Lexica and Corpora for Speech-to-Speech Translation: A Trilingual Approach

David Conejero, Jesús Giménez, Victoria Arranz, Antonio Bonafonte,
Neus Pascual, Núria Castell, Asunción Moreno

Talp Research Center
Universitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona
david@talp.upc.es

Abstract

Creation of lexica and corpora for Catalan, Spanish and US-English is described. A lexicon is being created for speech recognition and synthesis including relevant information. The lexicon contains 50K common words selected to achieve a wide coverage on the chosen domains, and 50K additional entries including special application words, and proper nouns.

Furthermore, a large trilingual spontaneous speech corpus has been created. These corpora, together with other available US-English data, have been translated into their counterpart languages. This is being used to investigate the language resources requirements for statistical machine translation.

1. Introduction

The development of a Speech-to-Speech Machine Translation (SST) system involves high performance components for Speech Recognition (SR), Machine Translation (MT) and Text-to-Speech Synthesis (TTS). Therefore, most attempts to develop such systems have been performed in projects including various partners with expertise in these areas. Some well-known references are the C-Star Consortium [1] , and the NESPOLE [2], Verbmobil [3], and Eutrans [4] projects.

These projects showed that it is possible to develop robust SST systems for small- to medium-sized domains using sophisticated SR and MT technology. The major problems in this area are: the acquisition of domain-specific training data (either mono- or bilingual), the robust behaviour of the MT component for SR errors and spoken language phenomena, and the development of an efficient recognition and translation component.

Language resources for SR have proved to be very useful to improve SR technology. Oral databases have been created in a large number of different languages. Furthermore, some producing-tools and standards have been created. Many databases are publically available at the European Language Resources Association (ELRA) or the Linguistic Data Consortium (LDC). In the last few years, the performance of speech synthesizers has also improved dramatically due to the usage of oral language resources. However, large lexica specifically designed for speech applications are difficult to find. Speech recognizers and synthesizers need good lexica with a wide coverage of the language, appropriate morphosyntactic and phonetic information.

Moreover, parallel bilingual texts are needed so as to enhance the performance of Statistical Machine Translation (SMT) systems. Electronic texts are only available for a small number of languages and no standard exists so far. Regarding large lexica in Catalan and/or Spanish, EuroWordNet is the only resource available at ELRA that may be useful for this purpose, although it was not designed on the speech technology perspective. In what regards large bilingual corpora (and specially for spoken languages) these are very difficult to find.

This paper shows the creation of large lexica in Catalan and Spanish for SR and TTS. This can be seen in section 2, together with the description of the corpora and word lists used. The paper also describes (section 3) the creation of a large sentence-aligned trilingual corpus (Catalan, Spanish and US-English) from spontaneous speech recordings in a tourist domain. This corpus will be used to investigate the requirements for SST-oriented LR and then to establish the specifications for their development. Section 4 describes the recording platform, the design of the scenarios, as well as the collection, annotation and translation processes involved. Section 5 describes a speech-to-speech translation demonstrator based on the LR generated in the project. This work is being carried out within the framework of the LC-STAR[5] (*Lexica and Corpora for Speech-to-Speech Translation Components*) project.

2. Word Lists and Lexica for SR and SST

One aim of the project is to create large pronunciation lexica suited for speech recognition and synthesis. Each lexicon will consist of two parts: one covering common words and another one covering proper nouns from a broad range of domains. Both parts will comprise 50K entries each. The proper nouns part will consist of 45K names and 5K application words that were selected by the consortium. The first step taken to create the lexica was the definition of the word lists that would constitute them.

2.1. Corpus and Word Lists Creation

In order to generate the common word list, a corpus with six different domains was defined as listed in Table 1. According to those domains, the text corpora were collected from different sources: on-line newspapers, magazines and other websites.

Once the corpus was collected, it was cleaned and normalized: scripts, frames and html marks were removed. Then, the corpus was lemmatized and POS tagged using MACO+ [1]. Punctuation marks and proper nouns were dropped. Table 1

[1] http://www.c-star.org
[2] http://nespole.itc.it
[3] http://verbmobil.dfki.de/verbmobil
[4] http://www.zeres.de/Eutrans/eutrans.html
[5] http://www.lc-star.com

shows the list of domains, corpus size per domain and number of distinct words in the Spanish and Catalan corpora.

Domains	*Spanish*		*Catalan*	
	Words per Domain	*Diff. Words*	*Words per Domain*	*Diff. Words*
Sports/Games	4.49 MT	73 KT	1.73 MT	31 KT
News	18.80 MT	161 KT	9.98 MT	99 KT
Finance	3.79 MT	56 KT	1.62 MT	33 KT
Culture/Ent.	4.30 MT	87 KT	5.01 MT	85 KT
Consumer Inf.	1.58 MT	42 KT	1.31 MT	44 KT
Personal Com.	3.68 MT	106 KT	0.55 MT	74 KT
TOTAL	*36.65 MT*	*254 KT*	*20.20 MT*	*163 KT*

Table 1: Common words per domain distribution

The objective of the word list is to have a wide lexical coverage for every domain. The LC-STAR consortium agreed that the word list should contain at least 50K words reaching a target for each domain of at least 95% self coverage. To reach the 50K different entries, the self coverage target was increased up to 97,5% in Spanish and 98,6% in Catalan. The selection process is explained in [2]. At the end of the process, a manual revision was done to handle problems arisen on the automatic process. A summary of the figures for these word lists is shown in Table 2. Coverage has been calculated with singletons included in the corpus domains, but excluded from the word lists:

Domains	*Spanish*		*Catalan*	
	Selected Words	*Cover. %*	*Selected Words*	*Cover. %*
Sports/Games	21,192	98.74	13,672	99.47
News	28,457	98.41	31,463	98.98
Finance	15,301	99.00	15,470	99.45
Culture/Ent.	29,471	98.36	33,794	98.89
Consumer Inf.	19,600	98.83	24,969	98.89
Personal Com.	31,802	97.96	16,293	96.79
TOTAL	*55,788*	*98.48*	*53,225*	*99.30*

Table 2: Selected word lists and coverage per domain

The proper nouns word list contains more than 45K entries and is divided into 3 different domains: First and Last names, Place names and Organizations. In order to fulfill the specifications, each domain should cover a minimum of 10% and a maximum of 50% of the entries.

2.2. Lexica for Speech Recognition and Synthesis

Once the word lists in the previous section had been created, they were used as input for the creation of monolingual lexica for SR and TTS. This is a complex task that must take into account the requirements and needs imposed by these technologies. Furthermore, the lexica produced must also provide sufficient and appropriate monolingual information so as to be linked with the translation lexica and, thus, contribute to speech-to-speech translation. In order to achieve this, a number of tasks have been considered (for more details see [3]):

- **Creation of language-independent specifications for the content of the lexica as well as addition of some language-dependent grammatical information**
  This considers issues such as the required and optional orthographic, phonetic and morpho-syntactic properties of each entry in the lexicon, as well as the size of the lexica.
- **Creation of language-specific specifications for each language**
  This deals with the grammatical and morphological representation, as well as with the XML-based exchange format used. The following section will expand on this task.

The creation of the language-specific specifications is as follows: firstly, a detailed POS study and description for the 12 languages of the project have been carried out, considering all possible phenomena. A set of 21 general POS tags has been adopted, where each POS is divided into a detailed set of attributes (e.g., POS for nouns contains attributes such as *Class, Number, Gender, Person, Case, Appreciative*, etc.). Then, a DTD has been created that implements the basic POS scheme representing all necessary features. Each language has made use of this DTD as a reference point in order to build its own language-specific POS scheme, taking into account only those phenomena relevant for the language. This DTD considers simple entries for unique word forms and entry groups for complex forms (cf. below).

Several attributes require special attention, in particular those handling the complex construction of Catalan verbal forms with *clitic pronouns*[6]. When these pronouns follow verbal forms (up to 3 of them in Catalan), they are assimilated by the verbs building up unique forms. A sample XML entry for one of these complex forms (*dóna-m'ho* - give to me it) can be seen below:

```
<ENTRYGROUP orthography="dóna-m'ho">
 <ENTRY_COMP>
  <PHONETIC>" d o - n @ - m u</PHONETIC>
  <ENTRY_EL orthography="dona">
   <VER person="2" tense="present"
    number="singular" voice="active"
    mood="imperative"/>
   <LEMMA>donar</LEMMA>
  </ENTRY_EL>
  <ENTRY_EL orthography="m'">
   <PRO person="1" number="singular"
    gender="invariant" type="personal"/>
   <LEMMA>jo</LEMMA>
  </ENTRY_EL>
  <ENTRY_EL orthography="ho">
   <PRO gender="neuter" type="personal"
    person="3" number="invariant"></PRO>
   <LEMMA>ho</LEMMA>
  </ENTRY_EL>
 </ENTRY_COMP>
</ENTRYGROUP>
```

Some other interesting phenomena that take place in Catalan (mostly in Spanish too) and that have required special attention are: 1) *Prepositional expressions*, referring to phrases that function as prepositions and always end up in prepositions; 2) *Politeness attribute*, which takes place in personal pronouns *vostè/vostès*; 3) *Oblique case* for those personal pronouns *mi, ti, m', l'...* following a prepositition or apostrophed; etc. For a detailed description on the specifications generated for Catalan and Spanish languages, please refer to [3].

3. Parallel Corpora and Lexica for SST

The evaluation results obtained in the Verbmobil and the Eutrans projects showed that a very promising approach for the

[6]This phenomenon also takes place in Spanish.

translation component is statistical machine translation. This approach resulted in an error rate which is better by a factor of two in comparison to the other approaches investigated in Verbmobil. This is the approach that LC-STAR has undertaken.

Regarding the type of LRs necessary for speech centered translation, we can distinguish between corpora and lexica. Needless to say that the limitations caused by the present small reference LR are still an objective to be overcome, and this is one of the aims of LC-STAR. In particular, for SST, it is the bi- or multi-lingual collections of on-line data that are highly sought for.

The aim of the LC-STAR project is not only the creation of LRs for SST but also the specification of how lexica and corpora should be created. Taking into account the above mentioned problem of training data, the specification should consider features of LRs in order to be as useful as possible for statistical SST as well as features to reduce the amount of needed data. Always in the tourist domain, the final aligned text corpora (Catalan, Spanish and US-English) will have a size of 750K words, and the monolingual lexica will contain 10K entries per language (Catalan, Finnish, German, Hebrew, Italian, Russian, Spanish and US-English).

One part of the text corpora has been obtained from transcribed speech US-English corpora. The other part of text corpora has been obtained from the transcription of conversations recorded in Catalan and Spanish (see next section). The transcriptions in one language have been manually translated into the other two languages. The golden rule for translation has been to produce target sentences as similar as possible to the source ones, but being at the translator criterion likely to be ever uttered by a competent speaker of the target language. That is, translators tried to be as literal as possible but always preserving the meaning and correctness of the utterances. The translation is not required to be the best one.

This trilingual corpus is the main SMT system training data. The literality criterion was imposed in order to simplify the SMT system training, specially with respect to the alignment process[4]. The trilingual corpus will be adapted to further aligned corpus specification criteria.

From the US-English corpus and other sources, a reference list of 10K words has been created and will be translated into the seven mentioned languages. As the translation must be domain oriented, some 5-grams are provided with each US-English word in order to help human translators. These lists will be completed according to the further specification criteria so as to obtain the eight monolingual lexica for speech centered translation. The Catalan, Spanish and US-English lexica will be used to improve the statistical SST.

4. Oral Database

As already mentioned in the previous section, most of the text corpora come from the transcription of Catalan and Spanish spoken dialogues.

In the LC-STAR project it was decided to focus the research on the tourist domain. We have chosen a subset subset of the tourist domain, namely tourist-employee conversations. Four scenario categories were defined, namely *Hotel, Travel Agency, Tourism Office and Railway/Airline Company*.

4.1. Recording Sessions

In order to avoid non-verbal communication, we decided to record the spoken dialogues through the telephone network: the speakers were placed in different rooms and talked on the phone with each other. Dialogues would last ten minutes in average. In order to do so, a recording platform was set up. The platform can be used in two different modes. In the first one, the platform is transparent: the two speakers may overlap each other. But we agreed that this mode was not well matched to the translation system, where a machine is between the speakers. Therefore, we designed and used a second version which imposes a rigid turn strategy, i.e speakers are not allowd to speak both at the same time. The first turn is given to the speaker receiving the phone call. The speaker in possession of the turn has to indicate a turn exchange by pressing a key. Conversations contain some disfluences such as false starts, corrections, repetitions, filled pauses, and certain ungrammaticalities.

Oral DataBase	
Spanish	
speech raw time	31h:7m:32s
#speakers	77
#dialogues	217
#turns	10.998
#sentences	24.372
#words	349.970
#distinct words	11.714
Catalan	
speech raw time	23h:43m:55s
#speakers	56
#dialogues	172
#turns	9.321
#sentences	19.113
#words	277.777
#distinct words	10.057

Table 3: Oral Database: Some figures.

For the conversations to yield the pursued information some specific subscenarios were designed. A series of templates were built so as to assist speakers at conversation time. They were used as a draft or schema containing a description of the information to talk about in every subscenario. In most cases, if not all, *'Speaker 0'* played the role of a tourist/customer, and *'Speaker 1'* played the role of an employee. Below follows an example of one such situation:

Example: Hotel Reservation. *'Speaker 0'* impersonates a tourist trying to book accommodation at a given hotel for a certain date, a given number of people, under certain specific conditions. *'Speaker 1'* is acting as a hotel employee providing the required information.

4.2. Contents

It was agreed that at least 500K words should be recorded considering Catalan and Spanish as a whole. The total figures for both languages can be seen in Table 3. All dialogues have been manually transcribed and validated. During the transcription process, some tags have been added so as to encode acoustic and linguistic information.

Thus are the encoded acoustic phenomena:
- *Lengthening:* Lengthening of a sound within a word denoting hesitation.
- *Filled pause:* Pause filled by a vocalization. It may carry some meaning, i.e. affirmation, negation, hesitation, etcetera.
- *Bad pronunciation:* Speaker mispronounciations possibly

causing some uncertainty in the transcription.
- *Unidentifiable:* Incomprehensible phonetic strings.
- *Hard to Identify:* Those words which the transcriber was highly positive for but not completely sure.
- *Speaker noise:* Meaningless sounds produced by the speaker, for instance, when laughing, swallowing or coughing.
- *Stationary noise:* Persistent background noise.
- *Impulsive noise:* Noises not produced by the speaker but due to technical instruments, i.e. telephone ringing, dropping things, etcetera.
- *Technical interruptions:* Temporary interruption of the audio signal caused by technical problems.

As to the encoded linguistic phenomena:
- *Punctuation Marks:* Standardized marks ('¿', '?', '¡', '!', ',', '.').
- *Foreign Words:* Words not belonging to the source language.
- *Named Entities:* Proper Nouns (names of people, places, institutions, etc.)
- *Letter Spelling:* Act of spelling out a word letter by letter.
- *Abbreviations:* Acronyms optionally spelled.
- *Neologisms:* Words either made up by the speaker or simply not appearing in the dictionary.
- *Repetition or Correction:* The speaker either repeats or corrects himself.
- *False Start:* The speaker gives up the the production of an utterance before it is completed and starts a new one from scratch.

It was agreed to use the Extensible Markup Language (XML) to represent this information. All the relevant XML tags have been mainteined during the manual translation process.

5. Demonstrator

A platform is being developed in order to show the language transfer. It is based on speech-to-speech translation components using the tourist domain LRs that have been generated for Catalan, Spanish and US-English. The platform acts as a telephone server. In first place, the user acting as a tourist calls the platform, then selects the language she is to speak and then dials the destination number. Then her counterpart, who receives the phone call, selects the language she's to speak, too. The dialogue starts having the SST system in between as an interpreter. Some of the LRs produced have already been used in the SST components. The rest of them are expected to be incorporated very soon.

5.1. Speech Recognition

The main ingredients to build a SR system are acoustic models, language models and phonetic information. The baseline acoustic models are estimated from SpeechDat databases [5]. However, most of the SpeechDat contents are read style and not from the tourist domain. Much better results are obtained when the oral database produced in the project is used to estimate the acoustic models. Using less than a half of training data, the word error rate evaluated on an independent test set is reduced by 30%. The transcriptions of the oral database are being used to estimate a stochastic language model. The size of the lexicon is 10K and the perplexity evaluated over the test set is 72. Class based n-grams will be used to improve the probability estimation for named entities (cities, hotels, etcetera). The pronunciation will be derived from the lexica as soon as they are finished. We expect that to produce some improvements, specially for US-English and for foreign words.

5.2. Speech Translation

A statistical machine translation system trained on the trilingual corpus, constitutes the baseline system. The LC-STAR project will study how the performance of this component can be improved using new information, such as the multilingual lexicon which is currently being produced.

5.3. Text to Speech

The text-to-speech component will also benefit from a wide coverage lexicon. The lexicon would be used to derive better pronunciations. Particularly, acronyms, abbreviations, proper nouns and foreign words are tokens that cause many pronunciation problems to our baseline TTS system. The use of the morphosyntactic information (part-of-speech) in the lexicon will allow us to produce better phrasing algorithms. This information will also be used to determine for instance the gender in which the numbers have to be read.

6. Summary

In this paper we have described the process of creation of large lexica for Speech Recognition and for text-to-speech Synthesis. It is shown how 50K words allow a self coverage larger than 97%. We have also explained the process of design, recording and labeling of an oral database consisiting of nearly 400 spoken dialogues. The oral corpora have been recorded either in Catalan or Spanish and have been and are being translated to their two counterparts so as to generate a large trilingual corpus (Catalan/Spanish/US-English). The corpus size is 24 hours of speech for Catalan, i.e. 278K words, and 31 hours of speech for Spanish, i.e. 350K words. A reference US-English list has been created using the corpora. A lexicon containing useful information for machine translation is also being produced. A demonstration platform has been implemented. It is being constantly updated as the language resources are available.

7. References

[1] Carmona, J. et al, "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain, 1998.

[2] Hartikainen, E., Maltese G., Moreno, A., Shammass, S., Ziegenhain, U. "Large Lexica for Speech-to-Speech Translation: from specification to creation" REF : submitted to this conference.

[3] Maltese, G. et al "General and language-specific specification of contents of lexica", LC-STAR Technical Report D2.1, D2.2, D2.3, D2.4, V1.0, 2003

[4] Och, F.J., Ney, H. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, Vol.29, N.1, Pages 19-51, 2003.

[5] Draxler C. et al "Experiences in creating Large Multilingual Speech Databases for Teleservices", Proceedings LREC98, Vol.I, pp 361-366, 1998, Granada, Spain.