Generation of Language Resources for the Development of Speech Technologies in Catalan

A. Moreno[1], Albert Febrer[2], Lluis Márquez[3]

(1) TALP Research Center. Universitat Politècnica de Catalunya. Barcelona, Spain

(2) Applied Technologies on Language and Speech” (ATLAS). Barcelona, Spain

Abstract

This paper describes a joint initiative of the Catalan and Spanish Government to produce Language Resources for the Catalan language. A similar methodology to the Basic Language Resource Kit (BLARK) concept was applied to determine the priorities on the production of the Language Resources. The paper shows the LR and tools currently available for the Catalan Language both for Language and Speech technologies. The production of large databases for Automatic Speech Recognition purposes already started. All the resources generated in the project follow EU standards, will be validated by an external centre and will be free and public available through ELRA.

1. Introduction

Catalan is a European language spoken in the East of Spain (Catalonia, Balearic islands and Valencia), Andorra, and South of France. Catalan is an official language in Spain and currently is spoken by 6 million people.

The Catalan Government jointly the Spanish Ministry of Industry, Tourism and Commerce, launched during 2005 a program to promote the Catalan language in the scope of the Information Society.

Inside the program, some priority areas were identified: Education in the University context, automatic translation and applications, help to disabled people, accessibility to Internet, and improving Human Language Technologies (HLT).

Concerning HLT, the main goal is to contribute to the progress of HLT for the Catalan language. The objective is to make available for companies and research centres the necessary language resources to:

1. Place the Catalan language to the same level of development as other European languages in availability, accessibility and quality of language resources.
2. Promote portability of the technologies developed in other languages to Catalan.
3. Attract technological business to develop products in Catalan.
4. Make accessible to the scientific community tools and linguistic resources.
5. Promote the research in the fields of speech processing and natural language processing in Catalan.

The goals are close to those found in the Dutch-Flemish HLT program (Cucchiarini and D'Halleweyn 1994) and linked to other initiatives as the Basic Language Resource Kit (BLARK). Indeed, Catalan is a minority language, the EU projects oriented to collect large LRs rarely included minority languages, the potential market is small to make companies to invest in such minority languages and the development of the technology in this field should be partially promoted from the local Governments and official institutions.

This paper is organized as follows: in Section 2 we summarize the BLARK initiative. In section 3 we describe the current LRs available for the Catalan language. In Section 3 we describe the LRs already initiated in the project and finally, the paper ends with some future work.

2. The BLARK initiative

In 1998, Steven Krauwer launched in the ELRA newsletters the BLARK concept. The Basic Language Resource Kit (BLARK) is the minimal set of language resources (LR) necessary to do research in the field of HLT. A BLARK comprises:

- Basic language resources:
    - written language corpora
    - spoken language corpora
    - bilingual (written) corpora
    - mono- and bilingual dictionaries
    - terminology collections
    - grammars
- Benchmarks for evaluation
- Basic tools:
    - modules (e.g. taggers, morphological analyzer, parsers, speech front-ends, grapheme-to-phoneme converters, statistical disambiguators, …)
    - annotation standards and tools
    - corpus exploration and exploitation tools
    - etc

The BLARK initiative is a starting point for academia and pre-competitive industry research in language and speech technologies.

3. Existing Resources in Catalan

There are several research groups in Catalonia working in language and speech technologies. A first step in the project was to collect information about which language resources and tools were available. Language Resources for Language and Speech Processing are summarized separately. Concerning language technologies, the resources can be summarized as follows:

3.1 Resources for Language Processing

A quantity of language resources and tools exist for the Catalan language. Most of the LR and tools have been generated under UE funded projects: Onomastica, EuroWordnet, Namic, Meaning, etc. We distinguish three categories where we found resources in Catalan

3.1.1. Language Databases

A- morphological (words, lemmas, part-of-speech, gender, number, person, etc.)
B- morpho-syntactic (locutions, periphrasis, etc.)
C- grammars (syntactic rules, syntactic phrases, etc.)
D- semantics (concepts, ontologies, thesaurus, etc.)
E- terminology

3.1.2- Corpora

A- raw texts
B- annotated with different kind of information: morpho-syntactic, syntactic, semantic, etc.
C- multilingual comparable corpora (e.g., two newspapers of different languages covering a similar time period)
D- multilingual parallel corpora (e.g., the same newspaper in two languages)

3.1.3 – Basic Processing Tools:

A– morphological analyzers (i.e., tools that use 3.1.1A to analyze a text)
B– multiword detectors (i.e., tools that use 3.1.1B to recognize person/place/organization proper nouns, locutions, numeric expressions, dates, hours, etc.)
C – morpho-syntactic disambiguators (i.e., tools that use 3.1.2B as training corpus to later perform POS tagging)
D – syntactic analyzers (i.e., tools that use 3.1.1C and/or corpora 3.1.2B to parse sentences, phrases, etc.)

The main applications were these Language Resources are currently used are:

- Automatic Machine Translation, (both rule and statistical based approaches)
- Information Retrieval (intelligent document searchers)
- Question Answering systems
- Information Extraction, IE (i.e., systems that are able to automatically gather structured information from texts, webs, etc.)
- Text Classification, both by topic and subjective intention.
- Supporting tools for lexicographers and linguists.

3.2 Speech Processing Language Resources

For speech technologies development, the following public available databases have been created.

1. Microphone database: 100 speakers recorded during 1 hour each. The database is intended for dictate applications
2. Fixed telephone database. The database consist of read words and sentences recorded from 1000 speakers. The database fulfils the SpeechDat II specifications
3. Spontaneous telephone recordings. The database consists of 100 speakers recorded in the fixed telephone network.
4. Speech Synthesis databases. Small databases (one male, one female) for speech synthesis applications. The databases are segmented into diphones.

In addition, there are automatic grapheme-to-phoneme converters available.

4. Language Resources under Production

In a first stage, it was decided to collect Language Resources for Automatic Speech Recognition Purposes. The objective was to collect databases useful for research in academic and pre-competitive industrial fields, as in the BLARK model, but attractive as well for companies interested in porting their products. For this purpose, the language resources to be created in this project will be:

Available:
In general, the term available means that the LR can be found in the market. An important issue is the price of those databases. It was decided that the databases will be free of cost both, for research and for commercial use

Standard:
One of the goals of the project is to facilitate the portability of the available systems to Catalan. To facilitate this task, is sure that the database has to be compatible and consistent with other LR, in our case, with the major LR created in Europe. For this reason, it was decided to follow, for each resource created in the project, the available standard specifications created in previous projects (SpeechDat II, SpeeCon and SALA II)

Task:
It was decided to create databases needed for a wide number of applications and domains.

Quality:
All the final databases are checked against validation criteria. Most of the LR created in European projects (SpeechDat, Orientel, Speecon…) have been checked by an external center. The purpose is to assure that the database fulfils the specifications, the documentation is complete and the annotation is correct.

The following databases are created:

4.1 A fixed telephone database.

Consist of the recordings of 2000 speakers half male and half female. Recordings are made trough the fixed telephone network at 8KHz sampling rate and coded with A law. The corpus is designed to support the creation of teleservices commanded by voice. The database is read, each speaker read about 40 sentences and words including isolated and connected digits, natural numbers, money, spellings, dates, hours, yes/no sentences, names, surnames, city names, company names, common application words embedded in sentences and phonetically rich sentences and words. Speakers are distributed among all the dialectal regions where Catalan is spoken in Catalonia, Valencia and Balearic islands. The database is orthographically annotated with a set of clearly audible noises (breath, door slam…). The work is fully manual. An accompanying lexicon contains all the words uttered by the speakers and their phonetic transcription in SAMPA symbols. The database is extensively documented and validated. The database follows SpeechDat II (Winski R. 1997), (Senia F. 1997), (Senia, F., J.v. Velden J. v. 1997) specifications.

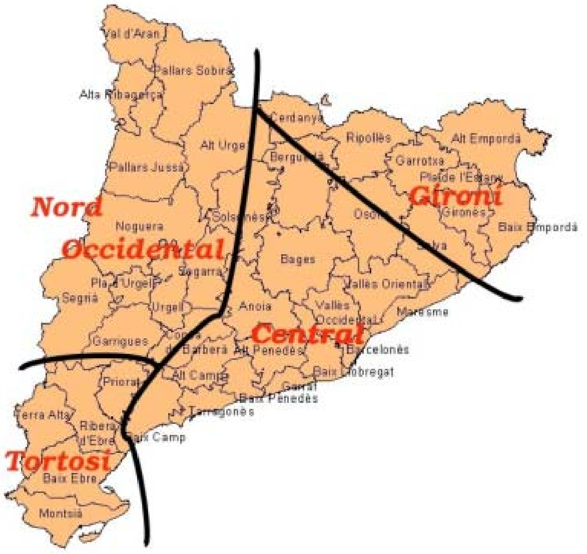

Figure 1. Catalonian dialectes

4.2 A mobile telephone database.

Consist of the recordings of 2000 speakers. The database is collected from mobile telephones. The corpus, speaker recruitment and documentation are very close to the above mentioned Fixed Network Database. In this case, the database follows the SALA II specifications (Moreno, A., Senia, F. 2002). The choice of SALA II instead of SpeechDat is due to the successive updates that are in the SALA II specifications (Moreno, A., Senia, F. 2002) for mobile telephone recordings.

Telephone	Environment	Full database distribution	Dialectal region distribution
Fixed (2000)	Home/Office	100%	100 to 600
Mobile (2000)	Car	20 % ± 5%	≥ 20 %
	Public place	25 % ± 5%	
	Street	25 % ± 5%	
	Home/Office	25 % ± 5%	≥ 20 %
	Car kit	5 % ± 1%	No constrain

Table 1. Environmental distribution per dialectal region and in total in the telephony databases.

4.3 A microphone database.

This database consists of the recordings of 550 adult speakers. Up to four microphones are recorded simultaneously in a mobile recording platform at 16 KHz sampling rate. One recording session consists of 291 read utterances plus 30 spontaneous spoken utterances.
In a session the following information is recorded:

Calibration: Acoustic measurements of the recording place

Free spontaneous speech: Five minutes of free spontaneous speech.

Short spontaneous utterances: dates, hours, proper nouns, city names, spellings, answer to easy questions, telephone numbers, language.

Basic read words and sentences: Consist in 31 words and general sentences as isolated digits, sequences of digits, telephone numbers, natural numbers, money amounts, dates, hours, spellings, proper nouns, cities, yes, no, web and email addresses.

Application words: In addition to the former words, each speaker read 208 application words and sentences for access to IVR systems, browser, edit, control, internet, audio, video,…

Phonetically rich words and sentences. Each speaker utters 30 phonetically rich sentences and some phonetically rich words.

Recordings take place in a wide quantity of environments. From office, 75 recordings are expected; from an entertainment environment (i.e. home) 200 speakers will be recorded. 75 recordings will be recorded from cars and 200 speakers will be recorded in public places.
The database is annotated orthographically. Some recorded noises are annotated as well by means of adequate codes. A lexicon containing each uttered word and their phonetic transcription is included in the

database. The database is fully documented. The database follows the Speecon documentation (Kiessling, A. et al.2001).
This database has a very interesting component for research (recognition, noise modelling, etc) and in addition has a clear interest for industries. In fact, the UE funded project Speecon has been very successful given the big number of databases created under the Speecon specifications

4.4 A car database.

This database contains the recordings of 600 sessions from 300 speakers. One session consist of 119 read utterances. In addition, 200 sessions contains spontaneous recordings. The recording platform is mobile and can record four microphones. Two microphones are placed in the car and the speaker wears other two microphones. The corpus consist of keywords, isolated and digits, telephone numbers, natural numbers, money amounts, dates, hours, spellings, proper nouns, cities, yes and no words, phonetically rich sentences and words and specific application words for the management of mobile telephone , IVR functions and car devices.
Recordings are made in different environments (city traffic, road, highway), with specific noise conditions (windows open/close, radio on/off ,…) The Database follows the SeechDat car specifications (S. Dufour, S. 1999), (Heuvel, V.d.H., 1999), (Draxler, C. et al. 2000), except the recording platform that is built as in the Speecon project.

All databases will be validated by an external center with the validation criteria of the corresponding above mentioned projects. Validation criteria is available at (Heuvel, V.d.H. 1997), (Heuvel, V.d.H.2002) , (Heuvel, V.d.H., Shammass, S., Moyal, A., Gedge, O. 2001), (Heuvel, V.d.H. 2000)and the validation report will be included in the database documentation.
Distribution will be via ELRA. All databases will be public and free (except shipping costs). The project began on September 2005 and ends on mid 2006.

5. References

Cucchiarini C., and D'Halleweyn E. (1994) ."The new Dutch-Flemish HLT Programme: a concerted effort to stimulate the HLT sector" LREC 04

Draxler, C. et al. (2000) D133 - Specification of Database Interchange Format Deliverable of the SpeechDat car project 2000

Dufour, S. (1999) D112 - Specification of the car speech database (definition of corpus, scripts and standard), Car environments and speaker coverage. Deliverable of the SpeechDat car project

Krauwer, S., Maegaard, B., Choukri, K., Lise Damsgaard Jørgensen .(1994) . Report on Blark for Arabic. Internal report for NEMLAR project.

Heuvel, V.d.H. (1997) SD1.3.3 - Validation criteria for databases Deliverable of the SpeechDat project.

Heuvel, V.d.H., H. (1999) D132 - Orthographic transcription conventions. Deliverable of the SpeechDat car project

Heuvel, V.d.H. (2000) CD131 - Validation criteria Deliverable of the SpeechDat car project.
Heuvel, V.d.H.(2002) Proto_Design SALA II Databases. Deliverable of the SALA II project Nov 2002. http://www.sala.org

Heuvel, V.d.H., Shammass, S., Moyal, A., Gedge, O. (2001) D41 - Definition of Validation Criteria Deliverable of the Speecon project.

Kiessling, A. et al.(2001) D21 - Specification of Databases Speecon project . Deliverable of the Speecon project.

Moreno, A., Senia, F. (2002) The complete SALA II project specifications. Deliverable of the SALA II project Nov 2002. http://www.sala.org

Senia F. (1997) SD1.3.1 - Specification of speech database interchange format, Deliverable of the SpeechDat project.

Senia, F., J.v. Velden J. v. (1997) SD1.3.2 - Specification of orthographic transcription and lexicon conventions Deliverable of the SpeechDat project.

Winski R. (1997) SD1.1.1 Definition of corpus, scripts and standards for Fixed Networks Deliverable of the SpeechDat project.