AUTOMATIC MORPHOLOGICAL SEGMENTATION FOR CONTINUOUS SPEECH RECOGNITION OF BASQUE

K. López de Ipiña[1], N.Ezeiza[2], G.Bordel[3]

[1]Sistemen Ingeniaritza eta Automatika Saila Gasteiz. Email: isplopek@vc.ehu.es

[2] IXA Taldea. Email: jipecran@sc.ehu.es

[3] Elektrika eta Elektronika Saila, Bilbo. Email: german@we.lc.ehu.es

University of the Basque Country.

ABSTRACT

The selection of appropriate Lexical Units (LUs) is an important issue in the development of Continuous Speech Recognition (CSR) systems. Word has been used classically as unit in most of them. However, proposals of non-word units have begun to arise. Since the subject of this study is the Basque language, which is an agglutinative language with a complex structure inside words, non-word units could be an appropriate choice. In this work an automatic morphological segmentation tool oriented to CSR tasks is presented.

1. INTRODUCTION

The development of a Continuous Speech Recognition (CSR) system for a language involves the selection of a set of suitable Lexical Units (LUs). These LUs are used not only in Language Modelling, but also to define the dictionaries where the acoustic-phonetic models could be integrated. Classically, words have been used as LUs in most of the CSR systems, but for languages whose words are not clearly delimited inside sentences as Japanese, or with words with certain structure within them as Finish, German, Basque etc., the use non-word alternative units seems to be more accurate. There have been several proposals for alternative LUs, such as morphemes (Otsuki et al, 1999), automatically selected non-word units (Peñagarikano et al, 1999), etc.

Since the language of this study is Basque, which is an agglutinative language with a certain structure inside words, the use of non-word units could be an appropriate election.

Basque is a Pre-Indo-European language with an unknown origin and has about 1.000.000 speakers in Basque Country. This language presents a wide dialectal distribution, being 8 the main dialectal variants. This dialectal diversity involves differences at phonetic, phonologic and morphological levels. Moreover, it is relevant the existence of the unified Basque, a standardisation of the language created with the aim of overcoming dialectal differences. Nowadays, a significant amount of speakers and most of mass media uses this standard. Thus, in this work the unified Basque is the main reference.

The following section describes the main morphological features of the language. Section 3 describes the design of rules to create a automatic morphological segmentation tool oriented to CSR. Section 4 details the statistical analysis of morphemes using three different text samples Finally, conclusions are summarised in section 5.

2. MORPHOLOGICAL FEATURES OF BASQUE

In order to deal with the development of a CSR system of Basque, adequate LUs have to be chosen taking into account some features of the language (Alegria, 1996):

1) It is an agglutinative language; the determiner, the number and the declension case are appended to the last element of the phrase and always in this order (deep morphological structure).

2) Basque has a unique declension system, with 15 cases, their morphemes are always added to other elements.

3) Prepositional functions are realised by case suffixes inside word-forms. Thus, Basque presents a relatively high power to generate inflected word-forms.

4) In Basque morphology becomes in fact morphosyntax. For instance, the case morpheme adds syntactic information inside the word-form.

5) Word-formation is very productive in Basque and is very usual to create new compound words as well as derivatives words.

6) The grammatical gender does not exist in Basque; there are not masculine and feminine. However, the verb system uses the difference sometimes, depending on the receiver and the grade of familiarity.

7) Verb forms are composed of a main verb and an auxiliary finite form and it is often found in a single finite verb form, morphemes corresponding to ergative, nominative and dative cases.

In next sentence some of these features can be observed:

Etxekoak jolas-tokira joango dira

Etxe+ko+a+k jolas+toki+ra joan+go dira

House+of+the+ones play+place+to go+will go-they

Those of the house will go to the place to play

3. AUTOMATIC MORPHOLOGICAL SEGMENTATION TOOL ORIENTED TO CSR PURPOSES.

For Basque, an appropriate unit for LUs could be the morpheme, which is the minimal grammatical unit of a language and cannot be divided into smaller grammatical parts.

In order to obtain an automatic morphological segmentation oriented to the CSR purposes, it was first used the automatic morphological segmentation tool of MORFEUS (Alegria, 1996), a robust and wide coverage morphosyntactic analyser for Basque.

MORFEUS divides every word into its constituent morphemes and assigns each morpheme all the morphological features. This process is performed in an incremental way (Fig. 1):

1) The standard analyser processes words according to the standard lexicon and rules of the language.

2) The analyser of linguistic variants analyses dialectal variants and competence errors. This module is very useful since Basque is still undergoing a normalisation process.

3) The analyser of unknown words or *guesser* processes the remaining words.

Nevertheless, the "linguistic" segmentation given by this system presents some problems when the result have to be used in a CSR task, since the result does not correspond exactly to the message of the speech signal.

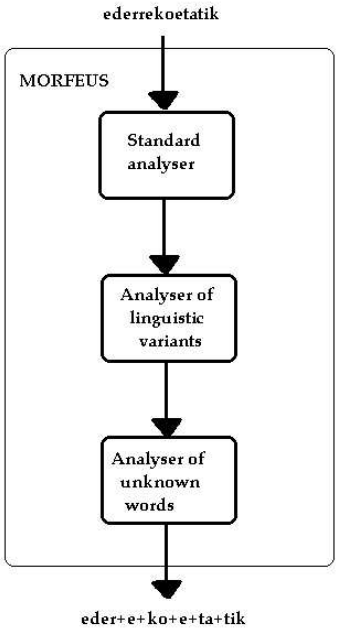

Figure 1. *Diagram of the automatic morphological segmentation tool of MORFEUS.*

In order to solve the problem pointed, an altered set of rules has been designed for the automatic morphological segmentation tool (Tab. 1). The use of these rules provides appropriate word segmentation for CSR purposes.

Thus, the obtained units are not "linguistic" morphemes but pseudo-morphemes oriented to CSR. These pseudo-morphemes could be a first approach of LUs for Basque.

Word	Segmentation of MORFEUS	Rule	Segmentation oriented to CSR
Osaba	Osaba+a	A+a=a	osab+a
Osabei	Osaba+ei	A+e=e	osab+ei
Ulertzeko	Ulertu+tze+ko	tu+tze=tze	Uler+tze+ko
Ulertzen	Ulertu+tzen	tu+tzen=tzen	Uler+tzen
Ulertzea	Ulertu+zea	tu+tzea=tzea	Uler+tze+a
Esertzeko	Eseri+tze+ko	i+tze=tze	Eser+tze+ko
Esertzen	Eseri+tzen	i+tzen=tzen	Eser+tzen
Esertzea	Eseri+tzea	i+tzea=tzea	Eser+tzea
Horretara	Hori+ta+ra	Hori+ta=horr+e+ta	Horr+e+ta+ra
Honetara	Hau+ta+ra	Hau+ta=hon+e+ta	Hon+e+ta+ra
Horri	Hori+i	Hori+i=horr+i	Horr+i
Honi	Hau+i	Hau+i=hon+i	Hon+i
ulertarazi	Ulertu+arazi	tu+arazi=tarazi	Ulert+arazi
Ulergarri	Ulertu+garri	tu+garri=garri	Uler+garri

Table 1. *Rules for the automatic morphological segmentation tool oriented to CSR purposes.*

	||WORD||	||MORPH||	|WORD|	|MORPH|
STBASQUE	197,589	346,232	50,121	20,117
NEWSPAPER	166,972	304,767	38,696	15,302
BCNEWS	180,221	324,126	41,085	17,983

Table 2. *Statistics of LUs for the three text samples.*

4. STATISTICAL ANALYSIS OF MORPHEMES

Taking into account the features of the language a statistical analysis of morphemes of Basque has to be carried out to select appropriate LUs (Lopez de Ipina et al, 2000).

Three text samples of different kinds of language have been randomly created to carry out the statistical analysis. All of them are written in unified Basque:

1) **STBASQUE**: a sample of generic texts and narrative in standard Basque extracted from the database created by the Basque institute of lexicography UZEI (www.uzei.com). The sample has 197,589 words, being its vocabulary of 50,121 words (Tab. 2).

2) **NEWSPAPER**: a sample of general news, sports and financial information from the Basque newspaper EGUNKARIA (www.egunkaria.com). This sample consists of 166,972 words being its vocabulary of 38,696 words (Tab. 2). Usually the newspaper text has been used in CSR because it is available in large amounts and present rich variability of subjects (Otsuki et al, 1999).

3) **BCNEWS**: a sample of general news, sports and financial information from the public Basque radio and television EITB (www.eitb.com). This sample consists of 39,221 and has a vocabulary of 41,085 words (Tab. 2) and includes texts in various styles (planned, spontaneous, interviews, etc.). The use of Broadcast news is extended in CSR because provides speech from different environments (studio, relay, telephone, etc.). There is also a large variety of speakers and speaking styles (planned, spontaneous, interviews, etc.). It also contains spontaneous speech phenomena (Garofolo et al, 1997; Gauvain et al, 1997; Otsuki et al, 1999).

All texts were corrected and pre-processed before the morphological analysis to reduce the morphological analysis error:

1. We discarded marks that are not normally pronounced in spoken communications such as: ?, !, >, ", etc.
2. Words and sentences from another languages were replaced by appropriated terms in Basque language.
3. Comments of the news writers and reporters related to the news and video inputs were removed.
4. In some cases numbers and dates were orthographically transcribed.

Finally, all texts were processing by the spelling Checker/corrector for Basque XUXEN.

The three text samples were automatically morphological segmented by previously designed automatic morphological segmentation tool.

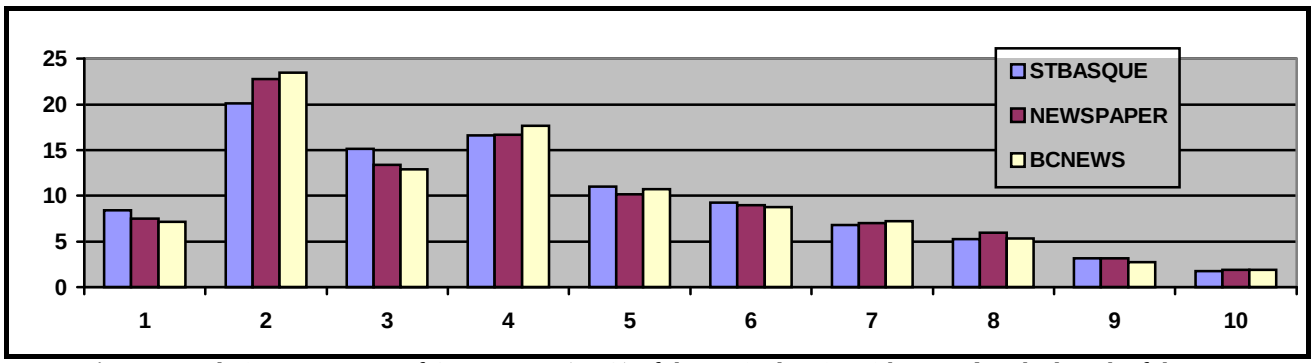

Figure 2. *Relative Frequency of Occurrence (RFO) of the Lexical Units with regard to the length of the LU.*

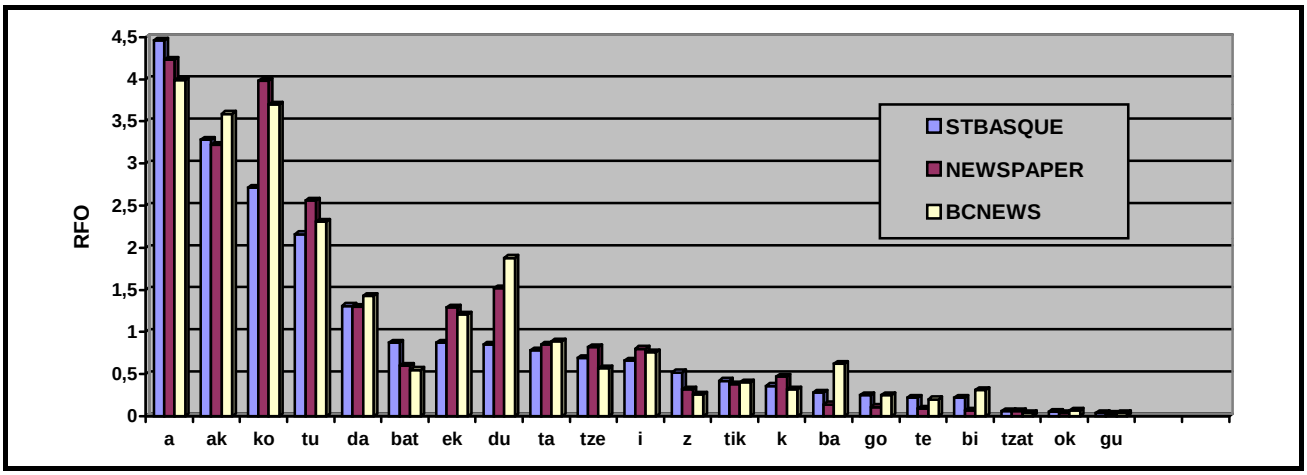

Figure 3. *Relative Frequency of Occurrence (RFO) of short and acoustically similar pseudo-morphemes*

Evaluating the obtained sets of pseudo-morphemes, several main statistic features oriented to CSR purposes can be extracted:

- The vocabulary size of pseudo-morphemes is reduced about 60% in all cases with regard to the vocabulary size of words (Tab. 2).

- About 10% of the vocabulary of pseudo-morphemes (Fig. 2) have less than 4 characters and have an Accumulated Frequency (AF) of 40%, being AF the sum of the Relative Frequency of Occurrence (RFO) of the pseudo-morphemes.

- A significant amount of acoustically similar pseudo-morphemes (Fig. 3) appear in the set of units. These morphemes have very short length and in some cases plosives at unit-boundaries. Moreover their AF is 25%.

The previously analysed morphological features of the language make a priori difficult the selection of appropriate LUs for CSR. Furthermore, evaluating the statistical measures of morphemes (Fig. 2 and 3) it can be observed that the performance of a CSR system could potentially be worse due to several factors:

- On the one hand, acoustically similar morphemes could lead to increase the acoustic confusion.

- On the other hand, the amount of short units could also increase the amount of insertions.

5. CONCLUDING REMARKS.

In this work we have developed an automatic morphological segmentation tool for Basque oriented to Continuous Speech recognition tasks. In a first approach the automatic morphological segmentation tool of MORFEUS was used. Then an altered set of rules was designed in order to adequate the provided segmentation to an appropriated segmentation oriented to CSR. Three text samples of different kinds of language were created to analyze the morphological structure of the language. Finally several statistic morphological analysis were carried out in order to extract features of the language oriented to the selection of appropriated LUs for CSR.

ACKNOWLEDGEMENTS

The authors would like to thank all people have collaborated in the development of this work. We thank specially to UZEI, the newspaper Euskaldunon Egunkaria and the public Basque radio and television EITB for providing us the textual databases.

REFERENCES

Alegria I. et al (1996). *"Automatic morphological analysis of Basque"*, Literary & Linguistic Computing Vol,11, No, 4, pp. 193-203, Oxford University Press.

Garofolo, J.S., Fiscus, J.G., Fisher, W.M., (1997). *"Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora"* In: Proceeding DARPA Speech Recognition Workshop, pp. 15-21.

Gauvain, J.L. et al (1997*). "Transcription of broadcast news"*, Proceedings of Eurospeech, pp. 907-910.

Lopez de Ipiña et al (2000) Lopez de Ipiña K. et al. *"First Selection of Lexical Units for Continuos Speech Recognition of Basque"*, Proceedings of ICSLP. Beijing 2000, vol II, pp. 531-535.

Otsuki K. et al (1999). *"Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcast news"*, Speech Communication. Vol 28, pp 155-166.

Peñagarikano M. et al (1999). *"Using non-word Lexical Units in Automatic Speech Understanding"*, Proceedings of IEEE, ICASSP99, Phoenix, Arizona.