Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo)

Xavier Gómez Guinovart and Elena Sacau Fontenla

Seminario de Lingüística Informática
Universidade de Vigo
sli@uvigo.es

Abstract

In this paper, we present the methodology developed by the SLI (Computational Linguistics Group of the University of Vigo) for the building and processing of the CLUVI Corpus, showing the TMX-based XML specification designed to encode both morphosyntactic features and translation alignments in parallel corpora, and the solutions adopted for making the CLUVI parallel corpora freely available over the WWW (http://sli.uvigo.es/CLUVI/).

1. Introduction

The CLUVI (Linguistic Corpus of the University of Vigo) is an open textual corpus of specialized registers of contemporary oral and written Galician language. In its current state of development, the texts from the main section of the CLUVI belong to four specialized registers (from fiction, computing, journalism and legal-administrative fields) and three linguistic combinations related to Galician (monolingual Galician, Galician to Spanish translation and English to Galician translation), making up a corpus of four million words. The corpus is divided into four subcorpora of around one million words each: the TECTRA parallel corpus of English-Galician literary texts, the LEGA parallel corpus of Galician-Spanish legal-administrative texts, the XIGA monolingual corpus of texts about computing in Galician, and the MEGA monolingual Galician corpus of language from the media. The expansion of the CLUVI with Portuguese and French parallel texts and with English-Galician cinematographic parallel texts is currently in preparation.

In this paper, we will present the methodology developed by the SLI (Computational Linguistics Group of the University of Vigo) for the building and processing of the CLUVI Corpus, showing the solutions adopted for the encoding of the TECTRA and LEGA parallel corpora, and for making the CLUVI parallel corpora freely available over the WWW (http://sli.uvigo.es/CLUVI/).

The notation of the TECTRA and LEGA parallel corpora presents two different aspects: the morphosyntactic tagging and lexical lemmatization, and the encoding of the translation alignments. For the lemmatization and morphosyntactic tagging of the texts we employ the XML standard and the parts-of-speech standard tagsets proposed by EAGLES. For the tagging of the texts in Galician we use the parts-of-speech tagset elaborated by the SLI following the standard guidelines of EAGLES. The format chosen for storing the aligned parallel texts is the TMX format, as it is the XML encoding standard for translation memories and parallel corpora, regardless of the application used. We will present the solutions adopted in the CLUVI Corpus for the encoding of translation equivalences in TMX, when the correspondence between original and translation is not direct due to the omission, addition or reordering of phrases in the translation. We will also present the solution adopted for uniting the morphosyntactic information and the information about the translation equivalences in the TMX encoding, as well as the computational and linguistic techniques employed for the generation of corpus-based bilingual dictionaries.

Finally, we will present the Web application designed by the SLI for the searching and browsing of the CLUVI parallel corpora. This utility, freely available via the SLI website (http://webs.uvigo.es/sli/), permits the study of bilingual equivalences in real texts with academic purposes of research and teaching, and is equally suitable as a translation aid.

2. Tagging Alignments

The basic segmentation unit for the alignment of the CLUVI bitexts is the orthographic sentence of the source text. Therefore, the correspondence between source and target text will always be of the 1:n type. Frequently one sentence of the source text corresponds with one sentence of the translation (1:1). Nevertheless, there are cases in which a source sentence is not translated (1:0), or in which a source sentence corresponds with half a sentence (1:1/2) or with two sentences of the translation (1:2), or even in which a sentence of the translation does not correspond with any source sentence (0:1). Moreover, translating sometimes implies movements of sentences, or movements of source fragments from their original sentences to other sentences in translation. These movements are reordered in the target section of CLUVI parallel corpora to fit with the 1:n alignment criterion that preserves the integrity and the order of the translation units of the source text. This criterion is crucial when applied to the processing of multilingual corpora, where source sentences must permit to establish correspondences among equivalent sentences in various languages.

The TMX specification does not consider the encoding of these aspects of translations, because it has been designed for the storage and exchange of translation memories, and not for the representation of equivalent segments in parallel corpora. The TMX-based CLUVI encoding system uses an adapted version of some tags which are part of the TMX 1.4 specification (Savourel, 2002) in order to represent the not-1:1 correspondences and reorderings encoded in the CLUVI parallel corpora. The aspects of translation encoded in the CLUVI corpora can be described as either omission, addition or reordering, and will be tagged using an adapted version of TMX 1.4 content elements `<hi>` and `<ph>`.

2.1. Omission

There is an omission when a piece of the source text does not correspond with any piece of the target text, that is, when a sentence or part of a sentence is not translated. Omission is encoded in the CLUVI parallel corpora by means of the <hi> element. According to the TMX 1.4 specification, the <hi> (or highlight) element “delimits a section of text that has special meaning, such as a terminological unit, a proper name, an item that should not be modified, etc.” (Savourel, 2002). In the TMX-based CLUVI encoding, the <hi> element marks in the source text the fragment that is omitted in the target text. This use of the <hi> tag is noted by means of the type attribute with the "supr" value. For instance, the following English-Galician aligned sentences would be encoded as shown below:

‘Hello’, I said. [English]
-Ola. [Galician]

```
<tu>
<tuv xml:lang=″en″>
<seg>′Hello′,<hi type=″supr″>I said.</hi></seg>
</tuv>
<tuv xml:lang=″gl″>
<seg>-Ola.</seg>
</tuv>
</tu>
```

2.2. Addition

The translation process of addition implies the insertion of target text fragments without a correspondence in the source text. Addition is also encoded in the CLUVI by means of the <hi> element, so that it highlights the inserted fragment in the target text. This use of the <hi> tag is indicated by means of the type attribute with the "incl" value. The added fragment joins the translation unit into which it is inserted. If the new fragment is a sentence (or a sequence of sentences), then it joins either the preceding or the following translation unit, according to its context, thus respecting the 1:n alignment criterion. For instance, the following alignment would be encoded in this way:

‘Hello.’
-Ola - dixen.

```
<tu>
<tuv xml:lang=″en″>
<seg>′Hello.′</seg>
</tuv>
<tuv xml:lang=″gl″>
<seg>-Ola <hi type=″incl″>- dixen.</hi>
</tuv>
</tu>
```

2.3. Reordering

The reordering in translation implies movements of sentences, or movements of source fragments from their original sentences to other sentences in translation. These movements are reordered in the target section of CLUVI parallel corpora to fit with the 1:n alignment criterion that preserves the integrity and the order of the translation units of the source text. Reordering is encoded in the CLUVI by means of a combination of the <hi> element and the <ph> element. The phrase or sentence moved is tagged with a <hi> element, with a type attribute with the "reord" value, as well as with an x attribute with a numeric value acting as an unambiguous index. Moreover, the place in the texts from where the segment was moved is indicated by means of a <ph> element. According to the TMX 1.4 specification, the <ph> (or placeholder) element is used “to delimit a sequence of native standalone codes in the segment. Standalone codes are codes that are not opening or closing of a pair, for example empty elements in XML” (Savourel, 2002). In the TMX-based CLUVI encoding, the adapted <ph> element marks the departure point of the movement, and the relationship between the element moved and its place of origin is encoded in the <ph> element by means of an x attribute that shares its value with the index encoded in the <hi> element of the segment moved. Obviously, the tag in the place of origin is always an empty tag. As a tagging criterion for the CLUVI encoding, in order to avoid inconsistencies between different encoders, reordering segments will always be moved up. As a consequence, in the CLUVI there is no sequence like <ph x=″n″/> [...] <hi type=″reord″ x=″n″>Reordered element</hi>; instead, they are all like <hi type=″reord″ x=″n″>Reordered element</hi> [...] <ph x=″n″/>. Here is a simple example of reordering codification:

‘The front door!’ she said in this loud whisper. ‘It’s them!’
-A porta de fóra. ¡Son eles! - murmurou bastante alto.

```
<tu>
<tuv xml:lang=″en″>
<seg>′The front door!′ she said in this loud
whisper.</seg>
</tuv>
<tuv xml:lang=″gl″>
<seg>-A porta de fóra.<hi type=″reord″ x=″1″>-
murmurou bastante alto.</hi></seg>
</tuv>
</tu>
<tu>
<tuv xml:lang=″en″>
<seg>It′s them.</seg>
</tuv>
<tuv xml:lang=″gl″>
<seg>¡Son eles!<ph x=″1″/></seg>
</tuv>
</tu>
```

Additional reorderings would be marked with the <x=″2″>, <x=″3″>, ..., <x=″n″> attributes, as is shown in the following example:

‘Leave him alone, hey’ Sunny said. ‘C’mon, hey. We got the dough he owes us. Let’s go.’
-Déixao. Imos logo. Xa témo-lo que nos debe - dicía Sunny.

```
<tu>
<tuv xml:lang=″en″>
<seg>′Leave him alone, hey′ Sunny said.</seg>
</tuv>
<tuv xml:lang=″gl″>
```

```
<seg>-Déixao. <hi type=”reord” x=”1”>- dicía
Sunny.</hi></seg>
</tuv>
</tu>
<tu>
<tuv xml:lang=”en”>
<seg><hi type=”supr”>’C’mon, hey.</hi></seg>
</tuv>
<tu xml:lang=”gl”>
<seg></seg>
</tuv>
</tu>
<tu>
<tuv xml:lang=”en”>
<seg>We got the dough he owes us.</seg>
</tuv>
<tuv xml:lang=”gl>
<seg><hi type=”reord” x=”2”>Xa témo-lo que nos
debe<ph x=”1”/> </hi></seg>
</tuv>
</tu>
<tu>
<tuv xml:lang=”en”>
<seg>Let’s go.’</seg>
</tuv>
<tuv xml:lang=”gl”>
<seg>Imos logo.<ph x=”2”/></seg>
</tuv>
</tu>
```

3. Parallel PoS Tagging

PoS tagging of the CLUVI parallel corpora is encoded in XML using the PoS tagset for Galician elaborated by the SLI (Aguirre et al., 2003) following the EAGLES guidelines (Leech and Wilson 1996; Monachini and Calzolari 1996). The probabilistic system of tagging and disambiguation developed by the SLI and Imaxin Software uses a Galician computational lexicon which contains the PoS specifications defined by the SLI tagset. The English section of the CLUVI parallel corpora is tagged with the Trigram's Tags (TnT) tagger (Brants, 2000), as done in the IJS-ELAN parallel corpus (Erjavec, 2002).

The SLI tagset for Galician follows the EAGLES general guidelines about the grammatical categories and morphosyntactic features which should be distinguished. In its design we have strictly applied the EAGLES set of categories and attributes-and-values scheme recommended by Leech and Wilson (1996), adapting it to Galician as has been done in other languages like Italian or German (Teufel, 1996).

Another essential aspect in the design of the SLI Galician tagset is its correspondence with the EAGLES intermediate tagset. The intermediate tagset is a linguistically neutral representation which describes the linguistic features (attribute-value pairs) included in a tagset, which helps to trace relationships between different tagsets (Leech and Wilson, 1996). Thanks to the intermediate tagset, we can tag a Galician text with a tagset defined according to the grammatical terminology of Galician, and an English text with a tagset from the tradition of English corpus linguistics, and to finally turn both of them into the EAGLES standard intermediate tagset. In this way it is feasible to observe the unambiguous correspondences between the grammatical information of Galician and English texts in the English-Galician section of the CLUVI parallel corpora. These correspondences can be exploited later in the extraction of bilingual lexical information, both contextual and phraseological. In a broader sense, the correspondence between the Galician tagset and the intermediate tagset makes it possible to re-use the tagged texts in applications adapted to the EAGLES standard.

4. The CLUVI Specification

The XML specification for the CLUVI parallel corpora is strongly based on the TMX 1.4 specification. The main difference is that the CLUVI specification includes the morphosyntactic information in a <ling> element absent from the TMX specification. This <ling> element tags all the words and punctuation marks of the <seg> elements of the TMX original structure, and contains a <mor> element to encode the added morphosyntactic information, and an <ort> element to enclose the orthographic form of lexical tokens. This is in essence the document type definition for the CLUVI parallel corpora:

```
<!-- CLUVI_TMX DTD -->
<!ELEMENT cluvi_tmx (header, body) >
<!ATTLIST cluvi_tmx
        version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ATTLIST header
        creationtool CDATA #REQUIRED
        creationtoolversion CDATA #REQUIRED
        segtype (block|paragraph|sentence|phrase)
#REQUIRED
        o-tmf CDATA #REQUIRED
        adminlang CDATA #REQUIRED
        srclang CDATA #REQUIRED
        datatype  CDATA #REQUIRED >
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
        xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | ph | hi | ling)*>
<!ELEMENT hi (#PCDATA | ling)*>
<!ATTLIST hi
        type CDATA #IMPLIED
        x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
        x CDATA #IMPLIED>
<!ELEMENT ling (mor, ort)>
<!ELEMENT mor EMPTY>
<!ATTLIST mor
        cat (ARDFP|ARDFS|...) #REQUIRED
        lema CDATA #REQUIRED
        lema2 CDATA #IMPLIED>
<!ELEMENT ort (#PCDATA)>
```

Finally, this is a fragment of the CLUVI parallel corpora where a <tuv> element PoS tagged and aligned is shown:

```
<tu> <tuv xml:lang="en"> <seg>In the town they
tell the story of the great pearl.</seg> </tuv>
```

```
<tuv xml:lang="gl"> <seg><ling> <mor lema="en"
lema2="o" cat="PREP_ARDFS"/> <ort>Na</ort>
</ling> <ling> <mor lema="cidade" cat="NCFS"/>
<ort>cidade</ort> </ling> <ling> <mor
lema="contar" lema2="se" cat="VIPRS3_PPS3AR"/>
<ort>cóntase</ort> </ling> <ling> <mor lema="o"
cat="ARDFS"/> <ort>a</ort> </ling> <ling> <mor
lema="historia" cat="NCFS"/> <ort>historia</ort>
</ling> <ling> <mor lema="de" lema2="o"
cat="PREP_ARDFS"/> <ort>da</ort> </ling> <ling>
<mor lema="grande" cat="AXAPFS"/> <ort>gran</ort>
</ling> <ling> <mor lema="perla" cat="NCFS"/>
<ort>perla</ort> </ling> <ling> <mor lema="."
cat="PUNTO"/> <ort>.</ort> </ling> </seg> </tuv>
</tu>
```

5. Browsing the CLUVI in the Internet

Since September 2003, the SLI offers the possibility of searching and browsing the CLUVI parallel corpora in the Web at the URL http://sli.uvigo.es/CLUVI/. The parallel corpora managed by the SLI Web application are stored in the XML CLUVI specification, whereas the searching and browsing tool designed in PHP by the SLI was specifically created to carry out bilingual searches in tagged texts conformant to the TMX specification (including CLUVI format). This PHP application permits both simple and very complex searches of isolated words or sequences of words, and shows the bilingual equivalences of the terms in context, as found in real and referenced translations. The terms searched can correspond to either of the two languages of the translation, but it is also possible to carry out true bilingual searches, that is, to simultaneously search one term from each of the two languages of translation.

The number of aligned works and language pairs available in the website increases regularly, since the CLUVI is a academic research project in progress and with great vitality. At the moment, the CLUVI Parallel Corpus webpage permits to search two major corpora —TECTRA and LEGA (of around one million words each)—, as well as other minor parallel corpora of the following language pairs: English-Portuguese, Portuguese-Spanish, English-Spanish and French-Galician. It should be pointed out that the CLUVI interface also permits to browse the Legebiduna Corpus of Basque-Spanish administrative texts developed at the U. of Deusto (Abaitua et al., 1997).

6. Conclusions and further research

At the moment, our research is focused on the lexical extraction of an English-Galician bilingual dictionary from the TECTRA parallel corpus, using the NATools word-alignment software (Simões and Almeida 2003) based on the Twente aligner (Hiemstra, 1998). We are also carrying out some experiments to improve the accuracy of the bilingual extraction results. We are testing the exploitation of the morphosyntactic markup of the CLUVI parallel corpus to solve problems of ambiguity in extracted lexical pairs. Besides, we are exploring the possibility of using a "clean" version of the CLUVI parallel corpus to facilitate bilingual extraction, for instance by means of removing certain elements as stopwords, or those items which are marked in the corpus as omissions and additions. Moreover, we are trying to establish an automated system to filter the results on the basis of their reliability, using such criteria as the number of ocurrences and probability rates (Vintar, 2001).

Other short-term research projects based on the CLUVI Corpus will deal with the automatical extraction of multi-word terms, with the processing of multilingual TMX of more than two languages, and with the use of the CLUVI Corpus as a resource in a CAT distributed environment of translation memories (Simões et al., 2004).

With this work we try to contribute to the progress of the research and development in the fields of corpus linguistics and linguistic technologies for the Galician language.

Acknowledgements

This work has been funded by the Xunta de Galicia, within the project "Study and acquisition of basic resources of computational linguistics of Galician for the elaboration and improvement of computer aplications of linguistic technologies" (ref. PGIDT01TICC06E); and also by the Ministerio de Ciencia y Tecnología (MCYT), the Fondo Europeo de Desenvolvemento Rexional (FEDER), the Xunta de Galicia and the Universidade de Vigo within the project "Linguistic-computational processing of the Linguistic Corpus of the University of Vigo (CLUVI)" (ref. BFF2002-01385).

References

Abaitua, J., Casillas, A., and Martínez, R. (1997). Segmentación de corpus paralelos para memorias de traducción. Procesamiento del Lenguaje Natural, 21, pp. 17-30.

Aguirre, J.L, Álvarez Lugrís, A., and Gómez Guinovart, X. (2003). Aplicación do etiquetario morfosintáctico do SLI ó corpus de traduccións TECTRA. Viceversa, 7-8, pp. 189-212.

Brants, T. (2000). TnT: A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000.

Erjavec, T. (2002). Compiling and Using the IJS-ELAN Parallel Corpus. Informatica, 26, pp. 299-307.

Hiemstra, D. (1998). Multilingual Domain Modeling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. In Proceedings of the 8th CLIN Meeting (pp. 41-58).

Leech, G. and Wilson, A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Guidelines.

Monachini, M. and Calzolari, N. (1996). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. EAGLES Guidelines.

Savourel, Y. (2002) TMX 1.4a Specification. Technical Report. Localisation Industry Standards Association.

Simões, A.M. and Almeida, J.J. (2003). NATools: A Statistical Word Aligner Workbench. Procesamiento del Lenguaje Natural, 31, pp. 217-224.

Simões, A.M., Almeida, J.J. and Gómez Guinovart, X. (2004). Memórias de tradução distribuídas. In Proceedings of XATA-2004: XML, Aplicações e Tecnologias Associadas (pp. 59-68). Porto: U. of Porto.

Teufel, S. (1996). ELM-DE: EAGLES Specifications for German Morphosyntax. EAGLES Guidelines.

Vintar, Š. (2001). Using parallel corpora for translation-oriented term extraction. Babel Journal, 47(2), pp. 121-132.