Language Resources for a Bilingual Automatic Index System of Broadcast News in Basque and Spanish

G. Bordel[1], A. Ezeiza[2], K. Lopez de Ipina[3], J.M. López[3], M. Peñagarikano[1], and E. Zulueta[3]

University of the Basque Country,
[1] Elektrizitate eta Elektronika Saila, Leioa
{german, mpenagar}@ehu.es
[2] Ixa taldea. Sistemen Ingeniaritza eta Automatika Saila, Donostia
aitzol.ezeiza@ehu.es
[3] Sistemen Ingeniaritza eta Automatika Saila, Gasteiz
{isplopek, josemanuel.lopez, iepzugue}@ehu.es

Abstract. Automatic Indexing of Broadcast News is a developing research area of great recent interest [1]. This paper describes the development steps for designing an automatic index system of broadcast news for both Basque and Spanish. This application requires of appropriate Language Resources to design all the components of the system. Nowadays, large and well-defined resources can be found in most widely used languages, but there is a lot of work to do with respect to minority languages. Even if Spanish has much more resources than Basque, this work has parallel efforts for both languages. These two languages have been chosen because they are evenly official in the Basque Autonomous Community and they are used in many mass media of the Community including the Basque Public Radio and Television EITB [2].

1 Introduction

Automatic Indexing of Broadcast News is a topic of growing interest for the mass media in order to take maximum output of their recorded resources. Actually, it is a challenging problem from researchers' point of view, due to many unresolved issues like speaker changes and overlapping, different background conditions, large vocabulary, etc. In order to achieve significant results in this area, high-quality language resources are required. Since the main goal of our project is the development of an index system of broadcast news in the Basque Country, our approach is to create resources for all the languages used in the mass media. The analysis of the specific linguistic problematic indicates that both Basque and Spanish are official in the Basque Autonomous Community and they are used in the Basque Public Radio and Television EITB [2] and in most of the mass media of the Basque Country (radios and newspapers). Thus it is clear that both languages have to be taken into account to develop an efficient index system. Therefore, all of the tools (ASR system, NLP system, index system) and resources (digital library, Lexicon) to be developed will be oriented to create a bilingual system in Basque and Spanish.

Spanish has been briefly studied for development of these kind of systems but the use of Basque language (a very odd minority language) introduces a new difficulty to

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 1047 – 1054, 2005.

the development of the system, since it needs specific tools and the resources available are fewer.

Basque is a Pre-Indo-European language of unknown origin and it has about 1.000.000 speakers in the Basque Country. It presents a wide dialectal distribution, including six the main dialects, and this dialectal variety entails phonetic, phonologic, and morphologic differences.

Moreover, since 1968 the Royal Academy of the Basque Language, Euskaltzaindia [3] has been involved in a standardisation process of Basque. At present, morphology, which is very rich in Basque, is completely standardised in the unified standard Basque, but the lexical standardization process is still going on. The standard Basque, called “Batua”, has nowadays a great importance in the Basque community, since the public institutions and most of the mass media use it. Furthermore, people who have studied Basque as a second language use "Batua" as well.

Hence, we have made use of the standard version of Basque as well as the standard Spanish in the development of the resources presented in this work.

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 resources developed. Section 4 describes the processing of the data. Finally, conclusions are summarised in section 5.

2 Morphological Features of Basque

Basque is an agglutinative language with a special morpho-syntactic structure inside the words [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 [5].

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 [6].

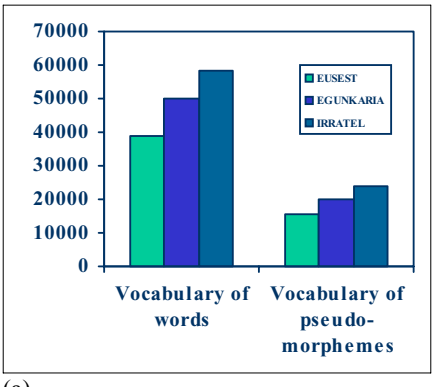
Table 1. Main characteristics of the textual databases for morphologic analysis

	STDBASQUE	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

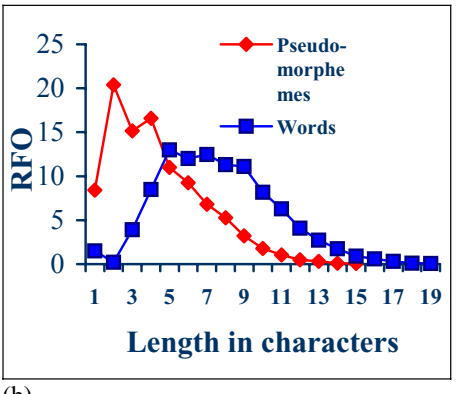
Table 1 shows the main features of the three textual samples relating to size, number of words and pseudo-morphemes and vocabulary size, both in words and pseudo-morphemes for each database [6].

Figure 1 shows some of the interesting conclusions derived of this analysis. The first important outcome of our analysis is that the vocabulary size of pseudo-morphemes is reduced about 60% (Fig. 1, a) in all cases relative to the vocabulary

size of words. Regarding the unit size, Fig. 1 (b) shows the plot of Relative Frequency of Occurrence (RFO) of the pseudo-morphemes and words versus their length in characters over the textual sample named STDBASQUE. Although only 10% of the pseudo-morphemes in the vocabulary have fewer than four characters, such small morphemes have an Accumulated Frequency of about 40% in the databases (the Accumulated Frequency is calculated as the sum of the individual pseudo-morphemes RFO) [7].



(a)



(b)

Fig. 1. (a) Vocabulary size of the words and pseudo-morphemes in the three textual samples and (b) Relative Frequency of Occurrence (RFO) of the words and pseudo-morphemes in relation to their length in characters (STDBASQUE sample)

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 units could lead to an increase of the acoustic confusion, and could also generate a high number of insertions (Fig. 2 over the textual sample EGUNKARIA[8]).

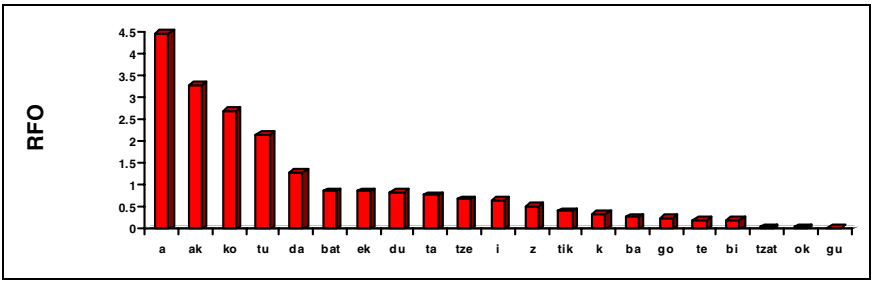


Fig. 2. Relative Frequency of Occurrence (RFO) of small and acoustically difficult recognition units in BCNEWS sample

Finally, Fig. 3 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.

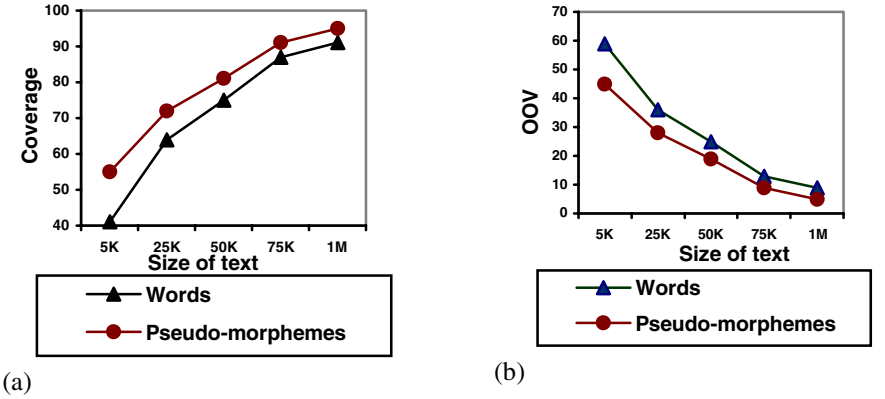


Fig. 3. Coverage (a) and OOV rate (b) for the textual sample BCNEWS

3 Resources Developed

Resources in Spanish

- 6 hours of video in MPEG4 (WMV 9) format of "Teleberri" program, the daily program of broadcast news in Spanish, directly provided by the Basque Public Radio and Television EITB [2].
- 6 hours of audio (WAV format) extracted from the video (MPEG4) files.
- 6 hours of audio transcription in XML format containing information about speaker changes, noises and music fragments, and each word's phonetic and morphologic information.
- 1 year of scripts, in text format, of the "Teleberri" program. The text is divided in sentences and paragraph.
- 1 year of local newspapers in Spanish Gara [9], in text format. The text is divided in sentences and paragraph.
- Lexicon extracted from the XML transcription files, including morphologic, phonologic and orthographic information.

Resources in Basque

- 6 hours of video in MPEG4 (WMV 9) format of "Gaur Egun" program, the daily program of broadcast news in Basque directly provided by the Basque Public Radio and Television EITB [2].
- 6 hours of audio (WAV format) extracted from the video (MPEG4) files.

- 6 hours of audio transcription in XML format containing information about speaker changes, noises and music fragments, and each word's phonetic and orthographic transcription including word's lemma and Part-Of-Speech disambiguated tags.
- 1 year of scripts, in text format, of the "Gaur Egun" program.
- 1 year of local newspapers in Basque (Euskaldunon Egunkaria [8]), in text format.
- Lexicon extracted from the XML transcription files, including phonologic, orthographic, and morphologic information.

4 Processing Methodology

Processing of the Video Data

The video data used in this work has been provided directly by the Basque Public Radio and Television. The format used to store the broadcast contents is MPEG4 (WMV 9), and the Basque Public Radio and Television has been very kind offering us all these resources.

The ASR system developed doesn't actually use the useful graphical information of the videos, but the images have been used thoroughly during transcription in order to find additional information that could enrich the transcription, as names and descriptions of speakers, translation of foreign speakers' words, description tables and maps, etc.

In the near future some specific image information retrieval techniques could be incorporated to the ASR system.

Processing of the Audio Data

The audio data has been extracted out from the MPEG4 video files, using FFmpeg free software[1]. The audio files have been stored in WAV format (16 KHz, linear, 16 bits).

When the audio data was ready, the XML label files were created manually, using the Transcriber free tool [10]. The XML files include information of distinct speakers, noises, and paragraphs of the broadcast news. The transcription files follow the conventions defined in the COST278 project and they contain extra phonetic and orthographic information of each of the words. Some of the recommendations and features described by the Linguistic Data Consortium in [11] have been also included for a better interpretation of the transcription files.

These features include identification of the dialect used by speakers, correct spelling of mispronounced words, language marks for any inclusion of foreign speech in the transcription, and identification numbers for related topics in both Basque and Spanish Broadcast News.

Table 2 shows a simplified sample of the enriched version of the transcription for Basque. Some of the morphological information has been deleted to easier reading of the example.

[1] Available online at http://ffmpeg.sourceforge.net

Table 2. Simplified sample of the output of the Transcriber free tool [10] enriched with morpho-syntactic information of Basque

```
<Sync time="333.439"/>
+horretarako /hortarako/<Word lemma=”hori” POS=”ADB”/>
+denok /danok/<Word lemma=”dena” POS=”IZL”/>
lagundu<Word lemma=”lagundu” POS=”ADI”/>
behar<Word lemma=”behar” POS=”ADI”/>
dugu<Word lemma=”*ukan” POS=”ADL”/>
.
</Turn>
<Turn mode="spontaneous" fidelity="high" start-
Time="335.182" endTime="336.065">
<Sync time="335.182"/>
^Batasunak<Word lemma=”9batasuna” POS=”IZB”/>
```

As Basque is an agglutinative language with very rich inflection variety [4], Basque XML files include morphologic information such as each word's lemma and Part-Of-Speech tag. This information could be very useful in the development of Language Models for the recognition of continuous Speech in this context.

Using this transcribed information, a Lexicon for each language has been extracted. The Lexicon stores information of each different word that appears in the transcription. This information could be very useful for developing speech recognition tools as well as many other NLP applications.

Processing of the Textual Data

There are two independent types of textual resources: The text extracted from the newspapers Gara [9] and Euskaldunon Egunkaria [8]), and the scripts of the "Teleberri" and "Gaur Egun" programs. These last resources are very interesting because they are directly related (date, program) with the texts read in the broadcast news both in Spanish and Basque.

All of them were processed to include morphologic information such as each word's lemma and Part-Of-Speech tag. Using all the information, a Lexicon for each language has been extracted taken into account the context of the word in order to eliminate the ambiguity. The Lexicon stores information of each different word that appears in the transcription, and this information could be very useful for developing speech recognition tools. Table 3 shows some examples of the lexicon information.

The first column of Table 3 shows some example of the words as they have been transcribed from the Broadcast News audio recording. The alternative transcriptions of the word are spotted in second place, and the morphological information is later added, and it includes morpho-syntactic information, lemma information [4] and its corresponding sub-lexical unit segmentation as explained in [6].

Table 3. Sample of the Lexicon for Basque, including information extracted of the morphologic analysis of the transcription

Input	Transcription	Morphological Analysis	LEMA	Morphological segmentation
euskaldunena	ewS.'kal.du.ne.'2na	ADJ IZO DEK GEN MG DEK ABS NUMS MUGM ;	euskaldun	euskaldun=en=a;
		ADJ IZO DEK GEN NUMP MUGM DEK ABS NUMS MUGM ;		euskaldun=en=a;
		ADJ IZO GRA SUP DEK ABS NUMS MUGM		euskaldun=en=a
margolarien	mar.'Go.la.r6i.'2en	IZE ARR DEK GEN NUMP MUGM ;	margolari	margolari=en;
		IZE ARR DEK GEN NUMP MUGM DEK ABS MG		margolari=en
margolaritzan	mar.'Go.la.r6i.'2t&san mar.'Go.la.r6i.'2t&c~an	IZE ARR DEK NUMS MUGM DEK INE	margolaritza	margolaritz=an
margolaritza	mar.'go.la.r6i.'2t&sa	IZE ARR; IZE ARR DEK ABS MG ;	margolaritza	margolaritza;
	mar.'go.la.r6i.'2t&c~a	IZE ARR DEK ABS NUMS MUGM		margolaritza;
				margolaritz=a

5 Concluding Remarks

In this paper a developing system for automatic indexing of bilingual Broadcast News has been presented. Its development entails the compilation of resources for both Basque and Spanish, which are the official languages in the Basque Country, and they are used in the Basque Public Radio and Television EITB [2] and in most of the mass media of the Basque Country.

Resources for Basque have been explained in more detail, since it is a minority language with special problematic. Since it is an agglutinative language, analysis of coverage and words OOV has been carried out in order to develop an appropriate Lexicon.

Finally, we would like to remark that lexicons are enriched using morphologic and phonetic information, not just extracting a word list, so this information could be useful in future development of more sophisticated approaches in ASR systems and transcription of Broadcast News.

Acknowledgements

We would like to thank UZEI for they help extracting information about RFO of phonemes. We thank also all the people and entities that have collaborated in the development of this work: EITB [2], Gara [9] and Euskaldunon Egunkaria [8].

References

1. Vandecatseye, A., J.P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, F.J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, C. Alexandris, 2004. The COST278 pan-European Broadcast News Database. In Proceedings of LREC 2004, Lisbon (Portugal).
2. EITB Basque Public Radio and Television, *http://www.eitb.com/*
3. Euskaltzaindia, *http://www.euskaltzaindia.net/*
4. Alegria I., Artola X., Sarasola K., Urkia M.: "Automatic morphological analysis of Basque", Literary & Linguistic Computing Vol,11, No, 4, 193-203, Oxford Univ Press, 1996.
5. Peñagarikano M., Bordel G., Varona A., Lopez de Ipina: "Using non-word Lexical Units in Automatic Speech Understanding", Proceedings of IEEE, ICASSP99, Phoenix, Arizona.
6. Lopez de Ipiña K., Graña M., Ezeiza N., Hernández M., Zulueta E., Ezeiza A., Tovar C.: "Selection of Lexical Units for Continuous Speech Recognition of Basque", Progress in Pattern Recognition, pp 244-250. Speech and Image Analysis, Springer. Berlin. 2003.
7. Lopez de Ipina K., Ezeiza N., Bordel. N., Graña M.: "Automatic Morphological Segmentation for Speech Processing in Basque" IEEE TTS Workshop. Santa Monica USA. 2002.
8. Egunkaria, Euskaldunon Egunkaria, the only newspaper in Basque, which has been recently replaced by Berria, online at *http://www.berria.info/*
9. GARA, local Basque Country newspaper in Spanish, online at *http://www.gara.net/*
10. Barras C., Geoffrois E., Wu Z., and Liberman M.: “Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech” First International Conference on Language Resources and Evaluation (LREC-1998).
11. Linguistic Data Consortium, Design Specifications for the Transcription of Spoken Language, available online at http://www.ldc.upenn.edu/Projects/Corpus_Cookbook.