

# Enriching parallel corpora with multimedia and lexical semantics

From the CLUVI Corpus to WordNet and SemCor

Xavier Gómez Guinovart

University of Vigo

In this chapter, I present the main characteristics of the CLUVI Corpus, an open collection of sentence-level aligned parallel corpora with over 44 million words in nine specialised domains (fiction, computing, popular science, biblical texts, law, consumer information, economy, tourism, and film subtitling) and different language combinations including Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque and Latin. Then, I present the methodology developed for extending the film subtitles section of the CLUVI Corpus with multimedia data. Finally, I discuss the resources and methods used to build the SensoGal Corpus, a SemCor-based English-Galician parallel corpus semantically annotated based on WordNet and aligned at the sentence and word levels.

**Keywords:** parallel corpora, multimedia, lexical semantics, WordNet, SemCor

## 1. Introduction

Parallel corpora are digitized collections of texts stored in both their original and translated version. Although it is possible to incorporate any type of linguistic information into parallel corpora, in general, parallel corpora are usually annotated with information on the equivalences of translation between the segments (words, phrases, sentences or other textual units). This process is called alignment, and these enriched parallel corpora are called aligned parallel corpora (Véronis 2000). Aligned parallel corpora have numerous applications in multilingual natural language processing, in the fields of machine translation (Koehn 2005), translation memories (Keshtkar & Mosavi Miangah 2012), lexical and terminological extraction (Tufiş 2007), second language teaching (Montero Perez et al. 2014), and contrastive linguistics (Santos 2004), among others. For a recent review of the state

of the art of this field, including an up-to-date survey of available parallel corpora, see the work of Mikhailov & Cooper (2016).

This chapter provides an overview of the research on parallel corpora carried out for more than a decade by the Seminario de Lingüística Informática (SLI) at the Universidade de Vigo in Galicia, Spain, with a focus on the development and exploitation of the CLUVI Corpus and on its multimedia extensions. I will also consider the most recent research on how to incorporate the advances into the design and study of computational lexical semantics to parallel corpora.<sup>1</sup>

## 2. The CLUVI Corpus

### 2.1 Corpus description

The CLUVI Corpus<sup>2</sup> is an open collection of human-annotated sentence-level aligned parallel corpora originally designed to cover specific areas of the contemporary Galician language in relation to other languages. With over 44 million words, the CLUVI collection currently comprises twenty parallel corpora in nine specialized registers or domains (fiction, computing, popular science, biblical texts, law, consumer information, economy, tourism, and film subtitling) and different language combinations with Galician, Spanish, English, French, Portuguese, Catalan, Italian, Basque, German and Latin. The coverage and size of the CLUVI Corpus are shown in Table 1, where the different subcorpora in the CLUVI collection are listed along with data about their current status.

At this moment, the CLUVI is the parallel corpus that contains the greatest number of translations from/to the Galician language and the widest thematic collection. Galician texts in CLUVI sum up about 11,000,000 words, which means a quarter of the total of the tokens in the corpus for the 10 languages gathered, and are representative of 9 types of specialized translation: legal, biblical, scientific-technical, literary, software localization, consumer information, film subtitling, economy and tourism.

---

1. This research has been carried out in the framework of the TUNER project (TIN2015-65308-C5-1-R) supported by the Ministry of Economy and Competitiveness of the Spanish Government and the European Fund for Regional Development (MINECO/FEDER).

2. International Standard Language Resource Number (ISLRN): 910-993-402-072-9.

**Table 1.** Coverage and size of the CLUVI Corpus

Language combinations and domains	Words
LEGA: Corpus of Galician–Spanish legal texts	6,582,415
BIBLOGAL: Corpus of Latin–Galician–Brazilian Portuguese–European Port–Catalan–Italian–Spanish–English biblical texts	5,489,607
UNESCO: Corpus of English–Galician–French–Spanish scientific-technical divulgation	3,724,620
LOGALIZA: Corpus of English–Galician software localization	3,706,242
CONSUMER: Corpus of Spanish–Galician–Catalan–Basque consumer information	5,586,431
TECTRA: Corpus of English–Galician literary texts	2,465,154
FEGA: Corpus of French–Galician literary texts	1,863,959
DEGA: Corpus of German–Galician literary texts	366,038
GALEA: Corpus of Galician–Spanish literary texts	162,795
PEGA: Corpus of Portuguese–Galician literary texts	68,431
EGAL: Corpus of Galician–Spanish economy texts	718,642
TURIGAL: Corpus of Spanish–Galician tourism texts	325,389
VEIGA: Corpus of English–Galician film subtitling	294,714
LEGE-BI: Corpus of Basque–Spanish legal texts	2,384,053
LOGALIZA: Corpus of English–Spanish software localization	4,992,133
TECTRA: Corpus of English–Spanish literary texts	2,108,141
PALOP: Corpus of Portuguese–Spanish postcolonial literature	566,590
TURIGAL: Corpus of Portuguese–English tourism texts	1,285,764
TECTRA: Corpus of English–Portuguese literary texts	875,595
SCOPE: Corpus of English–Spanish economy texts	1,151,544
<b>CLUVI Corpus total size:</b>	<b>44,163,783</b>

Other parallel corpus that currently facilitate access to translations to Galician are the Opus Corpus<sup>3</sup> (Tiedemann 2012) and the Per-Fide Corpus<sup>4</sup> (Almeida et al. 2014). On the one hand, the Opus collection provides Galician translated texts, mainly from English, taken from the web and automatically aligned at sentence level; it has an extension of around 7,600,000 tokens in Galician, and is taken primarily from the localization of the Linux operating system environment (Gnome, KDE4 and Ubuntu). On the other hand, the Per-Fide collection includes

3. Available at <<http://opus.lingfil.uu.se/>>.

4. Available at <<http://per-fide.di.uminho.pt/>>.

Portuguese–Galician software localization parallel texts derived from the Opus with about 400,000 tokens in Galician.

The format chosen for storing the aligned parallel texts in CLUVI is an adaptation of the TMX format (Savourel 2005), as this is the XML encoding standard for translation memories, regardless of the application used. A translation memory is a database that collects and records source text segments and their corresponding translated versions with the purpose of being reused for further translations via a computer-aided translation system. Albeit with some differences, an aligned parallel corpus is equivalent to a translation memory. Indeed, the last few years have seen an increasing number of TMX-encoded aligned parallel corpora, which offer the additional advantage that they can be used as translation memories for feeding computer-aided translation programs (Simões/Gómez Guinovart & Almeida 2004).

Since 2003, the SLI offers the possibility of searching and browsing the CLUVI parallel corpora online.<sup>5</sup> The parallel corpora managed by the web application are stored in the XML CLUVI specification, whereas the searching and browsing tool designed in PHP was specifically created to carry out bilingual searches in tagged texts that are conformant to this specification (Gómez Guinovart & Sacau Fontenla 2004b). This search application allows for very complex searches of isolated words or sequences of words, and shows the bilingual equivalences of the terms in context, as they appear in real and referenced translations. Due to copyright issues, it returns a maximum of 1,500 hits only, in order not to exceed the limits of the right to quote. Users can search for terms in either language of the corpus, although it is also possible to carry out true bilingual searches, that is, to search for bilingual segments that simultaneously contain a term in the source language and another term in the target language. Search results are displayed in a parallel fashion as a list of translation units. In addition, the LEGA Corpus from CLUVI can be downloaded via MetaShare<sup>6</sup> with CC BY-NC-SA 3.0 license.

## 2.2 Tagging the CLUVI Corpus

The basic segmentation unit for the alignment of the CLUVI parallel texts 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

---

5. The CLUVI online is available at <<http://sli.uvigo.gal/CLUVI/>>.

6. Available at <<http://hdl.handle.net/10230/20051>>.



source sentence corresponds with half a sentence (1: 1/2) or with two or more sentences of the translation (1: 2, 1: 3...), 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 stylistic 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 2005) in order to represent the non-1: 1 correspondences and reorderings encoded in the CLUVI parallel corpora. All these stylistic aspects of the corpus can be annotated according to the TMX-based CLUVI Corpus XML specification for parallel corpora which is summarized in Figure 1.

```
<!-- CLUVI_TMX DTD -->
<!ELEMENT cluvi_tmx (header, body) >
<!ATTLIST cluvi_tmx
    version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
    xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | hi | ph)*>
<!ELEMENT hi (#PCDATA)>
<!ATTLIST hi
    type CDATA #IMPLIED
    x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
    x CDATA #IMPLIED>
```

Figure 1. TMX-based CLUVI Corpus XML specification

The stylistic aspects of translation encoded in the CLUVI corpora can be described as either omissions, additions or reorderings, and will be tagged using an adapted version of the TMX 1.4 content elements *<hi>* (or highlight) and *<ph>* (or placeholder). An omission occurs when an item of the source text does not correspond

with any item of the target text, that is, when a sentence or part of a sentence is not translated. Omissions in the CLUVI parallel corpora are encoded by means of the *<hi>* element. According to the TMX 1.4 specification, the *<hi>* 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. It can be used for various processing tasks” (Savourel 2005). In the TMX-based CLUVI encoding, the *<hi>* element marks the piece in the source text that is omitted in the target text. This use of the *<hi>* tag is noted by means of the type attribute with the “*supr*” (deleted) value. For instance, the English–Galician aligned sentences in (1) would be encoded as the translation unit in (2).

- (1) ‘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>

On the other hand, the translation technique known as addition involves the insertion of elements in the target text that have no correspondence in the source text. Addition is also encoded in the CLUVI by means of the *<hi>* element, which highlights the inserted unit in the target text. This use of the *<hi>* tag is indicated by means of the type attribute with the “*incl*” (included) value. The added text joins the translation unit into which it is inserted. If the new element is a sentence (or a sequence of sentences), it joins either the preceding or the following translation unit, depending on its context, thus respecting the 1:n alignment criterion. For instance, the alignment in (3) would be encoded as (4).

- (2) ‘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>

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 text 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 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>* 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 2005). 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. The example in (5–6) shows how reorderings are encoded, and Figure 2 illustrates how the alignments are displayed by the web application when actually searching the corpus.

<u>VIX</u> (5781)	'The front door!' she said in this loud whisper.	-A porta de fóra. [[hi type='reord' x='16']] -murmurou bastante alto. [[/hi]]
<u>VIX</u> (5782)	'It's them!'	¡Son eles! [[ph x='16'/]]

Figure 2. Reordering in CLUVI

- (3) 'The front door!' she said in this loud whisper. 'It's them!' / -A porta de fóra. ¡Son eles! – murmurou bastante alto.
- (4) <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 = "16" > – 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 = "16"/></seg> </tuv> </tu>

### 2.3 Extending the CLUVI Corpus with multimedia data

Multimedia parallel corpora are a relatively scarce resource in the field of language technologies, due to the problem of obtaining translated multimedia materials, the difficulties of transcription and the technical complications of their processing. In turn, their use allows us to analyze aspects of multimedia translation that would be impossible to study from a merely textual perspective. In this section, I will present the methodology developed by the SLI for extending the CLUVI Corpus with multimedia data, focusing on the building of a multimedia extension of the VEIGA Corpus (Sotelo Dios & Gómez Guinovart 2012). The VEIGA Corpus is an English–Galician corpus consisting of 24 American, British, and Australian films subtitled in both English (intralingual subtitling) and Galician (interlingual subtitling) for DVD, cinema and Internet distribution. Developed under the broader framework of the CLUVI Corpus, VEIGA was born as a text-only corpus of subtitles. It was not until recently that we decided to make it multimedia, as soon as we found the appropriate tools to process the data and to make it accessible to the

public in what we considered to be an appropriate way. The VEIGA multimedia corpus of subtitles is currently available for public consultation at the CLUVI site.<sup>7</sup> However, it should be noted that only 13 of the 24 films are available in multimedia format at the time of writing.

The CLUVI Corpus functions as a repository of parallel corpora of different sizes and thematic fields, all of which undergo identical compiling and processing routines, and can be similarly accessed from one single search interface. Nonetheless, the VEIGA Corpus requires further processing in comparison to the other CLUVI corpora. Besides annotating stylistic aspects of translation such as omissions, additions and reordering of translation units, all the subtitles include both the in-cue and out-cue time (the time markers indicating where each subtitle should appear and disappear) and the line break indicator, allowing users to examine aspects which are inherent to subtitling practice, for example time and space constraints, segmentation, and condensation, among other specificities. In addition to this, the multimedia version of VEIGA enables users to stream the video clips corresponding to the bilingual pairs found in the search results, thus giving them access to the (co-)text in its original, multi-semiotic form. This means that wherever there is a result that matches the query in text format, the search interface shows a link to the corresponding video clips subtitled in each of the two languages involved (English and Galician) as shown in Figures 3–5.

All the above mentioned aspects of the VEIGA Corpus are annotated according an extended CLUVI XML specification, which is summarized in Figure 6.

Tagging the VEIGA Corpus at the textual/audiovisual interface level implies, on the one hand, tagging the correspondences between the English subtitles stored as XML textual data in the extended CLUVI XML specification and the equivalent segment of the original English-language film with English subtitles, and, on the other hand, tagging the correspondences between the Galician subtitles stored as XML textual data and the equivalent segment of the original English-language film with Galician subtitles. In order to be able to establish these textual/audiovisual correspondences, all of the VEIGA English-language films have been cut into video clips, each one corresponding to a subtitle. A first step is to check if the subtitles are in sync with the movie. In some cases, mostly when the subtitle file and the movie come from different sources, we need to edit the subtitles and add a time delay (forward or backward) so that their speed matches that of the video. Secondly, we embed the subtitles in the two languages in the original film. And finally, we edit each film subtitled both in English and in Galician and segment it into subtitles.

---

7. <<http://sli.uvigo.gal/CLUVI/vmm.html>>




335- <a href="#">CHU</a> (572)	You say <b>you'll</b> do anything for me, and this is what I get.	E ti dicías que <b>farías</b> calquera cousa por min...	
336- <a href="#">CHU</a> (573)	<b>You</b> don't think I had anything to do with that explosion?	¿Non crerás que teño algo que ver con esa explosión?	
337- <a href="#">CHU</a> (574)	<b>Didn't you?</b>	¿E logo non?	

Figure 3. Search results in VEIGA

**English subtitles**

Previous clip  
HTTP streaming  
English subtitles

**Current clip**  
HTTP streaming  
English subtitles

Next clip  
HTTP streaming  
English subtitles

Figure 4. English subtitles in VEIGA

**Galician subtitles**

Previous clip  
HTTP streaming  
Galician subtitles

**Current clip**  
HTTP streaming  
Galician subtitles

Next clip  
HTTP streaming  
Galician subtitles

Figure 5. Galician subtitles in VEIGA

Therefore, we come up with two subsets of subtitled video clips, one in English and the other in Galician, each made up of as many videos as subtitles in the corresponding film. Moreover, given that a high number of subtitles are not long enough to be played and watched properly (they are only one or two seconds long), each individual clip/subtitle is allotted ten extra seconds – five seconds before the subtitle shows up, and five seconds after it fades out –, thus providing the viewer with some context. Once two sets of subtitled clips for each film are obtained, we

```

<!-- VEIGA DTD -->
<!ELEMENT cluvi_veiga (header, body) >
<!ATTLIST cluvi_veiga
    version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
    xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | s | l | hi | ph)*>
<!ELEMENT hi (#PCDATA | l)*>
<!ATTLIST hi
    type CDATA #IMPLIED
    x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
    x CDATA #IMPLIED>
<!ELEMENT s EMPTY>
<!ATTLIST s
    n CDATA #IMPLIED
    d CDATA #IMPLIED
    a CDATA #IMPLIED>
<!ELEMENT l EMPTY>

```

Figure 6. VEIGA Corpus XML specification

link them to their corresponding text in the bitextual TMX-based CLUVI representation by means of their video clip identification tag, encoded both in the TMX file and in the video clip (in its file name).

These two sets of subtitled clips are stored as FLV files (because of their compression rate and small file size) in the server file system, where they are named – with a unique file name – according to their film title, their subtitle language (English or Galician), and their sequential number. Thus, whenever users search the VEIGA they get both the bilingual text pair and the clips where this text/subtitle appears. On the other hand, the bitextual TMX files are stored with a file name according to their film title, and include the tags of both the in-cue and out-cue time of each subtitle and their sequential number. This information is encoded in the VEIGA Corpus with a second element added to the tagging, the `<s>` element, which contains three attributes for each tagged subtitle: *s* for the sequential number, *d* for the in-cue time, and *a* for out-cue time. Furthermore, line breaks within subtitles are encoded with the `<l/>` tag, an element added to the TMX 1.4 specification to allow for the examination of aspects which are relevant to subtitling, such as typographical conventions and space constraints. To illustrate this tagging, Figure 7 shows a fragment of the code included in the TMX file named *peixe.tmx* (from the film entitled *Shooting Fish*, by Stefan Schwartz)

```

<tu>
  <tuv xml:lang="en"><seg><s n="848" d="01:07:32,351"
a="01:07:34,342 "/>We play our cards right,</>we could end up with... <s
n="849" d="01:07:34,431 " a="01:07:36,023 "/>two million pounds of
tobacco</>to spend it for us.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="808" d="01:07:33,271 "
a="01:07:36,946 "/>Podemos gastar 2 millóns en tabaco</>e pósters de Pamela
Anderson.</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><s n="850" d="01:07:36,511 " a="01:07:39,025 "/>I
meant to get someone</>to spend it for us.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="809" d="01:07:37,071 "
a="01:07:39,539 "/>Buscaremos alguén</>que o gaste por nós.</seg></tuv>
</tu>

```

**Figure 7.** Fragment of the VEIGA Corpus

that would correspond to the video clips stored in the file system as `peixe_en-848.flv`, `peixe_en-849.flv`, `peixe_gl-808.flv`; and `peixe_en-850.flv`, and `peixe_gl-809.flv`.

With regard to its applications, the VEIGA Corpus may serve a number of potential uses and purposes. First, it may be exploited as a reservoir of examples, offering researchers and scholars a database to analyse the different strategies and procedures used in both interlingual and intralingual subtitling and helping them substantiate their theoretical assumptions with practical evidence. From a pedagogical perspective, the VEIGA features suggest that it could be used for different purposes in various learning settings, ranging from general language courses dealing with pronunciation, register, collocations, and other features of oral and written discourse, to specialized courses in audiovisual translation with a focus on interlingual and intralingual subtitling (Sotelo Dios 2015; Sotelo Dios 2016). Concerning language learning, the use of assorted “real” texts, and particularly intralingual subtitles for L1 learning and interlingual subtitles for L2 learning, is likely to increase students’ motivation and cultural awareness, although careful selection, adaptation and designing of teaching materials and activities coupled with adequate teacher guidance need to be in place. At the same time, the VEIGA multimedia corpus may also prove a useful e-learning tool, since it would provide students with the possibility of exploring textual properties while listening to and watching film clips, which can be played and stopped at will, thus promoting autonomous learning. Finally, professional subtitlers could also benefit from the possibility to access a collection of ready-made subtitles, where they can look at how other practitioners solved particular subtitling challenges.

### 3. The SensoGal Corpus

This section will review the different methodologies and resources used to build the SensoGal Corpus,<sup>8</sup> an English–Galician parallel corpus semantically annotated with respect to WordNet and aligned at sentence and word level. The original English texts included in the SensoGal Corpus come from the SemCor Corpus, a textual corpus semantically annotated at lexical level and formed by 360,000 words distributed among 352 texts taken from the Brown Corpus (Landes/Leacock & Tengi 1998). The words of SemCor are tagged with an indication of the particular sense that they possess in their context of occurrence. This tagging uses the senses established in the English WordNet, a lexical resource elaborated by the same research team from the University of Princeton who carried out the annotation of the SemCor corpus. This is currently the largest semantically annotated and freely available corpus of real texts of a language, with 192,639 content words (nouns, verbs, adjectives and adverbs) annotated with their sense with respect to WordNet.

WordNet is a lexical database of the English language, organized as a semantic network where the nodes are concepts represented as sets of synonyms and the links between nodes are semantic relations between lexical concepts (Miller et al. 1990). The nodes contain nouns, verbs, adjectives and adverbs grouped by synonymy. In WordNet terminology, a set of synonyms is called a synset. Thus, each synset represents a distinct lexicalized concept and includes all the synonymous variants of this concept. In the WordNet model of lexical representation, the synsets are linked by means of lexical-semantic relations. Some of the most frequent relations represented in WordNet are hypernymy/hyponymy and holonymy/meronymy for nouns; antonymy and quasi-synonymy for adjectives; antonymy and derivation for adverbs; and entailment, hypernymy/hyponymy, cause and opposition for verbs.

WordNet, which was originally developed for English, is now available in many other languages, although the English WordNet still stands as the most complete reference version. Created and maintained at Princeton University since 1985, version 3.0 – used in the SensoGal Corpus – contains 206,941 lemmas, that is synonymous variants (155,287 of which are unique, non-homographic forms) grouped into 117,659 sets of synonyms or synsets. Many of the WordNet versions in languages other than English follow the design model of EuroWordNet (Vossen 2002), where the synsets of a particular language are linked to the synsets of the other languages through an InterLingual Index (ILI) that is unique to each concept, and which is mainly based on the synsets of the English WordNet. Therefore, the set of WordNet lexicons in different languages allows the connection between the synsets of any pair of languages via the ILI index, thus constituting a very

---

8. ISLRN: 653–144–288-768-2.



useful resource in applications of linguistic technologies dealing with multilingual processing such as SensoGal.

The semantic tagging in the SensoGal Corpus is based on English WordNet 3.0 and on Galnet,<sup>9</sup> the Galician WordNet. The goal of the Galnet project (Gómez Clemente et al. 2013; Solla Portela & Gómez Guinovart 2015; Álvarez de la Granja/ Gómez Clemente & Gómez Guinovart 2016), carried out at the SLI, is building a WordNet for Galician aligned with the English WordNet 3.0, following the expand model (Vossen 1998) for the creation of new wordnets, where the variants associated with the Princeton WordNet synsets are translated using different strategies. Table 2 shows the lexical coverage of English WordNet 3.0, and the current status for Galician in its development version 3.0.25 as available via Galnet's web interface.

Table 2. WordNet synsets and variants by language

	English (WordNet 3.0)		Galician (Galnet 3.0.25)	
	variants	synsets	variants	synsets
<b>Nouns</b>	146,312	82,115	45,810	30,565
<b>Verbs</b>	25,047	13,767	6,985	3,069
<b>Adjectives</b>	30,002	18,156	10,317	6,312
<b>Adverbs</b>	5,580	3,621	1,643	1,074
<b>Total</b>	206,941	117,659	64,755	41,020

The process of creation of the SensoGal Corpus begins with the adaptation to WordNet 3.0 of the semantic tagging of the SemCor Corpus, originally annotated with respect to WordNet 1.6. Then, the manual translation of the texts into Galician is carried out and, simultaneously, the new variants derived from the translation are introduced into the Galician WordNet. After the translation, the semantic labels of English are projected on to the Galician texts. Finally, a parallel English–Galician corpus is built up in TMX with the results of the semantic annotation of Galician texts (Solla Portela & Gómez Guinovart 2017).

The XML specification for SensoGal encoding follows the general conventions of the TMX format, as shown in the fragment of the corpus in Figure 8. The tagged versions of the original and translated sentences are stored as specialized variants of the translation unit (<tu> element) as character data (CDATA), in order to distinguish between the TMX structural markup and the internal sentence labelling. All content words in a sentence (nouns, verbs, adverbs and adjectives) are annotated with their lemma and tagged with their sense expressed as their

9. ISLRN: 544–286–653–437–9.

```

<tu>
  <prop type="group">br-g15:41</prop>
  <tuv xml:lang="en">
    <seg>This man's isolation is not merely momentary, it is permanent.</seg>
  </tuv>
  <tuv xml:lang="en-tag">
    <seg>
      <![CDATA[<wf pos="DT">This</wf> <wf lemma="man" ili="ili-30-
10287213 -n">man</wf> <wf pos="POS">'s</wf> <wf lemma="isolation" ili="ili-
30-14414715 -n">isolation</wf> <wf lemma="be" ili="ili-30-02604760 -
v">is</wf> <wf lemma="not" ili="ili-30-00024073 -r">not</wf> <wf
lemma="merely" ili="ili-30-00004722 -r">merely</wf> <wf lemma="momentary"
ili="ili-30-01443097 -a">momentary</wf> <punc>,</punc> <wf
pos="PRP">it</wf> <wf lemma="be" ili="ili-30-02604760 -v">is</wf> <wf
lemma="permanent" ili="ili-30-01754421 -a">permanent</wf>
<punc>.</punc>]]>
    </seg>
  </tuv>
  <tuv xml:lang="gl">
    <seg>Este illamento do home non é só momentáneo, é permanente.</seg>
  </tuv>
  <tuv xml:lang="gl-tag">
    <seg>
      <![CDATA[Este <wf lemma="illamento" ili="ili-30-14414715 -
n">illamento</wf> do <wf lemma="home" ili="ili-30-10287213 -n">home</wf>
<wf lemma="non" ili="ili-30-00024073 -r">non</wf> <wf lemma="ser" ili="ili-
30-02604760 -v">é</wf> <wf lemma="só" ili="ili-30-00004722 -r">só</wf> <wf
lemma="momentáneo" ili="ili-30-01443097 -a">momentáneo</wf> , <wf
lemma="ser" ili="ili-30-02604760 -v">é</wf> <wf lemma="permanente" ili="ili-
30-01754421 -a">permanente</wf>.</punc>]]>
    </seg>
  </tuv>
</tu>

```

Figure 8. Fragment of the SensoGal Corpus

WordNet ILI index. In addition, the *<prop>* element, used in TMX “to define the various properties of the parent element” (Savourel 2005), is used in the SensoGal encoding to identify all the phrases in the corpus with their bibliographical reference (abbreviation of text title and phrase number).

So far thirty Galician translations have been semantically tagged and aligned with their corresponding original English texts, totalling 2,734 translation units with 61,236 English words and 62,577 Galician words. The resulting parallel corpus can be accessed for consultation through a dedicated web interface.<sup>10</sup>

10. Available at <http://sli.uvigo.gal/SensoGal/>.

#### 4. Conclusion

Since its publication, the CLUVI Corpus has been used as the empirical basis for a wide range of academic studies in the fields of translational stylistics (Moreira 2010; Moreira 2011a; Sotelo Dios 2011), translation teaching (Sotelo Dios 2015; Sotelo Dios 2016), computational lexicology (Girju 2007a; Girju 2007b; Gómez Guinovart & Oliver 2014), terminology (Gómez Guinovart & Torres Padín 2006; Crespo et al. 2008; Gómez Guinovart & Simões 2009; Simões & Gómez Guinovart 2009; Moreira 2011b; Gómez Guinovart 2012; Moreira 2014) and multilingual lexicography (Gómez Guinovart & Sacau Fontenla 2004a; Gómez Guinovart & Sacau Fontenla 2005; Gómez Guinovart/Díaz Rodríguez & Álvarez Lugrís 2008; Gómez Guinovart & Simões 2010; Álvarez Lugrís & Gómez Guinovart 2014).

More generally, over the last decade, parallel corpora have proven very useful in many different applications in the fields of translation and language teaching, computational terminology and lexicography, applied linguistics, computer-aided translation and machine translation. The incorporation of multimedia data into parallel corpora will permit to transcend the traditional text-only approach to corpus design, reflecting the polisemiotic aspects of cultural products like film discourse and subtitling. Without any doubt, multimedia parallel corpora will represent in the very next future an important resource in the areas of language and cultural studies, second and foreign language teaching and in translation studies and professional practice.

Furthermore, the annotation of lexical meaning in parallel corpora increases their possibilities of exploitation in applied linguistics and language processing. Lexical-semantic processing is crucial for information society key technologies like conversational agents, information extraction and question-answering systems. Although much effort is still required to complete this task, the SemCor Corpus certainly represents a resource of vital importance for the development of Galician language technologies. Its exploitation should enable the construction of tools of great interest in the field of semantic processing – especially in tasks that require multilingual knowledge –, and the development of more efficient applications for language processing.

#### References

- Almeida, José João, Araújo, Sílvia, Simões, Alberto & Dias, Idalete. 2014. The Per-Fide Corpus: A New Resource for Corpus-based Terminology, Contrastive Linguistics and Translation Studies. In *Working with Portuguese Corpora*, Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds), 177–200. London: Bloomsbury Publishing.

- Álvarez de la Granja, María, Gómez Clemente, Xosé María & Gómez Guinovart, Xavier. 2016. Introducing idioms in the Galician wordnet: methods, problems and results. *Open Linguistics* 2: 253–286.
- Álvarez Lugrís, Alberto & Gómez Guinovart Xavier. 2014. Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Diccionario Moderno Inglés-Galego. In *Lexicografía de las lenguas románicas: Aproximaciones a la lexicografía moderna y contrastiva*, María José Domínguez Vázquez, Xavier Gómez Guinovart Xavier & Valcárcel Riveiro Carlos (eds), 31–48. Berlin/Boston: De Gruyter Mouton.
- Crespo Bastos, Ana, Gómez Clemente, Xosé María, Gómez Guinovart Xavier & López Fernández Susana. 2008. XML-based Extraction of Terminological Information from Corpora. In *Actas da 6ª Conferência Nacional XATA2008: XML, Aplicações e Tecnologias Associadas*, José Carlos Ramalho, João Correia Lopes & Salvador Abreu (eds), 28–39. Évora: Universidade de Évora.
- Girju, Roxana. 2007a. Experiments with an Annotation Scheme for a Knowledge-rich Noun Phrase Interpretation System. In *Proceedings of the Linguistic Annotation Workshop*, 168–175. Prague: ACL.
- Girju, Roxana. 2007b. Improving the Interpretation of Noun Phrases with Cross-linguistic Information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 568–575. Prague: ACL.
- Gómez Clemente, Xosé María, Gómez Guinovart, Xavier, González Pereira, Andrea & Verónica Taboada Lorenzo. 2013. Sinonimia e rexistros na construción do WordNet do galego. *Estudos de lingüística galega* 5: 27–42.
- Gómez Guinovart Xavier & Oliver, Antoni. 2014. Methodology and evaluation of the Galician WordNet expansion with the WN-toolkit. *Procesamiento del Lenguaje Natural* 53: 43–50.
- Gómez Guinovart Xavier & Sacau Fontenla Elena. 2004a. Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural* 33: 133–140.
- Gómez Guinovart Xavier & Sacau Fontenla Elena. 2004b. Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds), 1179–1182. Paris: ELRA.
- Gómez Guinovart Xavier & Sacau Fontenla Elena. 2005. Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Diccionario CLUVI Inglés-Galego. In *Viceversa* 11: 159–171.
- Gómez Guinovart Xavier & Simões, Alberto. 2009. Parallel corpus-based bilingual terminology extraction. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*. Toulouse: Université Paul Sabatier. <<https://www.irit.fr/TIA09/thekey/posters/simoes-guinovart.pdf>> (28 April 2017).
- Gómez Guinovart, Xavier & Simões, Alberto. 2010. Translation dictionaries triangulation. In *Proceedings of FALA2010: VI Jornadas en Tecnología del Habla & II Iberian SLTech*, Carmen García Mateo, Francisco Campillo Díaz & Francisco Méndez Pazó (eds), 171–174. Vigo: Universidade de Vigo.
- Gómez Guinovart, Xavier & Torres Padín, Ánxeles. 2006. Extracción dun vocabulario xurídico-administrativo galego-castelán a partir dun corpus paralelo. In *Terminología y derecho: la complejidad de la comunicación multilingüe*, M. Teresa Cabré, Carme Bach & Jaume Martí (eds), 175–188. Barcelona: Universitat Pompeu Fabra.

- Gómez Guinovart, Xavier, Díaz Rodríguez, Eva & Álvarez LUGRÍS, Alberto. 2008. Aplicación da lexicografía bilingüe baseada en corpora na elaboración do Dicionario CLUVI inglés-galego. *Viceversa* 14: 71–87.
- Gómez Guinovart, Xavier. 2012. A hybrid corpus-based approach to bilingual terminology extraction. In *Encoding the Past, Decoding the Future: Corpora in the 21st Century*, Isabel Moskowich-Spiegel Fandiño & Begoña Crespo (eds), 147–175. Newcastle upon Tyne: Cambridge Scholar Publishing.
- Keshtkar, Hossein & Mosavi Miangah, Tayebbeh. 2012. Using Bilingual Parallel Corpora in Translation Memory Systems. *International Journal of Applied Linguistics and English Literature* 1.5: 184–193.
- Koehn, Philipp. 2005. EuroParl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X: The Tenth Machine Translation Summit Proceedings*, 79–86. Tokyo: Asia-Pacific Association for Machine Translation.
- Landes, Shari, Leacock, Claudia & Tengi, Randee I. 1998. Building semantic concordances. In *WordNet: An Electronic Lexical Database*, Christiane Fellbaum (ed), 199–216. Cambridge: The MIT Press.
- Mikhailov, Mikhail & Cooper, Robert. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. Abingdon: Routledge.
- Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek & Miller, Katherine. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3: 235–244.
- Montero Perez, Maribel, Paulussen, Hans Macken, Lieve & Desmet, Piet. 2014. From input to output: the potential of parallel corpora for CALL. *Language Resources and Evaluation* 48.1: 165–189.
- Moreira, Adonay. 2010. Estratégias de tradução em sites das regiões de turismo de Portugal: estudo baseado em corpus. *Polissema: Revista de Letras do ISCAP* 10: 13–42.
- Moreira, Adonay. 2011a. The translator as cultural mediator: a corpus-based study of omissions and additions in translations of tourism brochures. *The Journal of Cultural Mediation* 1: 86–95.
- Moreira, Adonay. 2011b. Turigal: compilation of a parallel corpus for bilingual terminology extraction. In *Actas del III Congreso Internacional de Lingüística de Corpus: Las tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus*, María Luisa Carrió & Miguel Ángel Candel (eds), 33–42. València: Universitat Politècnica de València.
- Moreira, Adonay. 2014. A methodology for building a translator- and translation-oriented terminological resource. In *inTRAlinea Special Issue: Translation & Lexicography*, María Sánchez, María Porciel & Iris Serrat (eds). <<http://www.intraline.org/specials/article/2032>> (28 April 2017).
- Santos, Diana. 2004. *Translation-based Corpus Studies: Contrasting English and Portuguese Tense and Aspect Systems*. Amsterdam: Rodopi.
- Savourel, Yves. 2005. *TMX 1.4b Specification*. Localisation Industry Standards Association. <<https://www.gala-global.org/tmx-14b>> (28 April 2017).
- Simões, Alberto & Gómez Guinovart, Xavier. 2009. Terminology extraction from English–Portuguese and English–Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns. In *Proceedings of the Iberian SLTech 2009 - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, António Teixeira, Miguel Sales Dias & Daniela Braga (eds), 13–16. Porto Salvo: Designeed.

- Simões, Alberto, Gómez Guinovart, Xavier & Almeida, José João. 2004. Distributed translation memories implementation using WebServices. *Procesamiento del Lenguaje Natural* 33: 89–94.
- Solla Portela, Miguel Anxo & Gómez Guinovart, Xavier. 2015. Galnet: o WordNet do galego. *Aplicacións lexicolóxicas e terminolóxicas. Revista Galega de Filoloxía* 16: 169–201.
- Solla Portela, Miguel Anxo & Gómez Guinovart, Xavier. 2017. Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0. *Procesamiento del Lenguaje Natural* 59: 137–140.
- Sotelo Dios Patricia & Guinovart Xavier, Gómez. 2012. A multimedia parallel corpus of English–Galician film subtitling. In *1st Symposium on Languages, Applications and Technologies*, Alberto Simões, Ricardo Queirós & Daniela da Cruz (eds), 255–266. Saarbrücken: Dagstuhl Publishing.
- Sotelo Dios, Patricia. 2011. Using a multimedia parallel corpus to investigate English–Galician subtitling. In *Proceedings of the SDH 2011 Conference: Supporting Digital Humanities*, Bente Maegaard (ed). Copenhagen: University of Copenhagen. <<http://hnk.ffzg.hr/bibl/SDH-2011/proceedings.html>> (28 April 2017).
- Sotelo Dios, Patricia. 2015. Using a multimedia corpus of subtitles in translation training. In *Affordances of Language Corpora for Data-driven Learning*, Agnieszka Leńko-Szymańska & Alex Boulton (eds), 245–266. Amsterdam: John Benjamins.
- Sotelo Dios, Patricia. 2016. Adquisición de competencias en traducción audiovisual mediante un corpus multimedia. In *New Insights into Corpora and Translation*, Daniel Gallego Hernández (ed), 1–16. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 2214–2218. Istanbul: ELRA.
- Tufiş, Dan. 2007. Exploiting Aligned Parallel Corpora in Multilingual Studies and Applications. In *Intercultural Collaboration*, Toru Ishida, Susan R. Fussell & Peek Vossen (eds), 103–117. Berlin: Springer.
- Véronis, Jean, ed. 2000. *Parallel Text Processing: Aligement and Use of Translation Corpora*. Dordrecht: Kluwer.
- Vossen, Piek. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Norwell: Kluwer Academic Publishers.
- Vossen, Piek. 2002. WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée* 7: 27–38.