 omits more breaks than T1 (86 vs. 69), as it tries to minimize insertion errors. The breaks it omits are judged necessary by 42% of the evaluators. The reference labeller has only 3 wrong breaks which no evaluator has selected. This confirms that there is not only one correct way to insert breaks in a text, and that the agreement with our reference information is not complete. T2 has fewer wrong breaks than T1: it has been built to minimize insertion errors, so it inserts fewer correct breaks, but also fewer wrong breaks.

6 Discussion

After analysing the results of these break location prediction experiments, a first conclusion is that placing breaks is a speaker dependent task. There is only partial agreement among speakers about the places where breaks have to be inserted. Some break positions are very clear and the agreement is almost complete (in our test 21% of the breaks were marked by more than 80% of the evaluators), but some others have a strong dependency on the evaluator (19% of the breaks marked were selected by less than 20% of the evaluators).

Another important conclusion is that there are some places where a break cannot be made: insertion errors are more critical than deletion errors. The T2 algorithm, which places fewer breaks than the T1 algorithm, produces fewer wrong breaks. One break wrongly placed causes a bad subjective impression that degrades notably the quality of the synthetic speech signal. Although T1 has better objective results, taking into account the number of wrong breaks, in practice it may be better to use T2.

Among the considered morpho-syntactical features, information about the syntactic function of the word is the more determinant one. In standard Basque prepositional functions are realized by case suffixes inside word-forms and there are fewer function words than in other languages. POS information is not enough for the prediction of break locations and grammatical case has also to be taken into account.

It is extremely difficult to compare the results achieved in break location prediction with other similar works. On the one hand the databases used are very different in most cases. The treatment of punctuation information is also very different in each study: Some works include all the punctuation marks which have a great correlation with breaks and make the prediction task easier. Other works completely eliminate punctuation marks. In our case, only sentence internal punctuation marks have been eliminated, due to the considerations about the S2ST framework where our TTS system must work. Another issue that makes the comparison of the results more difficult are the measures presented to evaluate the performance of the algorithms. However, just for reference, we briefly state below some recent results in the prediction of break location. F score has been calculated when possible, if not directly provided by the authors.

Read and Cox (2007) achieve a best F of 80.19% with reduced POS tag set and syntactic information over MARSEC corpus. The predicting technique applied in this case was decision trees. Bell et al. (2006) working on the same MARSEC corpus got an F of 83.3% and an F of 82.1% on the Boston Radio Corpus, applying HMM in both cases. Pfitzinger and Reichel (2006) dealt with this problem in German. They used the IMS Radio News Corpus and predicted the ToBI break indices with a global F of 87.72%. Yi et al. (2006) applied a Logistic Generalised Linear Model to Mandarin assuming that phrase breaks follow a Bernoulli distribution. They got an F of 62.32% without any information about punctuation. Kim et al. (2006) worked for Korean which is an agglutinative language. They used a General Maximum Entropy Model with POS and global and second order features such as distance between current junction and previous or next phrase break. Their model has an F of 79.6%, without using information about punctuation. Another work dealing with break prediction for Korean is the one presented in Yoon (2006). He built a CART with POS and syntactic features and achieved an F of 71.3% for intonational break prediction, using information about punctuation. Zervas et al. (2005) use also decision trees and POS information and have a $\kappa > 0.75$ using a set of 11 POS labels. They worked for Modern Greek which is an inflected language. Oparin (2005) uses a rule based approach for the prediction of break location in Russian (an inflected language) with an F of 77.8%. Ingulfsen et al. (2005) using punctuation marks, POS and link grammar achieved an F of 79.4% over Boston Radio Corpus. Cox (2005) attained an F of 81.2% in the same Boston Radio Corpus, considering POS information and some other attributes derived with a sentence parser. In this work, Generalised Probabilistic Descent (GPD) was applied to a maximum entropy classifier.

Only most recent published works had been referred, but of course there have been many more works, which reflects the interest and importance of the problem.

Among the continuing lines of this work is the consideration of different window lengths and centrings for the labels considered in the prediction. Although most studies use a window centred in the word under study, some works indicate that considering more tags before the current word than after it can improve the results (Maragoudakis et al., 2003; Schmid and Atterer, 2004; Sun and Applebaum, 2001).

Finally, the subjective evaluation process could be improved by allowing the quantisation of the correctness of the inserted breaks. This would simplify the interpretation of the results.

7 Acknowledgements

We are grateful to Nerea Ezeiza and the IXA research group for the automatic morphological and syntactic labelling of the corpus and to Iñaki Gaminde for the manual labelling of the breaks.

The authors would also like to thank all the people that took part in the subjective evaluation.

This work has been partially founded by Spanish Ministry of Education and Sciences under grant TEC2006-13694-C03-02 (AVIVAVOZ project, http://www.avivavoz.es/) and by Basque Government under grant IE06-185 (ANHITZ project, http://www.anhitz.com/).

8 References

Aduriz I., Aranzabe M., Arriola J., Díaz de Ilarraza A., Gojenola K., Oronoz M., Uria L., 2004 A Cascaded Syntactic Analyser for Basque Lecture Notes on Computer Science, 2945, pp. 124-135.

Allen, J., Hunnicut, S., Klatt, D. 1987. From Text to Speech: the MITalk System. Cambridge University Press, Cambridge, United Kingdom.

Agresti, A., 1996. Introduction to Categorical Data Analysis. Wiley, New York.

Apel, J.; Neubarth, F.; Pirker, H.; Trost, H., 2004. Have a break! Modelling pauses in German speech, in: Buchberger, E. (Ed.), KONVENS, OEGAI, Vienna; Austria, pp. 5-12.

Bachenko, J., Fitzpatrick, E., 1990 A Computational Grammar of Discourse-Neutral Prosodic Parsing in English. Computational Linguistics, 16(3), pp. 155-170.

Bell, P., Burrows, T., Taylor, P., 2006. Adaptation of prosodic phrasing models, in: proceedings of Speech Prosody 2006.

Black, A.W.; Taylor, P., 1997. Assigning phrase breaks from part-of-speech sequences, in proceedings of Eurospeech'97, Rhodes, pp. 995-998.

Blum A, Langley P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, pp. 245-271.

Boula de Mareüil, P., d'Alessandro, C., 1998. Text Chunking for Prosodic Phrasing in French, in proceedings of 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 127-132.

Bonafonte, A., Agüero, P.D., 2004. Phrase break prediction using a finite state transducer, in AST 2004, 11th International Workshop on Advances in Speech Technology, Maribor, Slovenia.

Breiman, L., Friedman, J.H., Olsen., R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman&Hall, USA.

Busser, G., Daelemans, W., van den Bosch, A., 2001. Predicting phrase breaks with memory-based learning, in: proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis. Perthshire Scotland, pp. 29-34.

Carletta, J. C., 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), pp. 249-254.

Castejón, F., Escalada, J.G., Monzón, L., Rodríguez, M.A., Sanz, P., 1994. Un Conversor Texto-Voz para Español. Comunicaciones de Telefónica I+D, 10, paper 8.

Chen, C. J., 1999. Speech recognition with automatic punctuation, in: proceedings of EUROSPEECH'99, pp. 447-450.

Christensen, H.,Gotoh, Y., Renals, S., 2001. Punctuation Annotation using Statistical Prosody Models, in: proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 35-40.

Cox, S.,2005. A discriminative approach to phrase break modelling, in: proceedings of INTERSPEECH-2005, pp. 3229-3232.

Díaz de Ilarraza A., Sarasola K., A. Gurrutxaga, I. Hernaez, N. Lopez de Gereñu 2003. HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities, in: proceedings of Workshop on NLP of Minority Languages and Small Languages.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R., 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages, in: proceedings of COLING-ACL'98, pp 379-384.

Fordyce, C.S., Ostendorf, M., 1998. Prosody prediction for speech synthesis using transformational rule-based learning, in: proceedings of ICSLP-1998, vol. 3, pp. 843-846.

Frazier, L., Clifton, C., Carlson, K., 2004. Don't break, or do: prosodic boundary preferences. Lingua, vol. 114(1), 3-27.

Gotoh, Y., Renals, S., 2000. Sentence boundary detection in broadcast speech transcripts, in: proceedings of ASR-2000, pp. 228-235.

Hakkani-Tür, D., Tür, G., Stolcke, A., Shriberg, E., 1999. Combining words and prosody for information extraction from speech, in: proceedings of EUROSPEECH'99, pp. 1991-1994.

Hirschberg, J., Prieto, P., 1994. Training Intonational Phrasing Rules Automatically for English and Spanish Text-to-Speech, in: proceedings of Second ESCA/IEEE workshop on speech synthesis, pp. 159-162.

Hirschberg, J., Prieto, P., 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. Speech Communication, Vol. 18, pp. 281-290.

Hu, G.P., Chen, B.C., Fan, M., Wang, R.H., 2003. The contribution of mutual information in the intonational phrase prediction in Chinese text, in: proceedings of Natural Language Processing and Knowledge Engineering, pp. 407-412.

Hualde, J.I., Ortiz De Urbina, J. (eds.), 2003. A Grammar of Basque. Mouton de Gruyter, Berlin.

Ingulfsen, T., Burrows, T., Buchholz, S., 2005. Influence of syntax on prosodic boundary prediction, in: proceedings of INTERSPEECH-2005, pp. 1817-1820.

Kim, J., Woodland, P. C., 2001. The use of prosody in a combined system for punctuation generation and speech recognition, in: proceedings of EUROSPEECH-2001, pp. 2757-2760.

Kim, S., Lee, J., Kim, B., Lee, G., 2006. Incorporating second-order information into two-step major phrase break prediction for Korean, in: proceedings of INTERSPEECH-2006, paper 1487.

Koehn, P., Abney, S., Hirschberg, J., Collins, M., 2000. Improving Intonational Phrasing with Syntactic Information, in: proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 3, pp. 1289-1290.

Langley, P., 1994. Selection of relevant features in machine learning, in: proceedings of the AAAI Fall Symposium on Relevance, pp. 140-144.

Li, J., Hu, G., Wang, R., 2004. Chinese prosody phrase break prediction based on maximum entropy model, in: proceedings of INTERSPEECH-2004, pp. 729-732.

Liberman, M., Church, K., 1991. Text analysis and word pronunciation in text-to-speech synthesis, Advances in Speech Signal Processing, S. Furuy, Sondhi, M.M., eds., Dekker, New York, pp.791-831.

Maragoudakis, M., Zervas, P., Fakotakis, N., Kokkinakis, G., 2003. A Data-Driven Framework for Intonational Phrase Break Prediction. Lecture Notes in Artificial Intelligence, 2807, pp. 189-197.

Marsi, E., Busser, G.J., Daelemans, W., Hoste, V., Reynaert, M., van den Bosch, M., 2003. Learning to predict pitch accents and prosodic boundaries in Dutch, in: proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 489-496.

Navas, E., Hernáez, I., Ezeiza, N., 2002. Assigning Phrase Breaks using CART's in Basque TTS, in: proceedings of the 1st Int. Conf. on Speech Prosody, pp. 527-531.

Oparin, I., 2005. Robust Rule-Based Method for Automatic Break Assignment in Russian Texts. Lecture Notes in Computer Science, vol. 3658, pp. 356-363.

Ostendorf, M., Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computational Linguistics, 20(1), pp. 27-54.

Pfitzinger, H.R., Reichel U.D., 2006. Text-based and Signal-based Prediction of Break Indices and Pause Durations, in: proceedings of Speech Prosody.

Read, I., Cox, S., 2004. Using part-of-speech for predicting phrase breaks, in: proceedings of INTERSPEECH-2004, pp. 741-744.

Read, I. Cox, S., 2007. Stochastic and syntactic techniques for predicting phrase breaks. Computer Speech and Language, 21, pp. 519-542.

Salton, G., 1972. The ”Generality” Effect and the Retrieval Evaluation for Large Collections, Journal American Society for Information Science, pp. 11-22.

Sanders, E., 1995. Using probabilistic methods to predict phrase boundaries for a text-to-speech system. Master's thesis, University of Nijmegen.

Sangho, L., Yung-Hwan, O., 1999. Tree-based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems, Speech Communication, vol. 28, pp. 283-300.

Schmid, H., Atterer, M., 2004. New statistical methods for phrase break prediction, in: proceedings of the 20th international conference on Computational Linguistics, paper 659.

Siegel, S., Castellan Jr., N. J., 1988. Nonparametric Statistics for the Behavioral Sciences, Second Edition, McGraw-Hill, New York.

Silverman, K. E. A., Beckman, M., Pitrelli, J. F., Ostendorf, M. , Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: A standard for Labeling English Prosody, in: proceedings of the 1992 International Conference on Spoken Language Processing, vol. 2, pp. 867-870.

Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society, 36, pp. 111-147.

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

Taylor, P., Black, A.W., 1998. Assigning phrase breaks from part-of-speech sequences. Computer Speech and Language, Volume 12, Issue 2, pp. 99-117.

Tesprasit, V., Charoenpornsawat, P., Sornlertlamvanich, V., 2003. Learning phrase break detection in Thai text-to-speech, in: proceedings of EUROSPEECH-2003, pp. 325-328.

Torres, H. M., Gurlekian, J. A., 2004. Automatic determination of phrase breaks for Argentine Spanish, in: proceedings of Speech and Prosody-2004, pp. 553-556.

van Rijsbergen, C.J., 1979. Information Retrieval, 2nd edition, Butterworths, London

Veilleux N.M., Ostendorf M., Price P.J., Shattuck-Hufnagel, S., 1990. Markov Modeling of Prosodic Phrase Structure, in: proceeding of the 1990 International Conference on Acoustics, Speech and Signal Processing, Vol 2, pp. 777-780.

Wang, M.Q., Hirschberg, J., 1992. Automatic Classification of Intonational Phrase Boundaries, Computer Speech and Language, volume 6, number 2, pp. 175-196.

Yi, L., Li, J., Lou, X., Hao, J., 2006. Phrase break prediction using logistic generalized linear model, in: proceedings of INTERSPEECH-2006, pp. 1308-1311.

Yoon, K., 2006. A prosodic phrasing model for a Korean text-to-speech synthesis system. Computer Speech & Language, vol. 20(1), 69-79.

Zellner, B., 1994. Pauses and the temporal structure of speech, in Keller, E. (Ed.) Fundamentals of speech synthesis and speech recognition, John Wiley and Sons, Chichester, pp. 41-62.

Zervas, P., Maragoudakis, M., Fakotakis, N., Kokkinakis, G., 2003. Bayesian Induction of intonational phrase breaks, in: proceedings of Eurospeech 2003, pp. 113-116.

Zervas, P., Xydas, G., Fakotakis, N., Kokkinakis, G., Kouroupetroglou, G., 2005. Experimental Evaluation of Tree-Based Algorithms for Intonational Breaks Representation. Lecture Notes in Computer Science, Vol. 3658, pp. 334-341.

Zhao S., Tao J., Cai L., 2002. Prosodic Phrasing with Inductive Learning, in: proceedings of ICSLP2002 Denver, USA, pp. 2417-2420.

Zheng, Y, Kim, B., Lee, G., 2004. Using multiple linguistic features for Mandarin phrase break prediction in maximum-entropy classification framework, in: proceedings of INTERSPEECH-2004, pp. 737-741.

FIGURE LIST

TABLE LIST

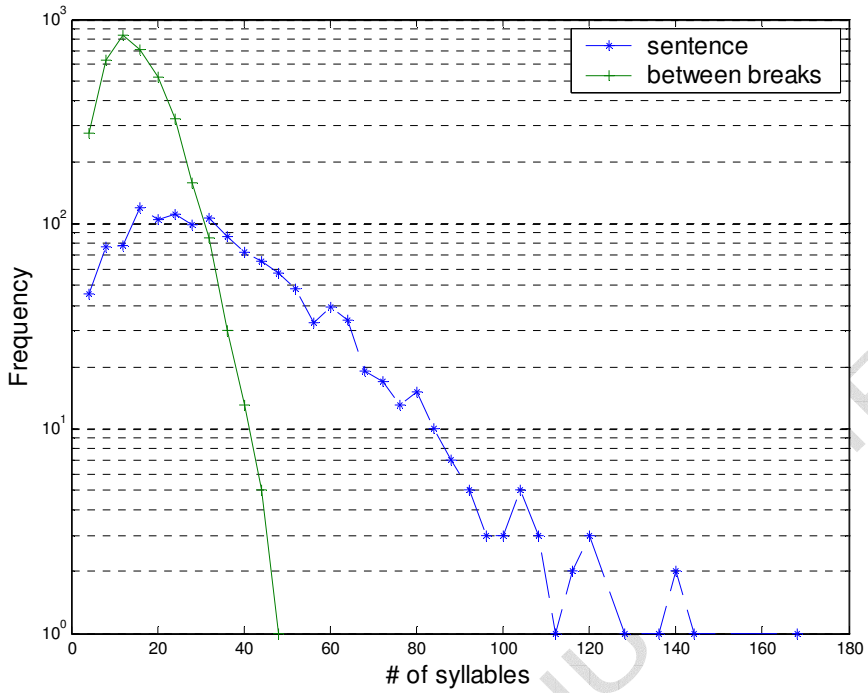


Fig 1. Histogram of the number of syllables between breaks compared with the histogram of the number of syllables in the sentence.

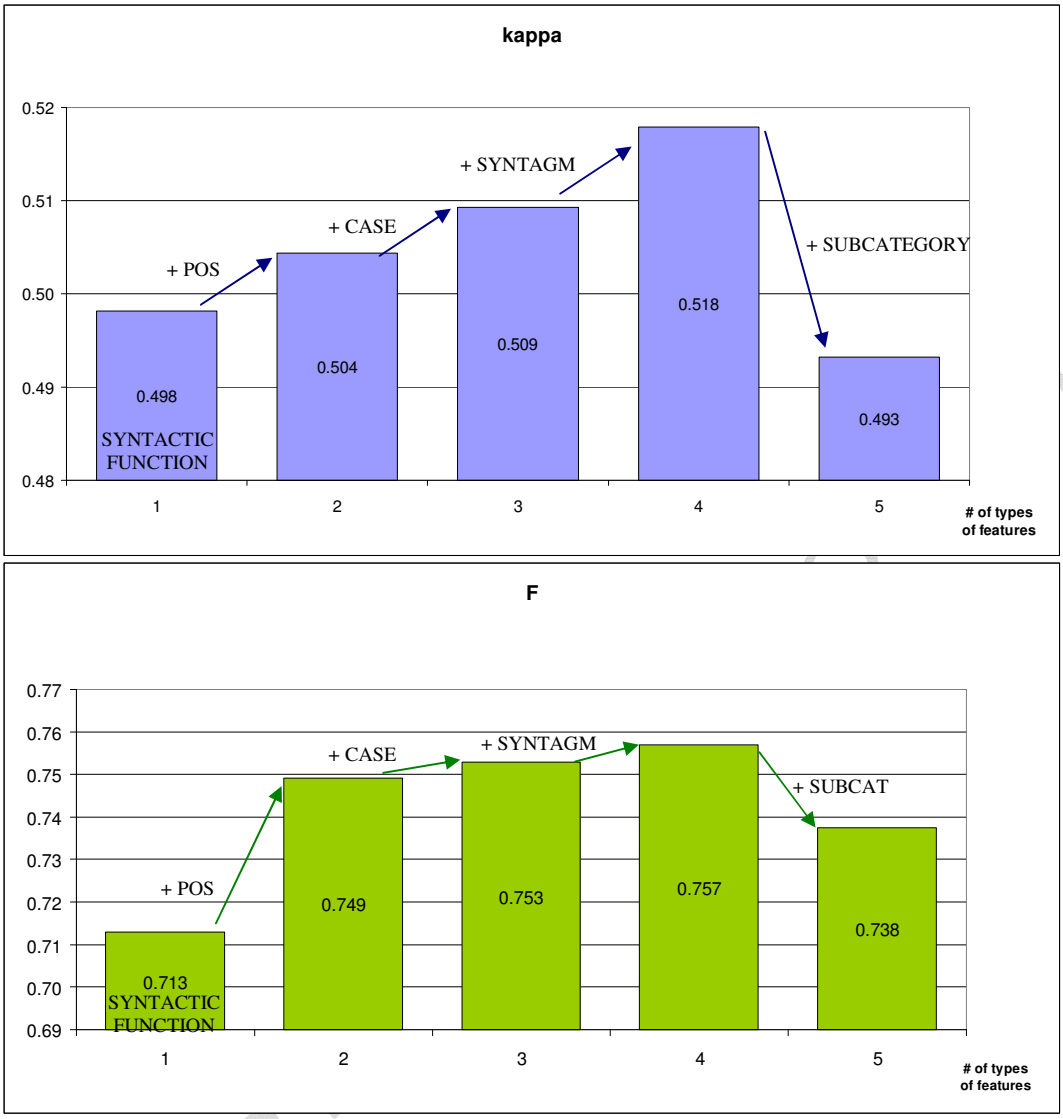


Fig 2. Best κ and F results of the trees built with different number of types of information.

Table 1
Characteristics of the corpus

Feature	Value
# of words	15 867
# of breaks	3 589
# of sentences	1 470
# of intra-sentence breaks	2 119
% of function words	17.5%
% of breaks	20.7%

Table 2
Category or Part of Speech (POS) tags

Label	Description	Example		%followed by break
ADB	Adverb	geroago	(later)	27.8%
ADI	Main verb	eman	(give)	17.0%
ADJ	Adjective	txiki	(little)	18.5%
ADL	Auxiliary verb	du	(has)	49.4%
ADT	Synthetic verb	dago	(is)	44.5%
BST	Others	ohi	(use to)	0.0%
DET	Determiner	hau	(this)	16.5%
ERL	Relation suffixes	badirudi	(it seems)	54.8%
IOR	Pronoun	gu	(we)	3.7%
ITJ	Interjection	kaixo	(hi)	66.7%
IZE	Noun	ordu	(hour)	18.6%
LAB	Abbreviation	etab	(etc)	100.0%
LOT	Conjunction	eta	(and)	18.0%
PRT	Particle	omen	(it seems)	11.4%
PUNT	Punctuation mark	.	(.)	99.2%
SIG	Acronym	AEB	(USA)	8.3%

Table 3
Subcategories considered

Label	Description	Example		%followed by break
ARR	Common noun	liburu	(book)	19.8%
GAL	Question pronoun/adjective	noiz	(when)	9.8%
ADP	Periphrastic verb	behar	(have to)	6.3%
DZG	Indeterminate quantifier	zenbait	(some)	11.4%
DZH	Cardinal	hiru	(three)	13.6%
ELK	Reciprocal pronoun	elkar	(one to another)	0.0%
FAK	Factitive verb	arazi	(make somebody do something)	0.0%
IGB	Indeterminate pronoun	norbait	(someone)	0.0%
IZB	Person proper noun	Jon	(John)	16.5%
IZO	Qualifying adjective	berria	(new)	21.0%
JNT	Coordinate	eta	(and)	2.4%
LIB	Place proper noun	Bilbao	(Bilbao)	20.3%
LOK	Connective	hala ere	(however)	71.7%
ORD	Ordinal	laugarren	(fourth)	6.9%
ORO	General numeral	guzti	(all)	26.8%
SIN	Simple verb	ekarri	(take)	17.1%
ZIU	Security adverb	ote	(maybe)	0.0%

Table 4
Grammatical cases considered

Label	Description	Example		%followed by break
ABL	Ablative case	etxetik	(from the house)	23.2%
ABS	Absolutive case	etxe	(house)	26.8%
ABZ	Orientative case	etxerantz	(towards the house)	16.7%
ABU	Terminative case	etxeraino	(as far as the house)	50.0%
ALA	Allative case	etxera	(to the house)	26.2%
DAT	Dative case	amari	(to the mother)	24.7%
DESK	Descriptive case	minutuko	(of minutes)	0.0%
DES	Benefactive case	amarentzat	(for the mother)	71.4%
ERG	Ergative case	amak	(the mother)	34.2%
GEL	Genitive case for places	etxeko	(of the house)	2.4%
GEN	Genitive case for people	amaren	(of the mother)	0.6%
INE	Inessive case	etxean	(in the house)	40.1%
INS	Instrumental case	kotxez	(by car)	38.1%
MOT	Causal case	etxeagatik	(because of the house)	52.0%
PAR	Partitive case	etxerik	(house)	19.0%
PRO	Prolative case	helburutzat	(as a goal)	20.0%
SOZ	Sociative case	amarekin	(with the mother)	27.0%

Table 5
Syntactic labels considered

LABEL	Description	Example		%followed by break
ADLG	Verbal complement	bakarrik	(only)	23.8%
JADLAG	Auxiliary verb	dute	(have)	25.5%
JADLAG_MP_ADLG	Auxiliary verb as verbal complement	dezagun	(let's have)	50.0%
JADLAG_MP_OBJ	Auxiliary verb as subordinate object	dezaten	(they made)	66.7%
JADNAG	Main verb	eman	(give)	9.3%
JADNAG_MP	Subordinate verb	begiratuta	(watched)	29.1%
JADNAG_MP_ADLG	Subordinate main verb as verbal complement	zerrendaturik	(listed)	8.7%
JADNAG_MP_KM	Subordinate main verb as noun	esate	(to say)	0.0%
JADNAG_MP_IZLG	Subordinate main verb as adjectival complement	erabaki	(decide)	2.5%
JADNAG_MP_OBJ	Subordinate main verb as object	dituen	(that has)	21.2%
JADNAG_IZLG	Main verb as adjectival complement	dutenek	(who have)	10.9%
LOK	Coordinate sentence	ere	(also)	59.2%
MP	Subordinate sentence	arren	(however)	100.0%
IZLG	Adjectival complement	Galiziako	(of Galicia)	4.8%
OBJ	Object	margolana	(painting)	14.4%
PJ	Coordinating conjunction	eta	(and)	1.2%
SUBJ	Subject	erretzaileak	(the smoker)	25.9%
ZOBJ	Indirect object	buruari	(to the head)	10.8%

Table 6
Labels about the grouping of words in syntagms

LABEL	Description	%followed by break
BEG	Beginning of the group	1.2%
CEN	Inside a group	3.9%
END	End of the group	24.5%
UNI	Group of only one word	18.5%

Table 7
Sakoda's contingency coefficient between the feature indicated in the row title and the feature indicated in the column title

FEATURE	C^* for BREAK	C^* for SUBCAT	C^* for CASE	C^* for SYNT	C^* for SYNTAG
POS	0.362	0.613	0.424	0.610	0.475
SUBCATEGORY	0.350	-	0.423	0.600	0.445
CASE	0.140	-	-	0.483	0.374
SYNTACTIC FUNCTION	0.368	-	-	-	0.564
SYNTAGM	0.277	-	-	-	-

Table 8
Results for the trees using only one type of information

TYPE OF INFORMATION	S	κ	P	R	F
POS	89.2%	0.429	0.948	0.454	0.614
SUBCATEGORY	89.8%	0.461	0.872	0.540	0.667
CASE	85.9%	0.253	0.795	0.341	0.478
SYNTACTIC FUNCTION	90.5%	0.498	0.833	0.623	0.713
SYNTAGM	87.6%	0.346	0.830	0.435	0.571

Table 9
Results for the baseline system and the prediction trees built

SYSTEM	S	κ	P	R	F
BASELINE	89.6%	0.402	1.000	0.402	0.574
T1	90.9%	0.518	0.764	0.750	0.757
T2	90.7%	0.511	0.828	0.644	0.725

Table 10
Result of the subjective evaluation

ALGORITHM	CORRECT BREAKS			OMITTED BREAKS			WRONG BREAKS
	#	>80%	Mean score	#	>80%	Mean score	#
T1	62	34	0.71	69	11	0.37	17
T2	45	27	0.74	86	18	0.42	10
REF. LABELLER	79	39	0.68	52	6	0.30	3