Bilingual Connections for Trilingual Corpora: An XML Approach

**Victoria Arranz, Núria Castell, Josep Maria Crego,
Jesús Giménez, Adrià de Gispert and Patrik Lambert**

TALP Research Center
Universitat Politècnica de Catalunya,
Jordi Girona Salgado, 1-3, 08034 Barcelona, Spain
{varranz, castell, jmcrego, jgimenez, agispert, lambert}@talp.upc.es

Abstract

This paper describes the design and development of a trilingual spontaneous speech corpus for statistical speech-to-speech translation. The languages considered are Catalan, Spanish and US-English. This corpus has been built bearing in mind the strong need for multilingual collections of on-line data within the area of statistical translation, as well as the need for data that are parallel or aligned, that contain different types of linguistic information and that can be used by different translation systems. For that reason, our aim has been the creation of a linguistically-enriched resource with an XML-based DTD that allows a useful, transparent and flexible storage of the data. Moreover, these resources are also valuable for a wide range of Natural Language Processing applications, such as multilingual resource acquisition or word sense discrimination, among others.

1. Introduction

A considerable effort has been devoted to the construction of language resources (LRs) in the past decade. These are extremely necessary for a large number of areas involved in language technology. The three International Conferences on Language Resources and Evaluation [1] that have taken place so far show the evolution and great interest in this area. Besides the need of multilingual resources like corpora for statistical speech-to-speech translation (SST), which is the area of interest in the current paper, these are also very valuable LRs for many other areas. Some of these areas are foreign language resource acquisition (Yarowsky and Ngai, 2001), multilingual resource acquisition (Lopez et al., 2002) or word sense discrimination (Ide et al., 2002).

With regard to statistical speech-to-speech translation, multilingual collections of on-line data are highly seeked for. Data used in the development of SST systems are usually multilingual, possibly aligned and sometimes generated from the annotation of speech recordings. However, these are particularly costly LRs that may need to handle the complexity of spontaneous speech when recording human conversations. Furthermore, the design of LRs needs to consider issues such as format, portability, what-to-include, etc.

Bearing these in mind, our aim is the creation of speech-to-speech centred and linguistically enriched trilingual LRs for Catalan, Spanish and US-English, which did not exist up to date. These resources are being developed within the frame of the LC-STAR [2] and ALIADO [3] projects. The trilingual corpus created is made up of two subsets, one based on already existing data and another one developed on the recordings carried out for that purpose. Further details on those subsets are provided below.

This paper describes the design and development of a trilingual spontaneous speech corpus, focusing on the importance of designing and creating transparent, useful, portable and robust resources. This has been achieved by designing a general XML-based DTD that is flexible enough to represent any type of corpus (mono- or multilingual; text or speech; parallel or aligned) in either broad- or limited-domains. This makes it particularly suitable for statistical SST oriented resources.

The paper is organised as follows: section 2. describes the expansion process carried out on some existing resource so as to build the first subset of our trilingual corpora. Section 3. explains the construction of the second subset of our trilingual corpora, which has been created from scratch. Section 4. proceeds onto the design issues of the DTD and the reasons to follow an XML approach. Section 5. elaborates on the usefulness of both this kind of resources and the use of XML for statistical SST. Finally, section 6. draws some conclusions and provides suggestions for future work.

2. Expanding an Existing Resource

In order to build the aimed resources, it was decided to combine the use of both existing and new data. That is, part of the new trilingual corpora would be developed based on some already existing resource (hereafter referred to as subset-1) and another part would be based on completely new data (hereafter referred to as subset-2). It was planned that the full trilingual data would add up to about 1,500k words.

As a starting point, data available from LR repositories such as ELRA [4] and LDC [5] were considered to create subset-1 of our trilingual corpus. Part of the recordings from the VerbMobil project were selected given that it was based on recorded conversations for a semantically restricted domain (the appointment scheduling domain). Our aim was to start from the US-English recordings and generate their counterparts in Catalan and Spanish by means of human translation.

Out of all the databases available from VerbMobil, 9 were selected that contained recordings in English. Initially, we aimed for US-English speakers' data, but then

[1] http://www.lrec-conf.org/

[2] http://www.lc-star.com/

[3] http://gps-tsc.upc.es/veu/aliado/

[4] http://www.elra.info

[5] http://www.ldc.upenn.edu

we also accepted Denglish (English spoken by Germans) so as to achieve a larger amount of data. A total amount of 287,655 tokens were collected from these databases, resulting in a vocabulay size of 3,333.

The VerbMobil corpus is annotated with a number of tags that mark different speech phenomena at word and utterance levels (e.g.: filled pauses, interjections, technical interruptions, turn breaks,...) as well as punctuation marks, proper names, foreign words, etc. Some of these tags were preserved to be later used during the translation task, in the sense that they could carry either some meaning or discourse information.

With regard to the translation of the data, several issues were taken into account that can be applied to the translation of both subset-1 and subset-2. These issues became a guideline or methodology for translation and can be summarised as follows: 1) As we want to train statistical SST systems, the main guideline for translators was to generate proper (correct and natural) sentences but translating as literally as possible. That implied preserving the source-language word order and syntax as much as possible as long as such literal translation neither changed the semantics nor added ambiguity. This type of translation produces a lower level of perplexity while still maintaining the competitiveness of those machine translation techniques which do not take advantage of any syntactic information. 2) When faced with ambiguity, human translators were asked to bear in mind the tourist-domain customer- employee- like dialogue scenario and their common sense and real world knowledge as the only additional arguments. 3) Some rules were also stated to handle difficult cases such as proper names, foreign words, neologisms and letter spellings. Generally, these were not tranlated. However, some exceptions were considered otherwise, such as *Christmas*, *AIDS* and *PM*. For a detailed description of the whole translation process, please refer to (Arranz et al., 2003) and (Arranz et al., 2004).

3. Creation of New Speech Corpora

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. The total figures for both languages can be seen in Table 1.

In order to avoid non-verbal communication, we decided to record the spoken dialogues through the telephone network: speakers were placed in different rooms and talked on the phone with each other. In order to do so, a phone-call recording platform was set up with a rigid turn strategy. No speaker overlapping was allowed given that this would not be well matched to the translation system, where a machine is between the speakers. Conversations between recruited volunteers would last ten minutes in average. Four scenarios were defined: Hotel, Travel Agency, Tourist Information Office and Railway/Airline Company. For the conversations to yield the pursued information some specific subscenarios were designed, such as *hotel booking* and *railway ticket reservation*. A series of templates were also built so as to assist speakers at conversation time. They provided speakers with draft descriptions of what to discuss in every subscenario.

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 1: Oral Database

A key factor that came out during recording time was the *speaker's motivation*. Situations were meant to be as realistic and natural as possible, but this turned out to be amazingly complicated due to the speakers' lack of interest as well as their getting nervous while being recorded. They were encouraged to talk about something closer to their own lives (booking their own summer holiday, etc.) so as to get more involved in this sort of role-play activity.

Likewise dialogue recording, manual transcriptions and translations were highly time-consuming tasks and numerous people were involved. It was thus very important to revise the generated material. That involved not only spell-checking but guaranteeing tag, token, sentence, turn and dialogue consistency among the three languages for the whole corpus.

Recordings were 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. Both subset-1 and subset-2 have been enriched with morphological information. Section 5. provides some details on the advantages of using POS information for statistical SST.

Finally, a DTD has been designed and created that focuses on the needs of data storage for statistical machine systems. It was agreed to use the Extensible Markup Language (XML) to represent all the encoded acoustic and linguistic phenomena in these corpora. The following sections explain the reasons behind the choice of XML, the DTD we have developed and its advantages for the storage of multilingual resources.

4. A General DTD

The Extensible Markup Language (XML) (Bray et al., 2000) has been succesfully applied to the creation of multilingual corpora. Many of the approaches are intended to comply, as far as possible, with the Text Enconding Initiative (TEI) guidelines for electronic text encoding and interchange (Sperberg-McQueen and L. Burnard (eds.), 2002). A well-known example is the Corpus Encoding Standard

```
<?xml version="1.0" encoding ="ISO-8859-1"?>
<<?xml version="1.0" encoding ="ISO-8859-1"?>
<!DOCTYPE lcstar SYSTEM "lcstar.dtd" []>
<DOC_REPOSITORY DATE="10/2/2004">
  <DOC NDOC="000">
    ...
    <SEC N="000" S="1">
      <SGM N="999">
        <LSGM LAN="CA">el dinou , d’ aquest mes ?</LSGM>
        <LSGM LAN="EN">nineteenth , of this
       month ?</LSGM>
        <LSGM LAN="ES">? el diecinueve , de
                       este mes ?</LSGM>
      </SGM>
    </SEC>
    ...
  </DOC>
</DOC_REPOSITORY>
```

Figure 1: A trilingual *section* aligned at *segment* level.

```
<SGM N="999">
<LSGM LAN="CA" S="1" W="7">
  <W L="el" P="DA0MS0">el</W>
  <W L="dinou" P="NCMS000">dinou</W>
  <W L="," P="Fc">,</W>
  <W L="de" P="SPS00">d’</W>
  <W L="aquest" P="DD0MS0">aquest</W>
  <W L="mes" P="NCMS000">mes</W>
  <W L="?" P="Fit">?</W></LSGM>
<LSGM LAN="EN" S="1" W="6">
  <W L="nineteen" P="JJ">nineteenth</W>
  <W L="," P=",">,</W>
  <W L="of" P="IN">of</W>
  <W L="this" P="DT">this</W>
  <W L="month" P="NN">month</W>
  <W L="?" P=".">?</W></LSGM>
<LSGM LAN="ES" S="1" W="8">
  <W L="?" P="Fia">?</W>
  <W L="el" P="DA0MS0">el</W>
  <W L="diecinueve" P="DN0CP0">diecinueve</W>
  <W L="," P="Fc">,</W>
  <W L="de" P="SPS00">de</W>
  <W L="este" P="DD0MS0">este</W>
  <W L="mes" P="NCMS000">mes</W>
  <W L="?" P="Fit">?</W></LSGM>
</SGM>
```

Figure 2: A trilingual *segment*.

for XML (XCES) (Ide et al., 2000), based on the Corpus Encoding Standard (CES) developed by the Expert Advisory Group on Language Engineering Standards (EAGLES). CES is an application of TEI, with a customization and some modifications appropriate to corpus-based language engineering. However, XCES is still under development and subject to change. The TEI system is more stable but it is difficult to use directly for speech corpora and the annotations are hardly readable for humans. Furthermore, a general purpose format like TEI can never be optimal for a specific application. With the aim of specifying a format suitable for state-of-the-art MT algorithms, we built a new XML-based document type definition (DTD). A customed format implies:

- Less redundant tags
- Elements having exactly the meaning we require
- The marked corpus is easy to read for its users

The use of XML implies a disk/memory space consumption overhead and an extra-effort to port data to XML format. However, it also involves important benefits:

- Human and machine readable documents
- Transparent access to data
- Usability and portability
- Flexibility and scalability
- Effectiveness and robustness

With our DTD, a corpus is a *document repository*, a collection of *documents*. A *document* is a general concept and can represent any kind of component of a corpus, such as a news item (in a news corpus) or a dialogue (in a dialogue corpus). Each *document* is divided into *sections*, such as a paragraph (in a news corpus) or a dialogue turn (in a dialogue corpus). Each *section* consists of one or more *segments* (that can be either a passage or a sentence). Each *segment* is divided into *lsegments* (one segment per language). This way, the alignment at segment level is guaranteed. This is the minimum for the multilingual data to be really valuable corpora.

A *segment* may either be simply a text string or consist of *words*, *compound-words*, *multi-words* and *acoustic-events*. Figure 1 is an example of the simplest structure for aligned segments, where segments are represented as text strings.

A *word* is a sequence of characters separated by blankspaces and it may carry linguistic features (lemma, part-of-speech, wordnet synset, is_a_neologism?, is_a_letter_spelling?, is_an_acronym?, is_a_foreign_word?). Moreover, a *word* may also have acoustic features (is_interrupted?, is_badly_pronounced?). A *compound-word* is a single word made up of several separate words, such as Catalan/Spanish verbal forms taking clitic pronouns (e.g.: *dá+me+lo* for “give + it + to me”). A *multi-word* may consist of several words grouped as a whole. They may form, for instance, named entities, dates, syntactic phrases, etcetera. Phenomena such as filled pauses, noises, repetitions or corrections, etc., are marked as *acoustic events*. Figure 2 is an example of a more complex structure, where segments are decomposed into words that carry morphological information (where **L** stands for lemma and **P** for POS[6]).

Furthermore, bilingual alignments may be defined among elements (words, compound words and multi-words) inside parallel *lsegments*. The alignment links will be stored only in one of the *lsegments*, although bi-directional redundancy is allowed.

5. Benefits for Statistical SST

Research on statistical speech-to-speech translation from a corpus-based approach always involves the necessary step of preparing training and test data, which is unfortunately quite complex due to the multilingual nature of the input data. Even if no further linguistic information is provided besides the words, the time and effort required to

[6] The use of different POS tags is due to the different tagging tools employed, MACO+(Carmona et al., 1998), for Catalan and Spanish, and Brill’s tagger(Brill, 1993), for English.

set this framework for developing translation models can be considerable, unless a unified tool is provided.

With regard to the use of XML, using a standard format for storing multilingual linguistic resources is a necessary step towards creating this efficient framework we are looking for in the development of linguistic-based statistical SST systems. By integrating a rich set of linguistic knowledge sources as well as bilingual alignment information, this XML format provides an easy and standard way of accessing the information needed to train and test statistical SST systems.

Regarding the use of POS information for statistical speech-to-speech translation, initial tests have proved their improvement of translation results (cf. (Ueffing and Ney., 2003) and (Toutanova et al., 2002), among others), and we can expect a more positive contribution of linguistic information (named entities, chunks, phrases, syntax, semantic tags,...) as further research is being carried out. The use of categorisation of time and date expressions, which is also included in the DTD here described can also provide a significative boost in translation performance (de Gispert and Mariño, 2003).

On the other hand, the availability of information on a word-level alignment allows us to design, test and compare automatic word alignment models in an efficient way.

To sum up, our DTD proposal provides a rich and flexible structure that is able to include this useful information, and that is specially designed to ease the incorporation of new information as soon as it may be required by progress on research. While allowing for the introduction of several information tags at a word level in an increasing fashion, it also accepts raw data when these are not in use, avoiding creating extra-large data files.

6. Conclusions

The paper has provided a description of on-going work towards the development of LRs for statistical SST, which aims at both creating corpora and lexica for SST and establishing criteria for future resource development. We have developed a trilingual aligned corpus for languages of very different morphological inflection (Catalan and Spanish versus US-English). This corpus has been enriched with POS information and its usefulness for statistical SST has been tested. An XML-based structure is proposed to represent either bilingual aligned corpora or multilingual paralell corpora. Further, this structure allows to store different types of data, such as text or speech, monolingual data, belonging to either a broad- or specific semantic domain. Finally, this structure also permits to store linguistic information as well as alignment links in a flexible and easy-to-use way. Last but not least, the combination of both this multilingual resource with the XML-based DTD designed makes this trilingual corpus a valuable contribution for the improvement of statistical speech-to-speech translation, as well as for research in other language technology areas.

Acknowledgements

This research has been done under the support of the European Commission (IST-2001-32216 project) and of the Spanish Ministry of Science and Technology (TIC2002-04447 project).

7. References

Arranz, Victoria, Núria Castell, and Jesús Giménez, 2003. Development of languages resources for speech-to-speech translation. In *RANLP'03*. Borovets, Bulgaria.

Arranz, Victoria, Núria Castell, and Jesús Giménez, 2004. Description of raw corpora. Technical Report Deliverable D5.3, LC-STAR project by the European Community (IST project ref. no. 2001-32216).

Bray, T., J. Paoli, C. M. Sperberg-McQueen, and E. Maler, ed., 2000. *Extensible Markup Language (XML) 1.0*. W3C, 2nd edition.

Brill, Eric, 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo, 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *Proc. 1st International Conference on Language Resources and Evaluation (LREC'98)*. Granada, Spain.

de Gispert, A. and José B. Mariño, 2003. Experiments in word-ordering and morphological preprocessing for transducer-based statistical machine translation. In *IEEE Automatic Speech and Understanding Workshop, ASRU'03*. St. Thomas, USA.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary, 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Ide, Nancy, Tomaz Erjavec, and Dan Tufis, 2002. Sense discrimination with parallel corpora. In *ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*. Philadelphia.

Lopez, Adam, Michael Nossal, Rebecca Hwa, and Philip Resnik, 2002. Word-level alignment for multilingual resource acquisition. In *2nd International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Spain.

Sperberg-McQueen, C. M. and L. Burnard (eds.), 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Toutanova, K., H. Tolga Ilhan, and C.D. Manning, 2002. Extensions to HMM-based statistical word alignment models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing, EMNLP'02*.

Ueffing, N. and H. Ney., 2003. Using POS information for statistical machine translation into morphologically rich languages. In *EACL*. Budapest, Hungary.

Yarowsky, David and Grace Ngai, 2001. 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)*. Carnegie Mellon University, Pittsburgh.