Selection of Lexical Units for Continuous Speech Recognition of Basque

K. López de Ipiña[1], M. Graña[2], N. Ezeiza[3], M. Hernández[2], E. Zulueta[1], A. Ezeiza[3], and C. Tovar[1]

[1]Sistemen Ingeniaritza eta Automatika Saila Gasteiz.
{isplopek, iepzugee}@vc.ehu.es

[2] Konputazio Zientziak eta Adimen Artifiziala Saila, Donostia. ccpgrrom@si.ehu.es

[3]IXA group, Donostia. aitzol@si.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. Words have been used classically as the recognition unit in most of them. However, proposals of non-word units are beginning to arise. Basque is an agglutinative language with some structure inside words, for which non-word morpheme like units could be an appropriate choice. In this work a statistical analysis of units obtained after morphological segmentation has been carried out. This analysis shows a potential gain of confusion rates in CSR systems, due to the growth of the set of acoustically similar and short morphemes. Thus, several proposals of Lexical Units are analysed to deal with the problem. Measures of Phonetic Perplexity and Speech Recognition rates have been computed using different sets of units and, based on these measures, a set of alternative non-word units have been selected.

Keywords: Lexical Units, CSR, aglutinative languages.

1 Introduction

This paper presents an approach to the selection of Lexical Units (LUs) for Continuous Speech Recognition (CSR) of Basque. 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 development of a CSR system for a language involves the selection of a set of suitable LUs. These LUs are used not only in Language Modelling, but also to define the dictionaries where the acoustic-phonetic models can be integrated. Classically, words have been used as LUs in most of the CSR systems. However, some recent proposals point out non-word units as alternative LUs for some languages. In fact for

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 244–250, 2003.

languages whose words are not clearly delimited inside sentences such as Japanese [1], or with words with some structure within them such as Finish, German, Basque etc., these alternative units seem to be more accurate. There have been several proposals for alternative LUs, such as morphemes [1], automatically selected non-word units [2], etc. Thus, taking into account the morphological structure of Basque, the use of morphemes seems to be an appropriate approach.

Table 1. Main characteristics of the textual databases

	STBASQUE	**NEWSPAPER**	**BCNEWS**
Text amount	**1,6M**	**1,3M**	**2,5M**
Number of words	197,589	166,972	210,221
Number of pseudo-morphemes	346,232	304,767	372,126
Number of sentences	15,384	13,572	19,230
Vocabulary size in words	50,121	38,696	58,085
Vocabulary size in pseudo-morphemes	20,117	15,302	23,983

The following section describes the main morphological features of the language and details the statistical analysis of morphemes using three different textual samples. Section 3 presents the experiments and the evaluation criteria that have been used. Finally, conclusions are summarised in section 4.

2 Morphological Features of Basque

Basque is an aglutinative language with a special morpho-syntatic structure inside the words [3][4] that may lead to intractable vocabularies of words for a CSR when the size of task is large. A first approach to the problem is to use morphemes instead of words in the system in order to define the system vocabulary [4]. This approach has been evaluated over three textual samples analysing both the coverage and the Out of Vocabulary rate, when we use words and pseudo-morphemes obtained by the automatic morphological segmentation tool AHOZATI [5].Table 1 shows the main features of the three textual samples relating to size, number of words and pseudo-morphemes and vocabulary size, both in words and pseudo-morphemes for each database. The first important outcome of our analysis is that the vocabulary size of pseudo-morphemes is reduced about 60% (Fig. 1) in all cases relative to the vocabulary size of words. Regarding the unit size, Fig. 2 shows the plot of Relative Frequency of Occurrence (RFO) of the pseudo-morphemes and words versus their length in characters over the textual sample STDBASQUE. Although only 10% of the pseudo-morphemes in the vocabulary have less than 4 characters, such small morphemes have an Accumulated Frequency of about 40% in the databases [5] (the Acumulated Frequency is calculated as the sum of the individual pseudo-morphemes RFO).

To check the validity of the unit inventory, units having less than 4 characters and having plosives at their boundaries were selected from the texts. They represent some 25% of the total. This high number of small and acoustically difficult recognition

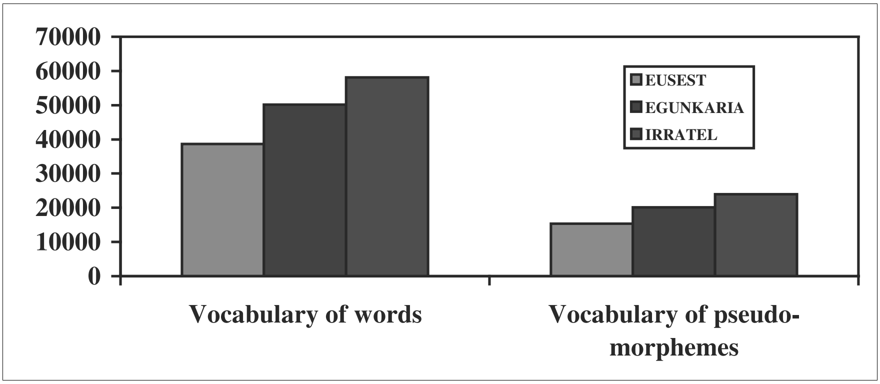


Fig. 1. Vocabulary size of the words and pseudo-morphemes

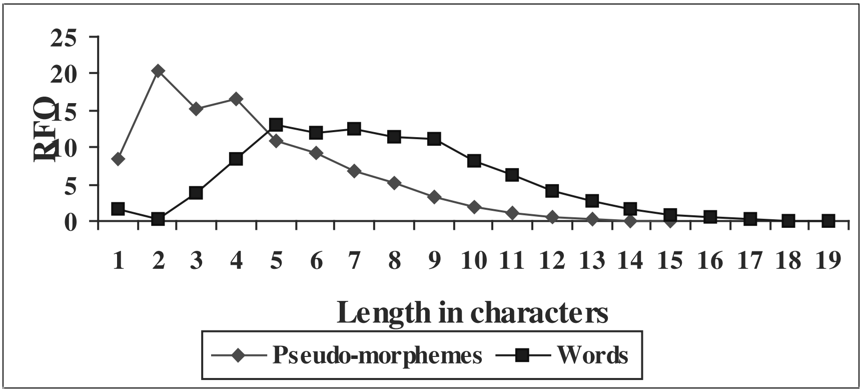


Fig. 2. Relative Frequency of Occurrence (RFO) of the words and pseudo-morphemes in relation to their length in characters (STDBASQUE sample)

units could lead to an increase of the acoustic confusion, and could also generate a high number of insertions (Fig. 3 over the textual sample EGUNKARIA).

Finally, Fig. 4 shows the analysis of coverage and Out of Vocabulary rate over the textual sample BCNEWS. When pseudo-morphemes are used, the coverage in texts is better and complete coverage is easily achieved. OOV rate is higher in this sample.

3 Experimentation

3.1 Description of the Tasks

Appropriate tasks with controlled vocabularies are required to test LM and/or LUs. Two tasks have been created [4] for this purpose:

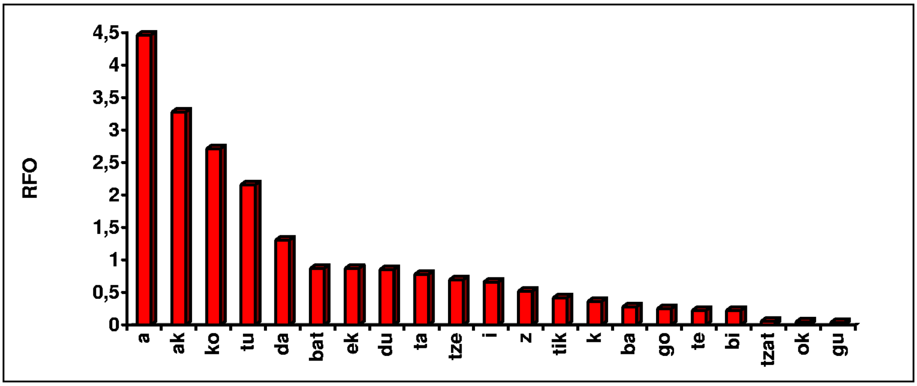
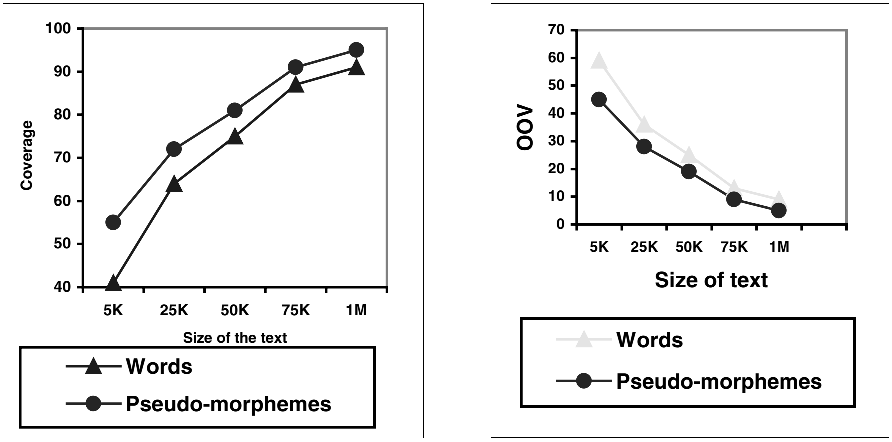


Fig. 3. Relative Frequency of Occurrence (RFO) of small and acoustically difficult recognition units (EGUNKARIA sample)



(a) (b)

Fig. 4. Coverage (a) and OOV rate (b) for the textual sample BCNEWS

a) Miniature Language Acquisition (MLA) task is the language used by a computer system to give examples of pictures paired with true statements about those pictures. The task in Basque has 15,000 sentences with about 150,000 words, being 47 the vocabulary size. It has very low perplexity and very restrictive vocabulary size. It was created for preliminary experiments of CSR.

b) Basic Vocabulary of Basque (BVB) is a task based on beginner's level of Basque. The task consists of 5,000 sentences with about 30,000 words, being 3,500 the vocabulary size. Most of the features of the language described in section 2 are present in this task. It has a high perplexity comparing to MLA task and, it was created to measure the precision of the system when a larger scale task is used.

Both tasks were automatic morphologically segmented into pseudo-morphemes by AHOZATI. The MLA task reduces its vocabulary size to 35 pseudo-morphemes and, BVB task to 1,900. Finally, a segmentation in N-WORDS was obtained resulting in, 40 and 2500 different vocabulary units for MLA and BVB tasks respectively. The sentences of MLA task were divided into 14,500 sentences for training and 500 for test and, the sentences of BVB task into 4,000 for training and 500 for test. 20 speakers, 10 males and 10 females, recorded both tasks, obtaining 400 sentences for MLA and 800 sentences for BVB. In the speech recognition experiments a subset of BVB (MBVB) was used. The subset has a vocabulary size of 550 for WORDS, 400 for PS-MORPHS and 500 for N-WORDS.

3.2 Evaluation Criteria

a) A perplexity function to evaluate the influence of the LUs in the LM. The classical perplexity function used to evaluate LMs might not be valid in this case. This function depends on the units used to compose sentences. Therefore, the evaluation must be based on an invariant unit, such as it is the phoneme. Thus, Phonetic Perplexity will be used to validate LUs. This perplexity is expressed as in [6]:

$$PP = 2^{\left[\frac{-1}{F}\sum_{i=1}^{N}\mathbf{log_2}\,\Pr ob(W_i|M)\right]} = P^{K/F} \quad (1)$$

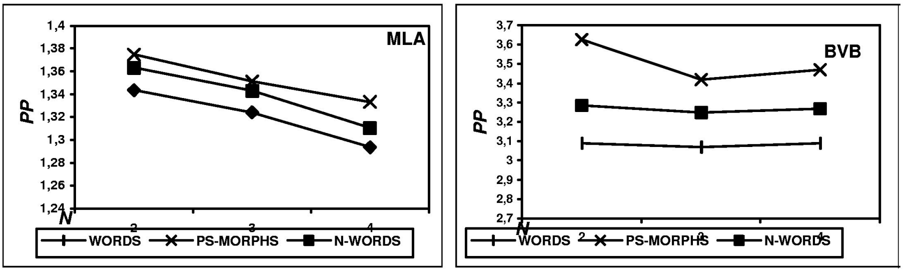
Where PP is the Phonetic Perplexity function, P is the perplexity and F and K are the number of phonemes and units composing the sentences, respectively. The CMU-Cambrige Toolkit [6] has been used to calculate both PP and P for different N-gram lengths.

b) Speech Recognition experiments without LM have been carried out to evaluate both the influence of acoustic confusion of LUs and the insertion of short LUs. Moreover, Recognition Rates for LUs (LURR) have been analysed using the raw stream of LUs (LURR-NA) and also the stream of words after the alignment of the non-words LUs (LURR-A) to words using simple information about the set of words. A set of 28 Contextual Independent Sublexical units modelled by Discrete HMMs with four codebooks will be used as acoustic models.

c) The computational cost of the experiments is also tested. We evaluate the Computational Time (CT) (the performance in msecs. Real time operative corresponding to 10 msecs) and the Time Weighted LURR (T-LURR).

3.3 Preliminary Experiment

The previously analysed morphological features of the language make difficult the selection of appropriate LUs for CSR. Furthermore, evaluating the statistical measures of morphemes, it can be observed that the performance of the Acoustic Phonetic Decoding system could potentially be worse due to several factors. On the one hand, acoustically similar morphemes could lead to increase acoustic confusion. On the other hand, the amount of short units could also increase the amount of insertions [5].



(a) (b)

Fig. 5. Measurements of Phonetic Perplexity for MLA (a) and BVB (b) tasks

Table 2. Recognition rates (LURR) using the three sets of lexical units, WORDS, N-WORDS andPS-MORPHS

	MLA					MBVB				
	LURR-NA	LURR-A	LURR-BIGR	CT	T-LURR	LURR-NA	LURR-A	LURR-BIGR	CT	T-LURR
WORDS	80,61	80,61	91,34	6	13,4	43,71	43,71	48,44	33	1,46
N-WORDS	74,84	76,30	88,82	5	14,96	30,09	32,60	42,07	28	1,50
PS-MORPHS	60,29	63,80	82,38	3	20,09	28,98	29,09	39,85	25	1,59

Three sets of LUs are used in the experiments [4]:

1. **WORDS**: words are our baseline LU set.
2. **PS-MORPHS**: these pseudo-morpheme units are morphemes automatically obtained and slightly transformed for Speech Recognition by ad-hoc rules [5].
3. **N-WORDS:** An alternative proposal. Pseudo-morphemes of length lower than 3 characters with a high level of confusion are merged with adjacent units [5]. This proposal reduces the vocabulary size about 25% with respect to WORDS.

3.4 Experimental Results

Experiments with WORDS and PS-MORPHS sets were carried out to analyse the influence of the morphological structure in the recognition of the LUs. Measures of PP were computed for different values of N. Fig. 5 shows lower PP of WORDS with respect to PS-MORPHS in both tasks. The results of the speech recognition experiments also show better performance for WORDS than for PS-MORPHS in both tasks (table 2) This is due to the frequent confusion and the high amount of insertion in the case of the shortest pseudo-morphemes. Consequently, the alignment improves the results and reduces the insertion of short LUs. However, WORDS still obtained better results than PS-MORPHS. With regard to the CT and T-LURR the advantage is for PS-MORPHS. Regarding BVB task, it can be observed that the overall results are worst than in MLA (table 2), but it must be taken into account that the perplexity of the task is considerably higher [4]. The results show that PS-MORPHS has worst result of recognition but better results with regard to the computational cost.

The experiments using the new LUs N-WORDS show that PP is lower than the one for PS-MORPHS (Fig. 5) and closer to the perplexity measure for WORDS. Table 2 indicates also that N-WORDS outperforms PS-MORPHS for MLA and MBVB tasks with or without alignment. Moreover the recognition rate of N-WORDS is closer to the rate for WORDS in both tasks. N-WORDS shows in table 2 the best balance of LURR and computational cost (CT and T-LURR).

Finally, table 2 shows the performance of the system with a bigram Language Model. The introduction of a Language Model improves all the results, but the increase in performance is more significant for non-word LUs.

4 Concluding Remarks

This work deals with the selection of appropriate LUs for Basque language. Since Basque is an agglutinative language, non-word units could be an adequate choice for LUs. First, morphemes and words have been tested, including a statistical analysis of morphemes in Basque. This analysis shows a large amount of short and acoustically similar morphemes, leading to a bad performance of the CSR system. Measures of phonetic perplexity, computational cost and speech recognition experiments have been completed to validate both proposals. Although word model obtains the best results, it becomes intractable for medium-large dictionaries. Thus, a new set of non-word units has been created based on morphemes. This proposal shows an appropriate performance of the system and reduces the problems raised by morphemes. In future works the obtained sets of LUs will be evaluated in a LVCSR system.

Acknowledgements. The authors would like to thank all the volunteer speakers that has collaborated recording the databases. We thank also all people have collaborated in the development of this work: Euskaldunon Egunkaria, UZEI [7], EITB [8] and the University of the Basque Country (this work has been partially supported by the University of the Basque Country, under project (UPV00147.345-E-14895/2002).

References

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

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

[3] Alegria I. et al. *"Automatic morphological analysis of Basque"*, Literary & Linguistic Computing Vol,11, No, 4, 193–203, Oxford University Press, 1996.

[4] Lopez de Ipiña K. et al. *"First Selection of Lexical Units for Continuous Speech Recognition of Basque"*, Proceedings of ICSLP. Beijing 2000, Vol II, pg. 531–535.

[5] Lopez de Ipina K., N. Ezeiza, G. Bordel. & M. Graña. "Automatic Morphological Segmentation for Speech Processing in Basque" Proceeding of IEEE TTS Workshop. Santa Monica USA. 2002.

[6] P.R. Clarkson and R. Rosenfeld. Statistical Language Modelling Using the CMU-Cambridge Toolkit From Proceedings ESCA Eurospeech 1997.

[7] http://www.uzei.com

[8] http://www.eitb.com