

# Development of Language Resources for Speech-to-Speech Translation

Victoria Arranz and Núria Castell and Jesús Giménez

Talp Research Centre

Universitat Politècnica de Catalunya

Jordi Girona, 1-3

08034 Barcelona, Spain

{varraz,castell,jgimenez}@talp.upc.es

## Abstract

This paper describes the creation of linguistically enriched aligned corpora for Catalan, Spanish and US-English for Speech-to-Speech Translation. These corpora are obtained from two different sources: US-English transcribed speech data and transcriptions of conversations recorded in Catalan and Spanish. After human translation, a large trilingual spontaneous speech corpus has been obtained. This corpus is being linguistically enriched in order to generate a valuable resource capable of helping to improve statistical machine translation.

## 1 Introduction

The past decade has experienced a boost in the creation and sharing of language resources (LR). A considerable effort has been invested in the design and construction of these resources, which are so necessary for the different areas involved in language technology. The International Conferences on Language Resources and Evaluation<sup>1</sup> shown the evolution of this area.

Regarding the type of LR necessary for speech centred translation, we can distinguish between corpora and lexica. As in textual data processing, both resource types can be mono- bi- or multilingual. However, when it comes to speech-to-speech translation (SST), it is the bi- or multi-lingual collections of on-line data that are highly sought for. Furthermore, multilingual parallel corpora are also very valuable resources for many other researches besides SST, such as foreign language resource acquisition (Yarowsky & Ngai 01), multilingual resource acquisition (Lopez *et al.* 02) or sense discrimination (Ide *et al.* 02).

In relation to the design and construction approaches to be followed in LR for SST, these differ from those adopted when generating on-line collections of textual data for general NLP purposes. The translation of spontaneous speech causes additional problems mainly related to speech recognition problems and phenomena in the speech in-

put for translation. Phenomena such as unconventional syntax, lexical variations and other disfluencies (Alvarez *et al.* 01) must be handled.

In order to tackle these problems, a translation system should be built robust enough to adjust to the incoming input. A number of approaches have been followed in SST, which can be generally classified into corpus-based methods ((Alshawi *et al.* 98), (Ney *et al.* 01)) and rule-based linguistic methods ((Levin *et al.* 00), (Lavie *et al.* 01)). It is mainly for the former approach that the current work is aiming for, i.e. helping to improve statistical SST. Work on SST can also be found in the results of large research projects such as C-star<sup>2</sup>, Nespole!<sup>3</sup>, Verbmobil<sup>4</sup>, and Eutrans<sup>5</sup>.

This document focuses on the work carried out within the frame of the LC-STAR project<sup>6</sup> (Lexica and Corpora for Speech to Speech Translation Components). It describes the design and construction of a trilingual spontaneous speech corpus. At present, a subset of this corpus has been enriched with morphological information. Some initial tests using this corpus have already been done by some other project partners, which have confirmed its usefulness to statistical SST.

The paper is organised as follows: section 2 describes the use and expansion processes carried out on some existing speech data towards the construction of subset-1 of our corpus. Section 3 explains the design and construction of subset-2 of the corpus. Section 4 proceeds onto the morphological annotation performed on subset-1. Section 5 draws some conclusions and provides suggestions for further work.

## 2 LRs Used and Expanded

Subset-1 of our trilingual corpus has been created based on some already available data. Before

<sup>1</sup><http://www.lrec-conf.org/>

<sup>2</sup><http://www.c-star.org>

<sup>3</sup><http://nespole.itc.it>

<sup>4</sup><http://verbmobil.dfki.de/verbmobil>

<sup>5</sup><http://www.zeres.de/Eutrans/eutrans.html>

<sup>6</sup>(IST-2001-32216): <http://www.lc-star.com>

taking any decisions on which existing resources to use for our purposes, those available from LR repositories such as ELRA<sup>7</sup> and LDC<sup>8</sup> were considered. It was decided to use part of the recordings from the Verbmobil project.

## 2.1 Verbmobil

Verbmobil was a long-term project of the German Federal Ministry of Education, Science, Research and Technology (BMBF, Projektträger DLR). One of its outputs was a valuable set of spontaneous speech databases, which are available via ELRA.

Our aim is the creation of speech-to-speech centred and linguistically enriched trilingual LR for Catalan, Spanish and US-English, which did not exist up to date in those mentioned repositories.

It was decided to take the Verbmobil corpus as starting point, given that it was based on recorded conversations for a semantically restricted domain. This has allowed us to focus on the appointment scheduling domain. At a later stage, further LR have been created from scratch (subset-2) focusing on some particular tourism-related subdomains.

## 2.2 Selected Material for Subset-1

Out of all the databases available from Verbmobil, we selected 9 of them that contained recordings only in US-English. The aim was to start from the US-English recordings and generate their counterparts in Catalan and Spanish by means of human translation. A total amount of 287,655 tokens has been collected, being 3,333 of them different.

## 2.3 Cleaning and Preservation Criteria

The Verbmobil<sup>9</sup> databases are annotated with a number of tags that mark different kinds of information about every utterance. These labels deal with speech phenomena at two different levels. At the word level these tags denote information related to the kind of token (i.e. punctuation, foreign word, interjection, proper noun, neologism, number, letter spelling, acronym) and to the token production inside the utterance (i.e. lengthening, unidentifiable or hard to identify word). At the utterance level these marks denote information regarding the utterance production as a whole (i.e. aborted articulations, articulatory and

technical interruptions, breathings, empty pauses, false starts, filled pauses, human noises, repetitions, corrections, noises, turn breaks).

Some of these tags were preserved because they were later to be needed during the translation task, in the sense that they could carry either some meaning or discourse information. This was the case for punctuation marks, filled pauses, foreign words, interjections, proper nouns, letter spelling, abbreviations, neologisms, technical interruptions, turn breaks, and unidentifiable or hard to identify words.

## 2.4 Translation of Subset-1

As already mentioned, Verbmobil's US-English dialogues in subset-1 focus on appointment scheduling and travel planning. One such dialogue looks as follows:

```
e001ach2_000_ANV_230000:
  hi ~Mary~, how are you.
e001ach1_001_SMA_230000:
  oh I am doing fine Mister ~Vandaloo~,
  how are you.
e001ach2_002_ANV_230000:
  pretty good, <uhm> I guess we need
  to figure out a day, today is the
  first of November, so within the next
  two months when we can make it to,
  ~Hanover~?
e001ach1_003_SMA_230000:
  that sounds like a good idea, let us
  see, I have, <uhm> the, nineteenth
  twentieth and twenty first I am
  available to travel.
```

Turn headers within the dialogues also contain useful information like *dialogue identifier*, *turn number* and *speaker*. Language markers have been added to each turn header so as to tag the turns for the three different languages. This simplifies any further search on the corpus.

Given the nature of the project, the translation task is perhaps the most critical issue since human translators are clearly the source of knowledge our system intends to acquire. Thus, a suitable translation methodology was designed to assist translators during the process of translation of both subset-1 and subset-2. Translators received an instruction manual. Bearing in mind that we want to train statistical SST systems, the main rule was:

“The target sentence should be as literal as possible in the sense of word-to-word translation and word order as long as it is a proper sentence in the target language.”

As a result of the translation process performed on subset-1, we have obtained a Spanish corpus

<sup>7</sup><http://www.elra.info>

<sup>8</sup><http://www ldc.upenn.edu>

<sup>9</sup>[http://www.is.cs.cmu.edu/trl\\_conventions/projects/verbmobil.html](http://www.is.cs.cmu.edu/trl_conventions/projects/verbmobil.html)

	Spanish	Catalan
<i>speech time</i>	31h:7m:32s	23h:43m:55s
<i>#speakers</i>	77	56
<i>#dialogues</i>	217	172
<i>#turns</i>	10,998	9,321
<i>#sentences</i>	24,372	19,113
<i>#tokens</i>	349,970	277,777
<i>#distinct</i>	11,714	10,057

Table 1: Oral Database

of 281,848 tokens (resulting in a vocabulary size of 5,149), and a Catalan corpus of 277,955 tokens (with a vocabulary size of 5,145).

An extra feature of this translation, and that of subset-2, is that the resulting resource is a trilingual corpus already aligned at sentence level. As mentioned in the introduction, this is highly valuable for statistical machine translation and other language technologies.

### 3 Speech Corpora Recorded

Subset-2 of our corpus has been created from scratch. Data come from the transcription of Catalan and Spanish spoken dialogues that focus on the tourist domain. Due to size restrictions, a subset of this domain was chosen: tourist-employee conversations. Further, four scenario categories were defined, namely *Hotel*, *Travel Agency*, *Tourist Information Office* and *Railway/Airline Company*. It was also an aim to collect data that would allow us to generate an appropriate language model of the given domain.

Some figures for both languages can be seen in Table 1. All dialogues have been manually transcribed, spell-checked, and then translated.

#### 3.1 Recording Procedure

Given the restrictions to record real telephone conversations (such as legal issues and lack of control over excessive non-domain input), but aiming to avoid non-verbal communication, we decided to record artificial telephone dialogues. This procedure is common in speech research projects.

A recording platform was set up. In the beginning it was designed to be used in two different modes, with overlapping and without. The second mode matches best the translation system, where a machine is between the speakers. In this rigid turn strategy speakers are not allowed to speak simultaneously. The first turn is given to

the speaker receiving the phone call. The speaker then indicates a turn exchange by pressing a key.

Volunteers were recruited in pairs, and placed in different rooms. Phone calls would last ten minutes in average.

#### 3.2 Pursued Information

The four scenarios mentioned in section 3 were further restricted by defining scenario situations in more detail:

- **Hotel** (Booking, Services, Leisure Activities)
- **Travel Agency** (Ticket and Hotel Reservation, Travel Planning, Reservation Changes, Cancellation, Modification)
- **Tourist Information Office** (about the City, Transportation, Accommodation, Gastronomy, Places to visit, Leisure, Shopping, Night Life)
- **Railway/Airline Company** (Ticket Reservation, Information, Services)

For the conversations to yield the pursued information some specific situations were designed. A series of templates containing descriptions, lists, tables and figures 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. This material was based on actual existing material (from hotels, travel agencies, the Internet, etcetera) and modified during recordings, mainly due to the suggestions of speakers. Below follows a simplified example of one such template:

*Hotel Reservation.* ‘*Speaker 0*’ acts as a tourist booking 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.

#### 3.3 Comments

A variety of problems arise when creating speech corpora in artificial situations and using *recruited speakers*. On one hand, domain and vocabulary should be controlled, and on the other, situations should be as realistic and natural as possible.

Moreover, speakers need to be motivated and not get nervous. Encouraging them to talk about either something they had experienced or something they were going soon to have to deal with, turned out to produce more realistic dialogues.

Likewise dialogue recording, manual transcriptions and translations are highly time-consuming tasks. A considerable number of people were involved in these tasks. In this situation, it is very

important to revise the generated transcriptions and translations. That involves not only spell checking but to guarantee tag, token, sentence, turn and dialogue consistency among the three languages for the whole corpus.

Along the process, it became clearer that the more restricted the domain the better the system would perform, the smaller the coverage would be though. Therefore, a subset of scenarios were selected to focus on, although never completely forgetting about the rest. Those scenarios were *accommodation booking*, *ticket reservation*, *travel planning*, and *asking for information*.

### 3.4 Transcription

During manual transcription, the information encoded was exactly the same than the one preserved for subset-1. It was agreed to use the Extensible Markup Language (XML). All relevant XML tags have been maintained during the human translation process. A guideline document was elaborated to help transcribers on their task.

### 3.5 Translation of Subset-2

Recordings are all in Catalan and Spanish. Thus, every utterance has been translated into English and either Spanish or Catalan, respectively, preserving the same translation style considered for subset-1. Recorded dialogues and their translations look as follows:

```
B0162_21_1_001_000_CAT:
  bon dia. agència de viatges "Sol i Mar".
  parla amb "Sebastià". com el puc ajudar?
B0162_21_1_001_000_SPA:
  buenos días. agencia de viajes "Sol i Mar".
  habla con "Sebastià". en qué puedo ayudarle?
B0162_21_1_001_000_ENG:
  good morning. "Sol i Mar" travel agency.
  "Sebastià" speaking. can I help you?
B0162_21_1_030_001_CAT:
  bon dia. voldria reservar un bitllet
  per "Milwaukee", "Estats Units".
B0162_21_1_030_001_SPA:
  buenos días. quería reservar un billete
  para "Milwaukee", "Estados Unidos".
B0162_21_1_030_001_ENG:
  good morning . I'd like to reserve a ticket
  for "Milwaukee" the "United States".
```

Turn headers contain again useful information: *dialogue identifier*, *scenario code*, *source language*, *speaker*, *turn number* and *turn language*.

## 4 Part-of-Speech Tagging

As already mentioned, good results have already been obtained in earlier attempts of statistical machine translation. However, all experts agree

on the fact that a crucial issue here is the type of data used for the training and knowledge learning. Thus, bearing in mind the needs of the expert researchers and developers with whom we are working closely, LC-STAR is developing more sophisticated and linguistically enriched LR. For that purpose, we have considered that providing part-of-speech (POS) information for the corpora developed was certainly the first step to take. So far, this annotation has been carried out for subset-1 of our corpus, but it is one aim of the project to enrich the whole trilingual corpus in the end. Moreover, the project also aims to establish the specifications for the design and development of SST-oriented LR. This is the reason why the corpus under development is already being tested, so as to evaluate its usefulness during construction and, thus, adjust its characteristics, according to requirements.

### 4.1 Morphosyntactic Annotation of Catalan and Spanish Data

The POS tagging of Catalan and Spanish data has been performed with our morphological analyser *MACO+* (*Morphological Analyzer Corpus Oriented*) (Carmona *et al.* 98), a robust and wide-coverage tool that accepts unrestricted text as input and provides all possible labels and lemmas for each word form.

The set of tags used to represent the morphological information is based on those proposed by EAGLES<sup>10</sup>. The output of *MACO+* is disambiguated by *RELAX* (*Relaxation Labelling Based Tagger*) (Padró 97). This tool selects (or aims to select) the correct POS and lemma for each word in the given context. Currently, it produces an output with over 97% accuracy.

Although both these tools are also available for English language, their performance is not as accurate as for Catalan and Spanish.

### 4.2 Morphosyntactic Annotation of English Data

After considering some tools available for English, it was decided to use Eric Brill's POS tagger (Brill 93). Its tags are based on the Penn Treebank project<sup>11</sup>. Brill's transformation-based error-driven learning tool was chosen because it is highly portable, it can be trained on small training tagged corpora and it is a simpler tool to use.

<sup>10</sup><http://www.ilc.cnr.it/EAGLES96/home.html>

<sup>11</sup><http://www.cis.upenn.edu/~treebank/home.html>

Before mentioning some of the problems encountered during our POS tagging process, it should just be added that as a first approach, Brill's tagger has been used as-is, so as to see what was capable of. As it will be seen in the following section, though, a few problems popped up that had been feared but not fully anticipated, and that had to be dealt with.

### 4.3 Problems Encountered during the POS Tagging Process

Conversations may contain some disfluencies such as false starts, corrections, repetitions, filled pauses, and certain ungrammaticalities. These phenomena cause the POS-tagger performance to significantly decrease. Unfortunately, the training data are very different from the data to be tagged, which is a major drawback in our research that must be overcome. If morphosyntactic information is to be used it should be taken into account that tagging errors are surely going to propagate into the rest of the system.

### 4.4 Usefulness of POS Tagging for SST

The trilingual corpus created represents a source of information on two different types of languages, those with little morphological inflection and those with a rather rich inflectional morphology. When pursuing SST, dealing with both such types can be very problematic. It seems reasonable, thus, to try and provide the system with the necessary information to achieve the correct translations.

Bearing this in mind, subset-1 of our corpus has been enriched with morphological information and its usefulness has already been tested. These initial tests (Ueffing & Ney 03) have shown that translation results have improved with the help of POS information. It is, thus, our immediate aim to provide POS information for subset-2.

## 5 Conclusions and Further Work

The paper has provided a description of on-going work towards the development of language resources for statistical machine translation, in the frame of the LC-STAR project, which aims at both creating corpora and lexica for SST and establishing criteria for future resource development.

As an initial stage, we have developed a trilingual aligned corpus for languages of very different morphological inflection. In order to improve

statistical SST for these languages, this corpus has been enriched with POS information. Initial experiments carried out by our project partners have proved the success of this approach. Thus, it is our objective to complete the task of enriching the corpus with POS information and study the addition of any other relevant linguistic knowledge.

As a further objective of the LC-STAR project, we also aim at developing lexica for SST. Criteria and contents for this are currently being considered and some experiments are being planned to test their usefulness. Last but not least, an important outcome of this work will be the specifications that should be later followed by other LR to be developed.

## References

- (Alshawi *et al.* 98) H. Alshawi, S. Bangalore, and S. Douglas. Automatic acquisition of hierarchical transduction models for machine translation. In *Proc. COLING-ACL'98: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 41–47, Montreal, Canada, August 1998.
- (Alvarez *et al.* 01) Jordi Alvarez, Victoria Arranz, Núria Castell, and Civit Montserrat. Linguistic and logical tools for an advanced interactive system in spanish. *Lecture Notes in Artificial Intelligence*, 2070:519–528, 2001.
- (Brill 93) Eric Brill. *A Corpus-Based Approach to Language Learning*. Unpublished PhD thesis, Department of Computer and Information Science, University of Pennsylvania, 1993.
- (Carmona *et al.* 98) J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An environment for morphosyntactic processing of unrestricted spanish text. In *Proc. 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, September 1998.
- (Ide *et al.* 02) Nancy Ide, Tomaz Erjavec, and Dan Tufis. Sense discrimination with parallel corpora. In *ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia, 2002.
- (Lavie *et al.* 01) A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazari, P. Coletti, L. Taddei, and F. Balducci. Architecture and design considerations in nespole!: a speech translation system for e-commerce applications. In *Proc. Human Language Technology Conference (HLT'01)*, San Diego, CA, March 2001.
- (Levin *et al.* 00) L. Levin, A. Lavie, M. Woszczyna, and Waibel A. The janus-iii translation system. *Machine Translation*, 15(1-2).
- (Lopez *et al.* 02) Adam Lopez, Michael Nossal, Rebecca Hwa, and Philip Resnik. Word-level alignment for multilingual resource acquisition. In *2nd International Conference on Language Resources and Evaluation (LREC'02)*, Spain, June 2002.
- (Ney *et al.* 01) Hermann Ney, Franz J. Och, and Stephen Vogel. The rwth system for statistical translation of spoken dialogues. In *Proc. Human Language Technology Conference (HLT'01)*, San Diego, CA, March 2001.
- (Padró 97) Lluís Padró. *A Hybrid Environment for Syntax-Semantic Tagging*. Unpublished PhD thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 1997.
- (Ueffing & Ney 03) Nicola Ueffing and Hermann Ney. Using pos information for statistical machine translation into morphologically rich languages. In *Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, Budapest, Hungary, April 2003.
- (Yarowsky & Ngai 01) David Yarowsky and Grace Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'01)*, Pittsburgh, USA, June 2001.