ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

FIRST APPROACH TO THE SELECTION OF LEXICAL UNITS FOR CONTINUOUS SPEECH RECOGNITION OF BASQUE.

M.K. López de Ipiña[1], I.Torres[2], L.Oñederra[3], A. Varona[2],
N. Ezeiza[2], M.Peñagarikano[2], M.Hernandez[4], L.J.Rodriguez[2].

[1]Sistemen Ingeniaritza eta Automatika Saila Gasteiz.
[2] Elektrika eta Elektronika Saila. Bilbo.[3] Euskal Filologi Saila. Gasteiz .
[4] Konputazio Zientziak eta Adimen Artifiziala. Donostia

University of the Basque Country. Spain.

email: karmele@we.lc.ehu.es

ABSTRACT[1]

The selection of appropriated Lexical Units is an important issue in the Language Model (LM) generation. Word has been used classically as unit in most of the Continuous Speech Recognition systems. However, during the last years proposals of non-word units have begun to appear. Since Basque is an agglutinative language with a certain structure inside the word, the non-word units could be an adequate option. In this work, a statistical analysis of the morphological structure of Basque has been carried out. This analysis shows a slight increment of the rates of confusion in Continuous Speech Recognition Systems due to the great increment of acoustically similar and short units. Finally several proposals of Lexical Units are analyzed to deal with the problem.

1. INTRODUCTION

In this paper a first approach to the selection of Lexical Units (LU) for Continuous Speech Recognition (CSR) of Basque is presented.

Basque is a Pre-Indo-European language with 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 [1]. This dialectal diversity involves differences at phonetic, phonological and morphological levels. Moreover, it is relevant the existence of the unified Basque, the so called *Batua*, an standardisation of the language created with the aim of overcoming dialectal differences. This standard has nowadays a great importance in the Basque community, being used by the public institutions and most of massmedia. The *Batua* is used also by people who have studied Basque as a second language the so-called *euskaldunberris*. In this work the standard language *Batua* has been used as a reference.

The development of a CSR system for a language involves not only the generation of a Language Model (LM) but also the selection of a set of suitable lexical units. Classically word has been used as lexical unit in most of the CSR systems.

However during the last years, different proposals of non-word units have begun to appear as alternative. In fact, for languages whose words are not clearly delimited in the sentences as Japanese [2], or with words with a certain structure within them as Finnish, German, Basque etc., this alternative units could be adequate. There are several proposal to deal with the problem, such as morphemes [2], automatic selected non-word units [3], etc. Taking into account the morphological structure of Basque, our first approach was based on morphemes.

Next section describes the main morphological features of the language. Section 3 describes the statistical analysis of the morphemes in Basque. In Section 4, different proposals of LUs are analyzed. The design of two different tasks is presented in section 5. Both tasks are evaluated using different lexical proposals in section 6. Finally, conclusions are summarized in section 7.

2. MORPHOLOGICAL STRUCTURE OF THE LANGUAGE

In order to deal with Basque Language Modeling, adequate LUs have to be chosen, taking into account some features of the language [4]:

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.
2. Basque has an unique declension system, with 15 cases; their morphemes are always added to other elements.
3. Prepositional functions are realized by case suffixes inside word-forms. Thus, Basque presents a relatively high power to generate inflected word-forms.
4. In Basque more than about morphology it can be spoken about morphosyntax. For instance, the case morpheme adds syntactic information inside the word-form.
5. Word-formation is very productive in Basque and it is very usual to create new compound words as well as derivative words.

[1] This work has been partially supported by Spanish CICYT under grant TIC98-0423-C06-03 and GV/PI98/111

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

The ones of the house will go to the playing place.

3. STATISTICAL ANALYSIS OF MORPHEMES

Bearing in mind these features, a brief statistical analysis of the morphemes in Basque was carried out, in order to select appropriated LUs,. For this purpose, it was used the automatic morphological analyzer MORFEUS [4], a robust and wide coverage analyzer for Basque. MORFEUS, in a first stage, divides every word into its constituent morphemes and assigns each morpheme all the morphological features. This process is performed in an incremental way:

1. The standard analyzer processes words according to the standard lexicon and standard rules of the language.
2. The analyzer of linguistic variants analyses dialectal variants and competence errors. This module is very useful since Basque is still in normalization process.
3. The analyzer of unknown words or guesser processes the remaining words.

Table 1. Relative Frequency of Occurrence (RFO) and Accumulated Frequency (AF) of Lexical Units, with regard to the length of them.

LU- length	RFO	AF	Count
1	22.90	22.90	10777
2	19.99	42.89	9405
3	17.26	60.15	8120
4	18.56	78.72	8735
5	11.02	89.73	5184
6	5.49	95.22	2582
7	2.53	97.75	1191
8	1.40	99.15	657
9	0.48	99.63	227
10	0.24	99.87	114
11	0.10	99.97	47
12	0.01	99.99	7
13	0.01	100.00	3

Thus for *ederrekoetatik* (from the ones of the pretty one) it is obtained:

Eder+e+ko+e+ta+tik *Pretty of the (pl.) from.*

A reference text sample of about 5000 sentences (30.000 words and 40.000 morphemes) has been used to carry out the statistical analysis. The following features are observed:

- The number of LUs is reduced from 3500 different words to 1917 different morphemes.

- Evaluating the obtained set of morphemes, several statistic features can be extracted:

· 15% of the different LUs have less than 3 characters and have an Accumulated Frequency (AF) of 42% (table 1) being AF the sum of the Relative Frequency of Occurrence (RFO) of the morphemes.

· A big number of acoustically very similar LUs (table 2) appear in the set of units. These morphemes have very short length and in some cases show plosives in the unit-boundaries. Moreover, their AF is 40% and they represent about the 2% of the different LUs.

Table 2. Relative Frequency of Occurrence (RFO) and Accumulated Frequency (AF) of acoustically similar Lexical Units.

LU	RFO	AF	Count
a	10.95	10.95	5153
k	4.71	15.67	2218
n	2.97	18.64	1399
du	2.83	21.47	1333
da	2.78	24.25	1309
i	1.97	26.22	926
e	1.82	28.05	858
ko	1.82	29.87	857
ek	1.65	32.90	778
tu	1.39	34.29	655
tik	0.47	34.76	219
ri	0.40	35.16	188
te	0.30	35.46	143
rik	0.30	35.76	139
z	0.25	36.01	118
tze	0.22	36.23	102
gu	0.21	36.44	98
ba	0.20	36.64	95
dik	0.19	36.83	89
ga	0.18	37.01	86
bi	0.14	37.15	67
ta	0.11	37.26	54
ok	0.09	37.37	42
ik	0.07	37.44	35
u	0.06	37.50	29
o	0.03	37.53	15

4. PROPOSALS OF LEXICAL UNITS

The previously analyzed morphological features of the language make difficult the selection of appropriated lexical units for CSR.

Furthermore, when the statistical measures of morphemes (table 2) are evaluated, it can be observed that the performance of a CSR system could worsen due to several factors related to the morphological structure of the language:

- Acoustically very similar morphemes could lead to an increment of the acoustic confusion.
- A great number of short units could lead to an increment of the number of insertions.

Due to these difficulties three sets of LUs are proposed:

1. WORD: based on words. Because of the described features, the vocabulary could became intractable when the task has medium-long size.
2. MORPH: based on morphemes. It reduces in about 45% the size of the vocabulary but it could lead to a worse performance of the system.
3. N-WORD: an alternative solution based on morphemes and acoustic criterion: the acoustically very similar or/and very short morphemes are not segmented. As a first approach, this work was made by hand. This proposal reduces in about 20% the size of the vocabulary regarding the first proposal, so the problems pointed out above in the second approach could be solved.

5. DESIGN OF TASKS

The Language Modeling requires the design of appropriated tasks with controlled vocabulary to test Language Models and/or LUs. In order to carry out a wide experimentation for Basque two tasks have been designed.

THE MINIATURE LANGUAGE ACQUISITION.

Miniature Language Acquisition (MLA) is a well defined and small task. This task is used by a computer system to give examples of pictures paired with true statements about those pictures. The task in Basque (Table 3) has 15.000 sentences with about 150.000 words and the vocabulary size is 47 words. These can be simple lemmas, declined words or lemma+morpheme particles:

Triangelu batek eta zirkulu batek karratu baten oso ezkerrera dagoen zirkulu txikia ikutzen dute.

Triangelu bat+ek eta zirkulu bat+ek karratu bat+en
Triangle one and circle one square one

oso ezkerre+ra dago+en zirkulu txiki+a iku+tzen dute.
far left to is which circle little the touch (pres) they
One triangle and one square touch the little circle which is far left of one square.

BASIC VOCABULARY OF BASQUE TASK

Basic Vocabulary of Basque (BVB) is a task based on a language for first level of Basque. The task consists on 5,000 sentences with about 30,000 words being the vocabulary size of 3,500 (table 3). In this task are presented most of the features of the language described in section 2.

Mikel etxera joango da anaiarekin.

Mikel etxe+ra joan+go da anai+a+re+kin.
Mikel house to go+will (he) brother+the+with

Mikel will go to the house with his brother.

Table 3. Statistics of the MLA and BVB tasks.

	SENTENCES.	WORDS	VOCABUL.
MLA	15,000	150,000	49
BVB	4,500	5,000	3,500

6. EXPERIMENTS

Several experiments have been carried out over both tasks in order to validate the three proposals of LUs.

- On the one hand, MLA task reduces the vocabulary to 35 units with MORPH (table 4) and to 40 units with N-WORDS (table 4). To learn the Language Model the set of 15,000 sentences was divided in 14,500 sentences for training and 500 sentences for test.
- On the other hand, BVB task reduces the vocabulary to 1,900 units with MORPH (table 4) and to 2,500 units with N-WORDS (table 4). To learn the Language Model the set of 4,500 sentences was divided in 4,500 sentences for training and 500 sentences for test.

Table 4. Vocabulary of MLA and BVB for WORD, N-WORD and MORPH proposals.

	WORD	MORPH	N-WORD
MLA	47	40	35
BVB	3,500	1,900	2,500

For both tasks Language Modeling was carried out using k-Testable in the Strict Sense LMs [5]. For several values of K, perplexity measures and number of states have been evaluated.

Figure 1 shows the obtained perplexity results and the number of states for MLA task. Clear decrements in perplexity values were obtained when morphemes were considered instead of words. This figure shows a wide gap of perplexity between both proposals (WORD and MORPH). When Non-word units were used, perplexity value situated within the previous observed gap and less acoustic confusion among the units, was observed.

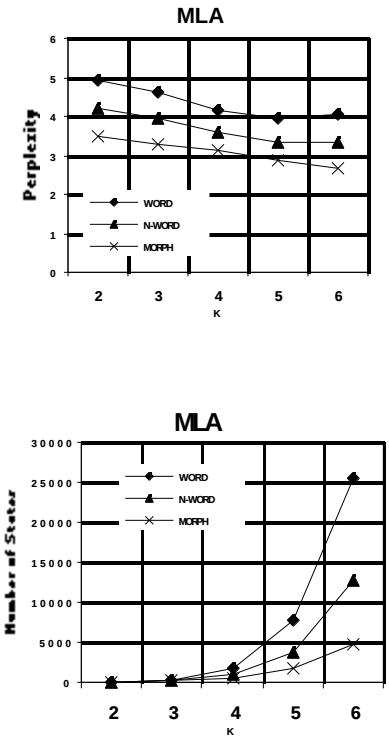

Figure 1. Measures of perplexity and number of states of the k-TSS LM with the three proposals of LUs (WORD, MORPH, N-WORD) for the MLA task.

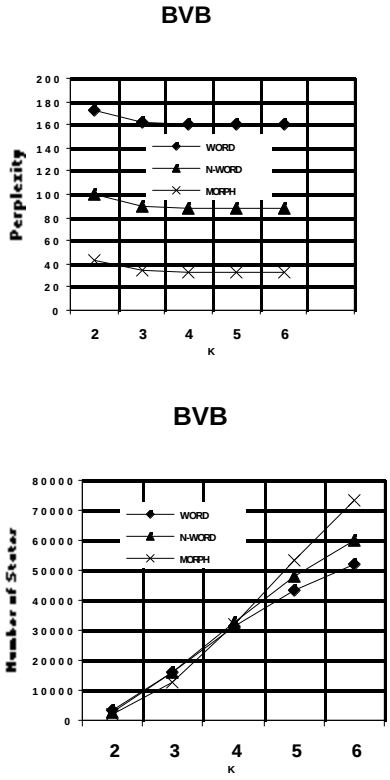

Figure 2. Measures of perplexity and number of states of the k-TSS LM with the three proposals of LUs (WORD, MORPH, N-WORD) for the BVB task.

Figure 2 shows the obtained perplexity results and the number of states for BVB task. The obtained perplexity is clearly bigger than in MLA (Figure 1). Moreover, a wider gap is observed between the two first proposals. Non-word units are also a valid proposal with a perplexity situated within this gap and less acoustic confusion. The bigger gap is observed the more possibilities in the election of suitable sets of LUs we have. The number of states in the LM increased with K in both tasks but only in BVB task there were an increment when Non-WORD and MORPH were considered. That is because the number of seen events in this task increases notably when shorter LU are evaluated.

7 CONCLUDING REMARKS

The selection of appropriated Lexical Units is an important issue in the Language Model (LM) generation. Since Basque is an agglutinative language with a certain structure into the word, the non-word units could be an adequate election. In this work several proposals of Lexical Units are analyzed to deal with the problem. Results show the importance of defining suitable LUs for Language Modeling of the Basque In future works an automatic classification of the LUs will carry out taking into account the acoustic confusion. Another solution based in the automatic generation of LUs will also be evaluated.

8 ACKNOWLEDGEMENTS

The author would like to thank to all people has collaborated in this work in the design of the task and with helpful suggestion about the development.

9 REFERENCES

[1]Mitxelena K., 1977, *"La lengua vasca"*, Leopoldo Zugaza, Durango 1977.

[2]Katsutoshi O., et al. *"Japanese large-vocabulary continuous-speech recognition using a newspaper corpus and broadcoast news"*. Speech Communication. Vol 28. pp 155-166. 1999.

[3]Peñagarikano M., Bordel G., Varona A., López de Ipiña K. *"Using non-word Lexical Units in Automatic Speech Understanding"*. Proceedings of IEEE. ICASSP99, Phoenix, Arizona.

[4]Alegria I., Artola X., Sarasola K., Urkia M. *"Automatic morphological analysis of Basque"*, Literary & Linguistic Computing Vol.11, No. 4, 193-203. Oxford University Press. 1996.

[5] Varona, A. and Torres, I. (1999): "*Using Smoothed K-TLSS(S) Language Models in Continuos Speech Recognition*". Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. Vol II, pp. 729-732