CHAPTER SIX

A HYBRID CORPUS-BASED APPROACH TO BILINGUAL TERMINOLOGY EXTRACTION

XAVIER GÓMEZ-GUINOVART

1. Introduction

The orientation towards applied research based on textual corpora has consolidated in recent years as a successful methodology for the description and analysis of linguistic phenomena in virtually all respects. This article will present an approach to parallel corpus-based research in the field of bilingual terminology, using the English-Galician software localization parallel corpus which is part of the CLUVI Corpus as the empirical source for the extraction of terminological information.

The study of language through textual corpora allows for an empirical approach to the actual use of language in context. The analysis of lexical units in a general corpus allows observation of their semantic potential, their frequency of use and their combinatory in a very realistic way, certainly unattainable from a purely introspective reflection on the functioning of language. Similarly, in the study of technical or specialised linguistic discourse, exploitation of technical corpora with proper tools makes it easy to identify the repertoire of lexical units with terminological content used in texts, to analyse their polysemy and synonymy, to check their frequency in the texts, to get real examples of usage and definitory contexts, and even to discover the semantic relations encoded in the texts between the terms associated with a speciality domain.

Traditionally, the authors of lexical repertoires have sought information sources on linguistic data in other lexical repertoires, in selected passages of the literary canon, or in their own intuition as speakers of the language. This method of work involved considerable constraints for lexicographical practice, because, first, the lexicographers' own reflections on the use of vocabulary might not be adjusted to the

linguistic reality; on the other hand, the manual collection of quotations from canonical texts was a slow and very improductive kind of work; and, finally, the dictionaries used as a source of inspiration tended to be out of date or, in the worst case, could contain errors accumulated due to their subsequent reproduction over time.

The introduction of the use of computerised textual corpora in lexicographical practice undoubtedly contributes to overcoming these limitations of traditional methodology, facilitating the observation of the vocabulary in real language use, in its textual practice. The first success in introducing the use of computerised corpora for the production of dictionaries is due to the University of Birmingham and Collins publishing house, who in 1987 released the first edition of the Cobuild dictionary (Sinclair, 1987). At the time, the Cobuild project was very innovative in that it represented the first use in lexicography of a representative corpus of texts to facilitate word meaning analysis, syntactic pattern recognition and the description of collocations and phraseology of a language, in this case, contemporary English. Following the success of Cobuild, the working methodology of corpus-based lexicography was followed by other major publishers such as Oxford University Press, Longman and Larousse (who collaborated in the development of the British National Corpus) and Cambridge University Press.

Although the theoretical bases for corpus-based terminology are similar to those for corpus-based lexicography, corpus-based terminology has taken longer to become a standard operating procedure, probably because these two disciplines work with corpora of a very different nature: whereas lexicography corpora are usually broad and general, terminology corpora, on the contrary, are much more oriented to specific domains where texts are often difficult to obtain.

However, at present, modern terminology holds theoretical and methodological principles that emphasise the importance of using large text repertoires for terminological work, because these repertoires facilitate the identification in texts of specialised content units, as well as the extraction of the terminological information associated with these units and encoded in texts. As the Communicative Theory of Terminology (Cabré, 1999) states, texts are conceived in modern terminology as the "natural habitat" of terms, the environment in which to observe the true nature of the units of specialised value. In this sense, the modern theory of terminology replaces the prescriptive paradigm of the General (or Traditional) Theory of Terminology with a descriptive overview of its subject, a perspective that our research group at the University of Vigo shares and that has naturally led us to the adoption of a corpus-based

approach in our research in the field of terminology for the Galician language.

I will now present the work our university research group is carrying out in the field of bilingual terminology extraction, using as source the English-Galician software localization parallel corpus which is part of the CLUVI Corpus, with the aim of extending the Termoteca, a terminological knowledge base for Galician built from specialised textual corpora.

2. Source and target data

First of all, I will explain the main characteristics of the CLUVI Corpus and of the Termoteca terminological database, which respectively constitute the source and target application selected to illustrate the hybrid method of bilingual terminology extraction that I am presenting here.

The CLUVI Corpus[1] is an open collection of textual corpora with translations in specific areas of the contemporary Galician language (Gómez-Guinovart 2003). With over 23 million words, the CLUVI Corpus comprises six main parallel corpora belonging to four specialised registers (from fiction, computing, popular science and legal-administrative fields) and five different language combinations with Galician (Galician-Spanish bilingual translation, English-Galician bilingual translation, French-Galician bilingual translation, English-Galician-French-Spanish tetralingual translation and Spanish-Galician-Catalan-Basque tetralingual translation). The format chosen for storing the aligned parallel texts is an adaptation of the TMX format (Gómez-Guinovart and Sacau 2004b), as this is the XML encoding standard for translation memories and parallel corpora, regardless of the application used. A translation memory is a database which holds the original and translated version for each of the sentences translated within a computer-aided translation system, with the aim of reusing translations by the program. With some differences, an aligned parallel corpus is equivalent to a translation memory and, in practice, there is a considerable number of TMX-encoded aligned parallel corpora, with the added advantage that these corpora can be used as translation memories for feeding computer-aided translation programs.

The Termoteca[2] is a terminological database based on the monolingual and parallel speciality texts collected in the corpora of the University of Vigo (Gómez-Clemente and Gómez-Guinovart 2006b), namely in the CLUVI Corpus and in the Galician Technical Corpus[3] (Gómez-Clemente and Gómez-Guinovart 2006a). This terminological database is freely accessible on the web, and currently has over 8,000 multilingual registers corresponding to 16,120 unique terms (or variants) documented in these

corpora belonging to the areas of law, sociology, economy, ecology, medicine and tourism. All terms in the Termoteca are documented in the Vigo corpora, the terminological inventories in the fields of computer science being in progress.

In the Termoteca, terminological information is structured around concepts. Each Termoteca record includes all the information relating to a concept expressed with a Galician term which can be recorded also with variants, both intralinguistic (synonymous terms, spelling variants, or dialectal variants) and interlinguistic (translations or, more properly, equivalences). The information collected for each variant (including the common or unmarked variant) includes the lemma of the term, its grammatical category, its definition, and a context of usage documented in the corpus. Registers or concepts in the database are grouped according to their thematic area within a branch of a hierarchical thematic tree of the topic. Also, the concepts in the database form a surfable lexical-semantic network where conceptual nodes interact with each other according to the semantic relations (antonymy, hyperonymy, holonymy, etc.) among them.

3. Extraction of term candidates

Terminology extraction from the CLUVI Parallel Corpus is applied from both a monolingual and a bilingual perspective. In both cases, terminology extraction may be focused on the extraction of monolexical term candidates or on the extraction of multilexical term candidates. We employ a set of different extraction techniques for each of these four tasks.

From now on I will illustrate the techniques applied to the LOGALIZA Corpus of English-Galician software localization which is part of the CLUVI Corpus using some standard tools for text processing in the Unix environment, as well as some applications included in the NATools software package. The LOGALIZA Corpus is a collection of English-Galician software localizations with 3,706,242 words (1,777,874 words from the English original texts, and 1,928,368 from the Galician translations) segmented in 284,341 translation units, and it includes translations of well-known software packages like Windows XP, OpenOffice 2.0, Gnome 2.18, Debian GNU/Linux 4.0r3 "etch", KDE 4.0.1, Firefox 3.0, Sunbird 0.8, Thunderbird 3, and Seamonkey 2.0. The NATools[4] is an open source workbench for parallel corpora processing which includes a sentence aligner, a probabilistic translation dictionaries extractor, a word aligner, a terminology extractor, and a set of other tools to study the aligned parallel corpora (Simões and Almeida 2003).

3.1. Monolingual perspective

Terminology extraction from a monolingual perspective works with the source and target language texts separately. For instance, we can easily obtain a pair of source and target language texts from a TMX-encoded aligned parallel corpus using the nat-tmx2pair script from the NATools workbench. Thus, from the English-Galician localization parallel corpus included in the CLUVI Corpus (say logaliza.tmx), we can generate one text per language (say logaliza.tmx-gl and logaliza.tmx-en) in a simple step as the basis for processing:

(1) nat-tmx2pair logaliza.tmx

3.1.1. Monolexical monolingual candidates

Monolexical candidates in the monolingual corpora generated in this way are identified taking into account their frequency, their specificity and their form. First of all, we generate a frequency dictionary for both of the monolingual corpora, discarding tokens with an absolute frequency lower than five. This frequency list can be obtained using free Unix standard tools as uniq, sort and awk, for instance:

(2) cat logaliza.tmx-gl | tr -cs "\'\-[:alnum:]" '\n' | tr '[:upper:]' '[:lower:]' | sort | uniq -c | sort -k 1,1nr -k 2 | awk '{if ($1 >= 5) print $2 "\t" $1 }' > logaliza-gl.freq.1

The frequency list created (with over 15,000 different lexical forms) would include a great number of non-terminological words. Therefore, we filter this file by deleting all the words which occur in a non-terminological corpus of the language. As for Galician, the exclusion corpus used in this task is a literary corpus of over 2 million words containing 145 fiction works (mainly novels and short stories) from the Virtual Library of Universal Literature in Galician.[5] The resulting filtered frequency list of about 8,000 different lexical elements contains more specific lexical items with a higher probability of being considered term candidates by the human terminologist, as can immediately be seen by comparing the very beginning of the raw frequency list (logaliza-gl.freq.1) with the corresponding fragment from the new file (logaliza-gl.freq.1.filtered).

(3)

Raw frequencies		Rank	Filtered frequencies	
de	106509	1	ficheiro	9815
a	55626	2	prema	5041
o	51898	3	ficheiros	4582
do	25552	4	formato	3253
para	25325	5	cartafol	3234
non	23415	6	usuario	2683
que	20611	7	seleccione	2616
e	20309	8	opcións	2515
se	19375	9	editar	2431
da	17904	10	icona	2338
un	16457	11	inserir	2325
os	14412	12	contrasinal	2072
en	13313	13	introduza	1971
as	11687	14	devolve	1835
unha	11219	15	computador	1759
ou	11056	16	mensaxes	1679
no	10744	17	certificado	1622
ficheiro	9815	18	windows	1607
é	9404	19	asistente	1583
ao	9135	20	seleccionar	1580
na	9003	21	seleccionado	1572
pode	8619	22	dispositivo	1517
nome	6932	23	selección	1395
por	6891	24	kde	1383
texto	6250	25	web	1370
datos	6166	26	valores	1335
como	6047	27	válido	1263
está	5583	28	enderezos	1256
esta	5380	29	comandos	1235
prema	5041	30	configurar	1192
con	4934	31	separador	1147
este	4676	32	directorio	1125
á	4673	33	activar	1114

Obviously not all the words listed in this filtered frequency list are good term candidates from the point of view of our intended application (widening the Termoteca). For this purpose, we carry out a human selection (by terminologists) from the list of proposed candidates, using

their form as well as their frequency as a key to clustering candidates in order to facilitate selection. For instance, the infinitive forms of the verbs are very likely candidates to enter the Termoteca from a textual corpus on software localization. A good amount of these verb forms can be retrieved according to their word ending, as could be done with this grep command:

(4) grep -E '[aei]r\b' logaliza-gl.freq.1.filtered > logaliza-gl.infinitives

This would produce a list of infinitive forms with a great number of reliable term candidates, as can be observed in the first lines of the generated file (logaliza-gl.infinitives):

(5)

-ar, -er, -ir endings		**Rank**
editar	2431	1
inserir	2325	2
seleccionar	1580	3
configurar	1192	4
activar	1114	5
definir	1110	6
estándar	695	7
actualizar	637	8
desactivar	624	9
exportar	613	10
cancelar	592	11
premer	582	12
conectar	544	13
localizar	477	14
especificar	447	15
restaurar	444	16
reiniciar	434	17
renomear	311	18
debuxar	304	19
desfacer	296	20
enter	281	21
personalizar	265	22

Other interesting clusters for terminological purposes may be set with the words beginning with certain frequent prefixes in this field like *hiper-*, *multi-*, *auto-*, *tele-* or *meta-*. It must be noted that non-infinitive Galician

forms (like *enter* or *estándar*) are also included in this list (logaliza-gl.infinitives) on the basis of their ending, since the grep search is applied to the untagged version of the corpus. At this level of analysis, which is focused on considering if a monolexical item is a term candidate by itself in a certain language, tagging the text would not mean a great advantage, given the kinds of texts present in the software localizations, which are characterised by short or very short sentences, the presence of programming code mixed with "real" text, and a high number of foreign words.

3.1.2. Multilexical monolingual candidates

Contrary to monolexical term candidates, multilexical candidates in monolingual corpora are identified by using a part-of-speech tagged version of the corpora, and by taking into account both their frequency and their syntactic pattern. In this section I will show how this task is accomplished using a version of Galician translations into the LOGALIZA Corpus annotated with word lemma and part-of-speech information with the FreeLing tagger for Galician language[6] (Padró *et al*. 2010).

First, what follows would be a little fragment of this annotated corpus tagged with FreeLing (logaliza.tmx-gl.tagged):

(6)

Actualizando_actualizar_VMG0000
información_información_NCFS000 de_de_SPS00
pistas_pista_NCFP000 en_en_SPS00 a_o_DA0FS0 '_'_Fe
Biblioteca_biblioteca_NCFS000
multimedia_multimedia_AQ0FS0 '_'_Fe e_e_CC en_en_SPS00
A_o_DA0FS0 miña_meu_DP1FSS música_música_NCFS000
._._Fp Esta_este_DD0FS0 operación_operación_NCFS000
poderá_poder_VMIF3S0 tardar_tardar_VMN0000
algúns_algún_DI0MP0 minutos_minuto_NCMP000 ._._Fp
Configure_configurar_VMSP1S0 os_o_DA0MP0
protocolos_protocolo_NCMP000 de_de_SPS00
rede_rede_NCFS000 e_e_CC a_o_DA0FS0
configuración_configuración_NCFS000 de_de_SPS00
o_o_DA0MS0 servidor_servidor_NCMS000
mandatario_mandatario_AQ0MS0 ._._Fp
Seleccione_seleccionar_VMSP1S0 os_o_DA0MP0
protocolos_protocolo_NCMP000 que_que_PR0CN000
desexa_desexar_VMIP3S0 utilizar_utilizar_VMN0000

para_para_SPS00 recibir_recibir_VMN0000
transmisión_transmisión_NCFS000 en_en_SPS00
secuencia_secuencia_NCFS000 :_:_Fd
Especificar_especificar_VMN0000 o_o_DA0MS0
lugar_lugar_NCMS000 onde_onde_PR0CN000
se_se_PP3CN000 almacena_almacenar_VMIP3S0 a_o_DA0FS0
música_música_NCFS000 e_e_CC
modificar_modificar_VMN0000 a_o_DA0FS0
configuración_configuración_NCFS000 de_de_SPS00
a_o_DA0FS0 copia_copia_NCFS000 ._._Fp
Formato_formato_NCMS000 de_de_SPS00 o_o_DA0MS0
ficheiro_ficheiro_NCMS000 :_:_Fd
Comparar_comparar_VMN0000 o_o_DA0MS0
formato_formato_NCMS000 WMA_wma_NP00000
con_con_SPS00 outros_outro_DI0MP0
formatos_formato_NCMP000 ._._Fp
Localización_localización_NCFS000 para_para_SPS00
copias_copia_NCFP000 de_de_SPS00
seguranza_seguranza_NCFS000 e_e_CC
restauración_restauración_NCFS000
Crear_crear_VMN0000 copia_copia_NCFS000 de_de_SPS00
seguranza_seguranza_NCFS000 agora_agora_RG

Syntactic patterns with a high degree of terminology are retrieved from this tagged version of the corpus. From work previously done on the Termoteca, we can know the frequency distribution of syntactic patterns in Galician terms (Crespo *et al*., 2008). The top three syntactic patterns at Termoteca for multilexical Galician terms are Noun+Adjective (as in Biblioteca_biblioteca_NCFS000 multimedia_multimedia_AQ0FS0), Noun+Preposition+Noun (as in copia_copia_NCFS000 de_de_SPS00 seguranza_seguranza_NCFS000) and Noun+Preposition+Article+Noun (as in Formato_formato_NCMS000 de_de_SPS00 o_o_DA0MS0 ficheiro_ficheiro_NCMS000), so we can retrieve all these multilexical candidate patterns from the corpus with a simple grep for them:

(7)

```
grep -Eo '\b\w+_\w+_NC..... \w+_\w+_AQ....'
logaliza.tmx-gl.tagged > logaliza.tmx-gl.patterns.NA

grep -Eo '\b\w+_\w+_NC..... \w+_\w+_SPS00
\w+_\w+_NC.....' logaliza.tmx-gl.tagged > logaliza.tmx-
gl.patterns.NPN
```

```
grep -Eo '\b\w+_\w+_NC..... \w+_\w+_SPS00
\w+_\w+_DA.... \w+_\w+_NC.....' logaliza.tmx-
gl.tagged > logaliza.tmx-gl.patterns.NPDN
```

Even better, we can group, count and sort the occurrences of these highly terminological patterns, generating in a few steps a list of very probable term candidates (with a limit of no less than five occurrences) for the consideration of terminologists:

(8)

```
tr '[:upper:]' '[:lower:]' < logaliza.tmx-gl.patterns.NA |
sort | uniq -c | sort -k 1,1nr -k 2 | perl -pe
"s/\_.+?_(nc.....|aq....)//g;" | awk '{if ($1 >= 5) print $2,
$3 "\t" $1 }'> logaliza-gl.freq.NA

tr '[:upper:]' '[:lower:]' < logaliza.tmx-gl.patterns.NPN |
sort | uniq -c | sort -k 1,1nr -k 2 | perl -pe
"s/\_.+?_(nc.....|sps00)//g;" | awk '{if ($1 >= 5) print $2,
$3, $4 "\t" $1 }'> logaliza-gl.freq.NPN

tr '[:upper:]' '[:lower:]' < logaliza.tmx-gl.patterns.NPDN |
sort | uniq -c | sort -k 1,1nr -k 2 | perl -pe
"s/\_.+?_(nc.....|sps00?|da....)//g;" | awk '{if ($1 >= 5)
print $2, $3, $4, $5 "\t" $1 }'> logaliza-gl.freq.NPDN

cat logaliza-gl.freq.N* | awk '{$0=$0 ; print $NF"|"$0}' |
sort -nr | cut -d"|" -f2 > logaliza.tmx-gl.patterns
```

The list generated with these patterns by order of frequency (logaliza.tmx-gl.patterns) would include a great amount of interesting terms, as shown in this fragment from the beginning of the list which contains all the candidate terms with 200 or more occurences automatically extracted from the LOGALIZA Corpus:

(9)

Candidates by pattern		Rank
caixa de diálogo	1378	1
base de datos	1142	2
correo electrónico	885	3
tipo de letra	806	4
barra de ferramentas	641	5
axenda de enderezos	466	6

nome de usuario	424	7
pé de páxina	372	8
nome do ficheiro	368	9
menú de contexto	338	10
fonte de datos	317	11
copia de seguranza	307	12
tipos de letra	306	13
caixa de verificación	281	14
nome de ficheiro	271	15
folla de cálculo	268	16
bases de datos	255	17
curva elíptica	242	18
liña de comandos	223	19
documento actual	218	20
enderezo de correo	203	21

With such a processing from the source and target text in the parallel corpus we obtain two separated lists of monolexical and multilexical term candidates for both source and target language which may then serve as a basis for further exploratory work on establishing translation equivalences.

3.2. Bilingual perspective

Terminology extraction from a bilingual perspective works directly with the TMX-encoded sentence-level aligned parallel corpus, exploiting in different ways the information about translation equivalences annotated in texts. The parallel corpora-based bilingual terminology extraction methods are based on probabilistic translation dictionaries (PTDs) generated from automatic lexical alignment by the NATools workbench (Simões and Almeida 2003).

3.2.1. Monolexical bilingual candidates

The NATools PTDs, automatically extracted from sentence-aligned parallel corpora, map words from a source language to a set of probable translations in a target language. Each of these translations have a probabilistic measure of translatability, that is, the mutual translation probability for each word translation equivalence. From the TMX-encoded LOGALIZA Corpus (logaliza.tmx), for instance, we can generate both source/target and target/source PTDs in a simple step, by using the nat-create script from the NATools workbench:

(10)

```
nat-create -tmx logaliza.tmx
```

This script creates a binary representation of both probabilistic translation dictionaries which can then be converted to a Perl-style human-readable list (say logaliza.ptds.txt) by executing the nat-dumpDicts script from the same package:

(11)

```
nat-dumpDicts -full source.lex source-target.bin
target.lex target-source.bin > logaliza.ptds.txt
```

What follows is a simplified example of lexical entries from this human- (and computer-) readable file (logaliza.ptds.txt) for both translation directions (English/Galician and Galician/English). I have eliminated the unlikely equivalences from the translation candidates in order to make data interpretation as easy as possible.

(12)

```
"updating" => {
  count => 107,
  trans => {
      "actualizar" => 0.39344776,
    "actualizando" => 0.20914072,
   "actualización" => 0.14384077,
         "anovar" => 0.05695909,
    },
},
"library" => {
  count => 434,
  trans => {
     "biblioteca" => 0.92184907,
    "bibliotecas" => 0.00992135,
      "libraría" => 0.00357169,
    },
},
"search" => {
  count => 1772,
  trans => {
          "busca" => 0.41747451,
         "buscar" => 0.24641067,
```

```
            "procura" => 0.09335117,
           "procurar" => 0.07416935,
           "procuras" => 0.01810816,
            "buscas" => 0.01018680,
    },
},
"files" => {
  count => 3622,
  trans => {
        "ficheiros" => 0.92156011,
         "ficheiro" => 0.01192421,
         "arquivos" => 0.00071648,
    },
},
"windows" => {
  count => 1686,
  trans => {
          "windows" => 0.52585220,
          "xanelas" => 0.22492823,
           "ventás" => 0.10177911,
          "fiestras" => 0.06585271,
    },
}
[...]
"actualizando" => {
  count => 36,
  trans => {
         "updating" => 0.81126708,
       "refreshing" => 0.14974207,
        "upgrading" => 0.03899088,
    },
},
"cartafol" => {
  count => 3111,
  trans => {
           "folder" => 0.89990991,
        "directory" => 0.06933586,
             "path" => 0.00053614,
    },
},
"premer" => {
```

```
  count => 567,
  trans => {
           "click" => 0.34901118,
           "press" => 0.13431254,
        "clicking" => 0.11403722,
        "pressing" => 0.05724910,
         "pressed" => 0.05285491,
         "clicked" => 0.01327109,
    },
},
"personalizar" => {
  count => 193,
  trans => {
       "customize" => 0.82002574,
          "custom" => 0.10193013,
     "personalize" => 0.01794922,
     "customizing" => 0.00513785,
         "specify" => 0.00252011,
    },
},
"ficheiro" => {
  count => 9084,
  trans => {
           "file" => 0.92257315,
       "filename" => 0.02437326,
          "files" => 0.00210973,
    },
}
```

Each lexical entry in the PTD shows the number of occurrences of the source lexical item, and their possible translations with their equivalence probability calculated from sentence alignments. Bilingual monolexical term candidates may then be automatically inferred from the PTD by combination of data from monolingual monolexical extraction with the probabilistic translation data contained in the PTD.

By way of illustration, a sample of the result of this bilingual combination is shown here, with the term candidate list logaliza-gl.infinitives (previously generated) complemented with the translation candidates which achieve a probability of equivalence higher than 0.3. At this level of equivalence probability, the precision of results is reported to be higher than 90% (Gómez-Guinovart and Sacau 2004a).

(13)

-ar, -er, -ir endings	PTD translation probability > 0.3
editar	"edit" => 0.90549797
inserir	"insert" => 0.80692744
seleccionar	"select" => 0.90574151
configurar	"configure" => 0.61180413
activar	"enable" => 0.63463140
definir	"set" => 0.65490377
estándar	"standard" => 0.87318891
actualizar	"update" => 0.64495027
desactivar	"disable" => 0.61879921
exportar	"export" => 0.84409225
cancelar	"cancel" => 0.79754061
premer	"click" => 0.34901118
conectar	"connect" => 0.57319689
localizar	"find" => 0.78408235
especificar	"specify" => 0.91307855
restaurar	"restore" => 0.82364434
reiniciar	"restart" => 0.44223973
renomear	"rename" => 0.77977806
debuxar	"draw" => 0.69876260
desfacer	"undo" => 0.58765787, "undone" => 0.32991102
personalizar	"customize" => 0.82002574

3.2.2. Multilexical bilingual candidates

The multilexical bilingual terminology extraction method used is also based on NATools probabilistic translation dictionaries, and was explained in some detail in previous works of the group (Gómez-Guinovart and Simões 2009; Simões and Gómez-Guinovart 2009). NATools dictionaries include a probabilistic measure of translatability for each word pair which enables the creation of an alignment matrix for any translation unit. This alignment matrix includes in each cell the mutual translation probability for each bilingual lexical equivalence, as shown in the following table:

(14)

	a	conta	de	usuario	actual	non	é	recoñecida
the	27.94	0.00	2.17	0.00	0.06	0.00	0.41	0.00
current	0.24	0.00	0.17	0.00	**72.60**	0.00	0.00	0.00
user	0.00	0.00	0.67	**80.62**	0.00	0.00	0.00	0.00
account	0.00	**78.99**	0.00	0.00	0.00	0.00	0.00	0.00
is	1.09	0.00	1.14	0.00	0.00	0.00	49.92	0.00
not	0.00	0.00	0.00	0.00	0.00	69.77	0.00	0.00
recognized	0.00	0.00	0.00	0.00	0.00	0.00	0.00	36.24

These matrixes can be used to extract bilingual terminology using translation patterns that specify how word order in the source language changes after translation takes place. For instance, this table illustrates a very frequent English/Galician alignment pattern (*current user account/conta de usuario actual*) which is formally represented in the NATools workbench as A B C = C "de" B A.

These NATools translation patterns may include morphological restrictions (for one or both languages) defining the morphological categories allowed for the words matching the pattern. This makes it possible, for instance, to specify that the sequence at the right side of the translation pattern has to correspond to a common noun + *de* + common noun + adjective. NATools relies on external morphological analysers like FreeLing to validate these morphological restrictions. Thus, using FreeLing for Galician tagging, the translation pattern with morphological restrictions may be coded in the NATools formalism as A B C = C[CAT<-/^NC/] "de" B[CAT<-/^NC/] A[CAT<-/^AQ0/], or including all the grammatical alternations in the use of preposition, A B C = C[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/] A[CAT<-/^AQ0/]. For each variable containing a morphological restriction the system will invoke the morphological analyser and ask for the specific word analysis.

In order to extract a (both quantitatively and qualitatively) representative set of multilexical bilingual term candidates from the LOGALIZA Corpus, we have identified the most relevant English/Galician alignment patterns in the parallel texts, taking into account their categorial restriction on the Galician side. This is the set of rules (en_gl.cat_patterns) coded in the NATools formalism, where each rule begins with a unique identifier and ends at semicolon:

(15)

```
[R0004] A B C = C[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das"
B[CAT<-/^NC/] A[CAT<-/^AQ0/];
[R0051] A "of"|"in"|"for" B "and"|"or" C D = A[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" D[CAT<-/^NC/] B[CAT<-/^AQ0/]
"e"|"ou" C[CAT<-/^AQ0/];
[R0005] A "of"|"in"|"for" B C = A[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" C[CAT<-/^NC/] B[CAT<-/^AQ0/];
[R0031] A "of"|"in"|"for" B "and"|"or" C = A[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/] "e"|"ou" C[CAT<-
/^NC/];
[R0032] A "of"|"in"|"for" B "of"|"in"|"for" C = A[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" C[CAT<-/^NC/];
[R0003] A "of"|"in"|"for" B = A[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/];
[R0006] A B C = C[CAT<-/^NC/] A[CAT<-/^AQ0/] B[CAT<-
/^AQ0/];
[R0021] A "and"|"or" B C = C[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" A[CAT<-/^NC/] "e"|"ou" B[CAT<-
/^NC/];
[R0022] A B C D = D[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das"
C[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/]
"de"|"do"|"da"|"dos"|"das" A[CAT<-/^NC/];
[R0023] A B C = C[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das"
B[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das" A[CAT<-/^NC/];
[R0002] A B = B[CAT<-/^NC/] "de"|"do"|"da"|"dos"|"das"
A[CAT<-/^NC/];
[R0001] A B = B[CAT<-/^NC/] A[CAT<-/^AQ0/];
```

Multilexical bilingual extraction is done from the parallel corpus using the nat-examplesExtractor script which is part of the NATools package:

(16)

```
nat-examplesExtractor -local=logaliza_en_gl -
rules=./en_gl.cat_patterns -chunk=1 -attraction
```

This script calculates a translation probability measure for each candidate pair of bilingual terminological equivalents. This value is based on the translation probabilities for each word pair, discarding probabilities for stop-words translation. Considering the previous pattern example, A B

C = C "de" B A, the translation probability of this equivalence is measured as the average of the mutual translation probability of the words matching the variables A, B, and C.

Moreover, following many other works on term extraction based on Dunning (1993), the system scores each term candidate with the log-likelihood measure, using the Text::NSP[7] package. The minimum value for the partial trigrams is used for terms with more than three constituents (Patry and Langlais 2005).

This process generates a very useful list of 5,356 different multilexical bilingual term candidates, corresponding to 14,894 parallel tokens. The output format produced as a result is: rule identifier (Id) + translation probability (Prob) + English term candidate (EN) + log-likelihood for the English candidate (EN_LLR) + Galician term candidate (GL) + log-likelihood for the Galician candidate (GL_LLR). This is a selection of the result of the processing, with only one example for each different rule:

(17)

Id	Prob	EN	EN LLR	GL	GL LLR
R0001	58.399	empty string	267.919	cadea baleira	128.618
R0002	28.510	power status	6.942	estado da enerxía	585.402
R0003	69.622	contents in cell	64.824	contido da cela	2797.236
R0004	48.116	complete search path	108.383	camiño de busca completo	1649.495
R0023	48.223	system volume controls	40.956	controis de volume do sistema	259.233
R0005	53.848	error of unknown origin	37.886	erro de orixe descoñecida	833.980
R0021	33.863	read and write privileges	114.732	privilexios de lectura e escritura	74.073
R0006	64.296	special internal memory	36.794	memoria especial interna	24.768

R0031	69.403	changes in size or resolution	14.320	alteracións de tamaño ou resolución	12.168
R0032	75.707	number of rows for array	21.846	número de filas da matriz	257.525
R0022	72.759	access control list size	37.478	tamaño da lista de control de acceso	1420.507
R0051	39.603	patterns of black and white dots	16.274	patróns dos puntos negros e brancos	67.590

Obviously not all the translation patterns used for bilingual terminological extraction have the same potential for term identification. Thus, the two most productive rules give account of 83% of all the lexical types, and the first five more than 98%. The different productivity of the patterns analysed is clearly illustrated by this summary table:

(18)

	Types	**Tokens**
R0002	2775	9115
R0001	1673	4322
R0003	342	624
R0004	253	372
R0023	226	327
R0005	36	49
R0021	18	37
R0006	13	22
R0031	8	11
R0032	7	9
R0022	4	5
R0051	1	1
Total	5356	14894

4. Extraction of term semantic relations and definitory contexts

The terminological work on corpora carried out by our research group at the University of Vigo is focused on the terminographic treatment of the terms identified in the framework of the Termoteca. This terminological database for Galician was created with an onomasiological approach, where concepts are the door to enter the term. This approach leads directly to focus our interest on the description of the conceptual relations between terms and on their definitions (Crespo *et al*. 2008).

With respect to that issue, the work undertaken so far has aimed at the search for linguistic and typographical patterns that can be discovered in corpora, both for semantic relations and definitions. We understand, first, that the semantic relations can be identified by textual segments which function as real anchors and also lead to retrieve textual information relevant to the semantic explicitation of a term, and, second, that when an author of a text defines a term, s/he does so through definitory contexts, considering as definitory context any textual fragment from a document which provides specialised information useful to define a term. All of this information can be automatically retrieved from an untagged corpus or (in a faster and clearer way) from a PoS-tagged corpus. I will show some examples of this kind of terminological information extraction from the annotated version of the LOGALIZA Corpus (tagged with FreeLing) using the perl standard tool for text processing in Unix.

For the identification of semantic relations, we draw on Feliu (2004), because in her work she describes the general framework of conceptual relations that we use in our analysis and she also presents textual markers that identify them in a Catalan corpus (textual markers adapted and supplemented by us for Galician). For definitory contexts we draw on the work done in the Corpógrafo[8] (Pinto and Oliveira 2004), Alarcón (2009), Sierra (2009), and especially on the classic work of Pearson (1998), which explains that when an author wants to define a term in a text, s/he may resort to typographical elements to highlight this term, and to the definition and definitory patterns to relate the term to its definition. In our research we also believe it is interesting to take advantage of any relevant information which, even without being a definition, can be related to semantic aspects of the term.

4.1. Patterns for semantic relations

In a pattern [X p Y] for the automatic extraction of semantic relations, both X and Y are well defined terms within a specific domain (and documented in our terminological database), and "p" is a linguistic pattern that can be formed by verbs, verbal phrases, connectors and typographical elements. Currently, the search pattern is based on [X p] to seek any textual segment including Y. Here is a sample of the linguistic patterns "p" we use to search for semantic relations in the tagged LOGALIZA Corpus expressed as regular expressions:

Antonymy:

```
\w+_ser_V.....(\so_o_DA0MS0)?\s\w+_(contrario|opost
o)_\w+\s\w+_(de|a)_SPS00
```

Hyponym-hypernym:

```
\w+_ser_V.....\s\w+_un_DI0.S0\s\w+_(tipo|clase)_NC...
..\sde_de_SPS00
```

Hypernym-hyponym:

```
,_,_Fc\scomo_como_PR0CN000\s\w+_un_DI0.S0
```

Meronymy:

```
\w+_estar_V.....\s\w+_(compoñer|formar|constituír)_V
MP....\s\w+_(de|por)_SPS00
\w+_constar_V.....\sde_de_SPS00
\w+_(abranguer|englobar)_V.....
```

Instrumentality:

```
\w+_(servir|usar|empregar|utilizar)_V.....\s\w+_(para|co
mo|de|en)_
```

We can see some examples of these semantic relations identified in the tagged LOGALIZA Corpus, followed by their untagged version with X and Y terms in italics and "p" patterns between angle brackets:

Antonymy:

```
As_o_DA0FP0 referencias_referencia_NCFP000
absolutas_absoluto_AQ0FP0 son_ser_VSIP1S0 o_o_DA0MS0
contrario_contrario_NCMS000 de_de_SPS00 o_o_DA0MS0
```

direccionamento_direccionamento_NCMS000 absoluto_absoluto_AQ0MS0 ._._Fp

As *referencias absolutas* <son o contrario do> *direccionamento absoluto*. Nas referencias absolutas colócase un sinal $ antes de cada letra e número nunha referencia absoluta, por exemplo, $A$1:$B$2. [OpenOffice]

Hyponym-hypernym:

Info_info_NP00000 é_ser_VSIP3S0 un_un_DI0MS0 tipo_tipo_NCMS000 de_de_SPS00 documentación_documentación_NCFS000 ._._Fp

Info <é un tipo de> *documentación*. Os documentos están nun formato chamado textinfo, e poden ser lidos na liña de comandos mediante o programa info. [KDE]

Hypernym-hyponym:

O_o_DA0MS0 carácter_carácter_NCMS000 de_de_SPS00 espazo_espazo_NCMS000 suprímese_suprímese_NCFS000 cando_cando_PR0CN000 o_o_DA0MS0 seguinte_seguinte_AQ0CS0 carácter_carácter_NCMS000 é_ser_VSIP3S0 un_un_DI0MS0 delimitador_delimitador_NCMS000 ,_,_Fc como_como_PR0CN000 un_un_DI0MS0 punto_punto_NCMS000 final_final_AQ0CS0 ou_ou_CC un_un_DI0MS0 novo_novo_AQ0MS0 carácter_carácter_NCMS000 de_de_SPS00 liña_liña_NCFS000 ._._Fp

O carácter de espazo engádese despois de teclear o primeiro carácter da palabra seguinte á que se completou de forma automática. O carácter de espazo suprímese cando o seguinte carácter é un *delimitador* <, como un> *punto final* ou un novo *carácter de liña*. [OpenOffice]

Meronymy:

UML_uml_NP00000 está_estar_VMIP3S0 composto_compoñer_VMP00SM de_de_SPS00 diversos_diverso_AQ0MP0 elementos_elemento_NCMP000 de_de_SPS00 modelo_modelo_NCCS000 que_que_PR0CN000 representan_representar_VMIP3P0 as_o_DA0FP0 diferentes_diferente_AQ0CP0 partes_parte_NCCP000

de_de_SPS00 un_un_DI0MS0 sistema_sistema_NCMS000 de_de_SPS00 sóftware_sóftware_NCMS000 ._._Fp

UML <está composto de> diversos *elementos de modelo* que representan as diferentes partes dun sistema de sóftware. Os elementos UML son usados para criar diagramas, que representan unha parte, ou un ponto de vista do sistema. [KDE]

Instrumentality:

As_o_DA0FP0 chaves_chave_NCFP000 primarias_primario_AQ0FP0 serven_servir_VMIP3P0 como_como _PR0CN000 identificadores_identificador_AQ0MP0 únicos_único_AQ0MP0 de_de_SPS00 os_o_DA0MP0 campos_campo_NCMP000 de_de_SPS00 as_o_DA0FP0 bases_base_NCFP000 de_de_SPS00 datos_dato_NCMP000 ._._Fp

As *chaves primarias* <serven como> *identificadores únicos* dos campos das bases de datos. A identificación única dos campos das bases de datos úsase nas bases de datos relacionais para acceder a datos doutras táboas. Cando se fai referencia a unha chave primaria doutra táboa, fálase de chave externa. [OpenOffice]

4.2. Patterns for definitions

In a pattern [X = Y] for the extraction of term definitions, "X" is a term from the terminological database, "=" is a definitory pattern based on verbs, linguistic or metalinguistic phrases (including reformulative markers) and typographical elements, and Y is the definition or the relevant syntactic elements that can lead to the creation of a definition. With regard to Y, it must be clear that it can also be a term that is the superordinate in the sort of classic definition based on gender and difference [X = Y [specific semantic features]]. Currently, the search pattern is based on [X =] to seek any textual segment including Y. Here is a sample of the patterns "p" we use to search for term definitions in the corpus expressed as regular expressions:

—Typographical elements:

:_:_Fd

—Verbal phrases:

\w+_ser_VSIP3.0\s\w+_un_DI0..0

—Reformulative markers:

,_,_Fc\sé_ser_VSIP3S0\sdicir_dicir_VMN0000\s,_,_Fc
,_,_Fc\sisto_isto_PD0CN000\sé_ser_VSIP3S0\s,_,_Fc

Here follow some examples of definitory contexts identified in the tagged LOGALIZA Corpus, followed by their untagged version:

—Typographical elements:

Ecualizar_ecualizar_NP00000 :_:_Fd Esta_este_DD0FS0 opción_opción_NCFS000 axusta_axustar_VMIP3S0 o_o_DA0MS0 brillo_brillo_NCMS000 de_de_SPS00 as_o_DA0FP0 cores_cor_NCFP000 de_de_SPS00 a_o_DA0FS0 imaxe_imaxe_NCFS000 de_de_SPS00 tal_tal_DI0CS0 maneira_maneira_NCFS000 que_que_PR0CN000 o_o_DA0MS0 histograma_histograma_NCMS000 de_de_SPS00 o_o_DA0MS0 canal_canal_NCMS000 de_de_SPS00 valor_valor_NCMS000 resulte_resultar_VMSP1S0 o_o_DA0MS0 máis_máis_RG chao_chao_AQ0MS0 posíbel_posíbel_AQ0CS0 ,_,_Fc isto_isto_PD0CN000 é_ser_VSIP3S0 ,_,_Fc que_que_PR0CN000 cada_cada_DI0NN0 posíbel_posíbel_AQ0CS0 valor_valor_NCMS000 de_de_SPS00 brillo_brillo_NCMS000 apareza_aparecer_VMSP1S0 máis_máis_RG ou_ou_CC menos_menos_RG o_o_DA0MS0 mesmo_mesmo_DI0MS0 número_número_NCMS000 de_de_SPS00 píxeles_píxeles_NP00000 que_que_CS o_o_DA0MS0 resto_resto_NCMS000 de_de_SPS00 valores_valor_NCMP000 ._._Fp

Ecualizar <:> *Esta opción axusta o brillo das cores da imaxe de tal maneira que o histograma do canal de valor resulte o máis chao posíbel, isto é, que cada posíbel valor de brillo apareza máis ou menos o mesmo número de píxeles que o resto de valores*. Ás veces Ecualizar funciona marabillosamente para mellorar os contraste nunha imaxe. Outras veces xera lavaduras. É unha operación moi poderosa, que pode ou facer milagres ou destruir unha imaxe. [KDE]

—Verbal phrases:

Un_un_DI0MS0 protocolo_protocolo_NCMS000 é_ser_VSIP3S0 unha_un_DI0FS0 linguaxe_linguaxe_NCFS000 que_que_PR0CN000 usa_usar_VMIP3S0 o_o_DA0MS0 seu_seu_DP3MS0 computador_computador_NCMS000 para_para_SPS00 se_se_PP3CN000 comunicar_comunicar_VMN0000 con_con_SPS00 outros_outro_DI0MP0 computadores_computador_NCMP000 ._._Fp

Un *protocolo* <é unha> *linguaxe que usa o seu computador para se comunicar con outros computadores*. [Windows]

—Reformulative markers:

As_o_DA0FP0 bases_base_NCFP000 de_de_SPS00 datos_dato_NCMP000 dBase_dbase_NCFS000 e_e_CC texto_texto_NCMS000 restrínxense_restrínxense_AQ0CS0 a_o_DA0FS0 conxuntos_conxunto_NCMP000 de_de_SPS00 caracteres_caracteres_NCMP000 con_con_SPS00 lonxitude_lonxitude_NCFS000 fixa_fixo_AQ0FS0 ,_,_Fc é_ser_VSIP3S0 dicir_dicir_VMN0000 ,_,_Fc con_con_SPS00 todos_todo_DI0MP0 os_o_DA0MP0 caracteres_caracteres_NCMP000 codificados_codificar_VMP00PM con_con_SPS00 o_o_DA0MS0 mesmo_mesmo_DI0MS0 número_número_NCMS000 de_de_SPS00 bytes_byte_NCMP000 ._._Fp

As bases de datos dBase e texto restrínxense a conxuntos de *caracteres con lonxitude fixa* <, é dicir,> *con todos os caracteres codificados co mesmo número de bytes*. [OpenOffice]

Insire_insire_NP00000 o_o_DA0MS0 campo_campo_NCMS000 como_como_PR0CN000 contido_contido_NCMS000 estático_estático_AQ0MS0 ,_,_Fc isto_isto_PD0CN000 é_ser_VSIP3S0 ,_,_Fc non_non_RN se_se_PP3CN000 pode_poder_VMIP3S0 actualizar_actualizar_VMN0000 ._._Fp __F

> Insire o campo como contido estático <, isto é,> non se pode actualizar. [OpenOffice]

It is important to note that the definitory contexts often include a hypernym or a list of meronyms, which is why these patterns can help to extract both definitions and other semantic relations, as can be seen in the following examples:

Definition by meronymy;

> Cambia_cambiar_VMIP3S0 a_o_DA0FS0 aparencia_aparencia_NCFS000 de_de_SPS00 o_o_DA0MS0 escritorio_escritorio_NCMS000 ,_,_Fc isto_isto_PD0CN000 é_ser_VSIP3S0 ,_,_Fc o_o_DA0MS0 fondo_fondo_NCMS000 ,_,_Fc o_o_DA0MS0 protector_protector_NCMS000 de_de_SPS00 pantalla_pantalla_NCFS000 ,_,_Fc as_o_DA0FP0 cores_cor_NCFP000 ,_,_Fc os_o_DA0MP0 tamaños_tamaño_NCMP000 de_de_SPS00 os_o_DA0MP0 tipos_tipo_NCMP000 de_de_SPS00 letra_letra_NCFS000 ,_,_Fc a_o_DA0FS0 resolución_resolución_NCFS000 de_de_SPS00 a_o_DA0FS0 pantalla_pantalla_NCFS000 ,_,_Fc etc_etc_NCCN000 ._._Fp
>
> Cambia a *aparencia do escritorio* <, isto é,> *o fondo, o protector de pantalla, as cores, os tamaños dos tipos de letra, a resolución da pantalla, etc.* [Windows]

Definition by hypernymy:

> Unha_un_DI0FS0 [lista_lista_NCFS000 de_de_SPS00 elementos_elemento_NCMP000] { é_ser_VSIP3S0 unha_un_DI0FS0 } listaxe_listaxe_NCFS000 que_que_PR0CN000 se_se_PP3CN000 utiliza_utilizar_VMIP3S0 cando_cando_PR0CN000 a_o_DA0FS0 orde_orde_NCFS000 de_de_SPS00 os_o_DA0MP0 elementos_elemento_NCMP000 non_non_RN é_ser_VSIP3S0 importante_importante_AQ0CS0 ._._Fp
>
> Unha *lista de elementos* <é> *unha listaxe que se utiliza cando a orde dos elementos non é importante*. [KDE]

5. Conclusions

The exploitation of technical corpora allows to directly observe the linguistic reality shaped in specialised texts, facilitating the empirical analysis of many pragmatic aspects of their terminology that could not be studied in another way without great difficulties (such as the frequency of use of the terms, their semantic potential, their textual dispersion, their dating or their combinatoriness).

Nevertheless, working with corpora imposes certain limitations the terminological research is not free from. First of all, it is necessary to take into account that basing the terminographical work on corpora requires the existence of enough textual material written in the specialised field and in the language you want to study. For example, the textual production in Galician in very recent technical areas or in fields of high speciality, like those of genomics, quantum mechanics or particle acceleration is very limited or almost non-existent, except in those cases in which textual production is impelled by the institutions, so corpus-based terminological research based on those fields is impractical. This limitation is even greater if you want to apply a corpus-based multilingual approach. For example, in Galician language, we have a certain textual production on climate change. However, the English-Galician parallel texts in this field are virtually non-existent. The incorporation of the translation factor limits Galician in almost every specialised field, with the exception of the juridical one in the Galician-Spanish pair, thanks to the existing legal imperative.

Another important limitation derived from corpus-based methodology is that sometimes, by chance or by limitations of the selection of the texts in the corpus, terms that we suspected to be frequent or normal in a particular area of specialty are not documented in the corpus. The reason is that, by the logic of statistics (do not forget that a corpus is a sample of a mostly unknown population), chances are that no corpus will contain all the terms of a field. In order to solve this problem, at least in part, it is possible to increase the size of the corpus and to diversify the thematic and register variety of collected texts, provided that it is possible.

Finally, although the semi-automatic extraction of terminological information from technical corpora complements with great efficiency the human research work, in no way does it make it unnecessary. Any methodology of automatic extraction of terminological information applied to corpora must be complemented by a long period of human labour of appraisal, reflection and decision making from the collected data.

Notes

[1] http://sli.uvigo.es/CLUVI/
[2] http://sli.uvigo.es/termoteca/
[3] http://sli.uvigo.es/CTG/
[4] http://natools.sourceforge.net/
[5] http://www.bivir.com/
[6] http://nlp.lsi.upc.edu/freeling/
[7] http://ngram.sourceforge.net/
[8] http://www.linguateca.pt/corpografo/

References

Alarcón, Rodrigo. "Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios." PhD diss., Universitat Pompeu Fabra, Barcelona, 2009.

Cabré, Teresa. *La terminología: representación y comunicación*. Barcelona: Universitat Pompeu Fabra, 1999.

Crespo-Bastos, Ana, Xosé María Gómez-Clemente, Xavier Gómez-Guinovart, and Susana López-Fernández. "XML-based extraction of terminological information from corpora." *Actas da 6ª Conferência Nacional de XML: Aplicações e Tecnologias Associadas (XATA2008)* (2008): 28-39.

Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics*, 19(1) (1993): 61-74.

Feliu, Judit. "Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica." PhD diss., Universitat Pompeu Fabra, Barcelona. 2004.

Gómez-Clemente, Xosé María, and Xavier Gómez-Guinovart. *Corpus Técnico do Galego*. http://sli.uvigo.es/CTG/. Vigo: Universidade de Vigo, 2006a.

Gómez-Clemente, Xosé María, and Xavier Gómez-Guinovart. *Termoteca (Banco de Datos Terminolóxico da Universidade de Vigo)*. http://sli.uvigo.es/termoteca/. Vigo: Universidade de Vigo, 2006b.

Gómez-Guinovart, Xavier. *Corpus CLUVI (Corpus Lingüístico da Universidade de Vigo)*. http://sli.uvigo.es/CLUVI/. Vigo: Universidade de Vigo, 2003.

Gómez-Guinovart, Xavier, and Elena Sacau-Fontenla. "Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos." *Procesamiento del Lenguaje Natural* 33 (2004a): 133-140.

Gómez-Guinovart, Xavier, and Elena Sacau-Fontenla. "Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo)." *Proceedings of the 4th International Conference on Language Resources and Evaluation* (2004b): 1179-1182.

Gómez-Guinovart, Xavier, and Alberto Simões. "Parallel corpus-based bilingual terminology extraction." *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence* (2009).

Padró, Lluís, Collado, Miquel, Reese, Samuel, Lloberes, Marina, and Irene Castellón. "FreeLing 2.1: five years of open-source language processing tools." *Proceedings of the 7th International Conference on Language Resources and Evaluation* (2010): 931-936.

Patry, Alexandre, and Philippe Langlais. "Corpus-based terminology extraction." *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering* (2005): 313-321.

Pearson, Jennifer. *Terms in context.* Amsterdam: John Benjamins, 1998.

Pinto, Ana Sofia, and Débora Oliveira. *Extracção de definições no Corpógrafo*. Technical report. Porto: Faculdade de Letras da Universidade do Porto, 2004.

Sierra, Gerardo. "Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos." *Linguamática* 1(2) (2009): 13-37.

Simões, Alberto Manuel, and José João Almeida. "NATools: a statistical word aligner workbench." *Procesamiento del Lenguaje Natural* 31 (2003): 217-224.

Simões, Alberto Manuel, and Xavier Gómez-Guinovart. "Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns." *Proceedings of the Iberian SLTech 2009 - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages* (2009): 13-16.

Sinclair, John. *Collins Cobuild English Language Dictionary*. London: Collins, 1987.