1. Introduction

The development of Information Retrieval systems requires appropriated digital resources. Since the main goal of our project is the development of an index system of broadcast news in the Basque Country, it is essential 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 (EITB) 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 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 (Euskaltzaindia) 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 "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 (Alegria et al., 1996) 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 (Peñagarikano et al., 1999).

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 (Lopez de Ipina et al., 2002). 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 (Lopez de Ipina et al., 2003).

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

<table>
<thead>
<tr>
<th></th>
<th>STDBASQUE</th>
<th>NEWSP</th>
<th>BCNEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text amount</td>
<td>1,6M</td>
<td>1,3M</td>
<td>2,5M</td>
</tr>
<tr>
<td>Number of words</td>
<td>197,589</td>
<td>166,972</td>
<td>210,221</td>
</tr>
<tr>
<td>Number of pseudo-morphemes</td>
<td>346,232</td>
<td>304,767</td>
<td>372,126</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>15,384</td>
<td>13,572</td>
<td>19,230</td>
</tr>
<tr>
<td>Vocabulary size in words</td>
<td>50,121</td>
<td>38,696</td>
<td>58,085</td>
</tr>
<tr>
<td>Vocabulary size in pseudo-morphemes</td>
<td>20,117</td>
<td>15,302</td>
<td>23,983</td>
</tr>
</tbody>
</table>

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 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) (Lopez de Ipina et al., 2002).

Figure 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).

Figure 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.

Figure 3: Coverage (a) and OOV rate (b) for the textual sample BCNEWS

3. Resources Developed

3.1 Resources in Spanish

- 6 hours of video in MPEG4 (WMV 9) format of “Teleberri” program, the daily program of broadcast news in Spanish.
- 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 “Telberri” program. The text is divided in sentences and paragraph.
- 1 year of local newspapers in Spanish Gara (GARA), in text format. The text is divided in sentences and paragraph.
- Lexicon extracted from the XML transcription files, including morphologic, phonologic and orthographic information.

3.2 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 (EITB).
- 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 (Egunkaria)), in text format.
- Lexicon extracted from the XML transcription files, including phonologic, orthographic, and morphologic information.

### 4. Processing Methodology

#### 4.1 Processing of the audio data

The audio data has been extracted out from the MPEG4 video files, using FFmpeg free software\(^3\). 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 (Barras et al., 1998). The XML files include information of distinct speakers, noises, and paragraphs of the broadcast news. The files also contain phonetic and orthographic information of each of the words. As Basque is an agglutinative language with very rich inflection variety (Alegria et al., 1996), Basque XML files include morphologic information such as each word’s lemma and Part-Of-Speech tag. 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.

#### 4.2 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.

#### 4.3 Processing of the textual data

There are two independent types of textual resources: The text extracted from the newspapers Gara (GARA) and Euskaldunon Egunkaria (Egunkaria)), and the scripts of the "Teleberri" and "Gaur Egun" programs. These last resources are very

\(^3\) Available online at http://ffmpeg.sourceforge.net
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.

5. Concluding Remarks

Basque and Spanish are both official languages in the Basque Country, and they
are used in the Basque Public Radio and Television EITB (EITB) and in most of
the mass media of the Basque Country. Hence, this work deals with the
development of appropriated resources for a bilingual automatic index system of
broadcast news in Basque and Spanish. Since Basque is an agglutinative language,
analysis of coverage and words OOV has been carried out in order to develop an
appropriate Lexicon. Finally, lexicons are created for both languages using
morphologic and phonetic information, not just extracting a word list.

Bibliography

- Informedia, digital video understanding research, http://www.informedia.cs.cmu.edu/
- Euskaltzaindia, http://www.euskaltzaindia.net/
- Peñagarikano M., Bordel G., Varona A., Lopez de Ipina: "Using non-word
Lexical Units in Automatic Speech Understanding", Proceedings of IEEE,
ICASSP99, Phoenix, Arizona.
- 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
Segmentation for Speech Processing in Basque" IEEE TTS Workshop. Santa
Monica USA. 2002.
- GARA, local Basque Country newspaper in Spanish, online at
http://www.gara.net/
- Egunkaria, Euskaldunon Egunkaria, the only newspaper in Basque, which has
been recently replaced by Berria, online at http://www.berria.info/


