XML-based Extraction of Terminological Information from Corpora

Ana Belén Crespo Bastos, Xosé María Gómez Clemente,
Xavier Gómez Guinovart, and Susana López Fernández

Grupo TALG
Tecnoloxías e Aplicacións da Lingua Galega
Universidade de Vigo
{acrespo,xgomez,xgg,susanalopez}@uvigo.es
http://sli.uvigo.es

Abstract. In this paper, we present a methodology for the extraction of terminological information from textual corpora, showing the processes we follow for identification of term candidates in corpora, and for recognition in textual data of term definitions and conceptual relations. Both the textual corpora that are used as the source for terminological information, as well as the terminological database we build from this information, are stored and maintained by linguists in XML format, and converted to MySQL format for consultation through a PHP-based web application.

Key words: Natural language processing, textual corpora, terminological databases, ontologies, information extraction

1 Introduction

In this paper[1] we present a methodology developed at the University of Vigo by the TALG Research Group ("Galician Language Technology and Applications") for the extraction of terminological information from textual corpora leading to the creation of linguistic resources in the field of terminology for Galician. We will explain the main characteristics of the CLUVI corpus and the CTG corpus which constitute the source of this work, the process followed for the preparation of the Terminological Databank of the University of Vigo (TUVI), as well as the results obtained so far and the tasks we are undertaking and the ones we have in prospect.

The Linguistic Corpus of the University of Vigo (CLUVI) is an open collection of textual corpora with translations in specific areas of the contemporary

[1] This work has been funded by the Ministerio de Educación y Ciencia and the Fondo Europeo de Desenvolvemento Rexional (FEDER) within the project "Diseño e implementación de un servidor de recursos integrados para el desarrollo de tecnologías de la lengua gallega (RILG)" (HUM2006-11125-C02-01/FILO), a coordinated project between the University of Vigo (TALG Research Group) and the University of Santiago de Compostela (Instituto da Lingua Galega).

Galician language, accessible in the web since September 2003 at `http://sli.uvigo.es/CLUVI`. With a current total length exceeding the 20 million words, the CLUVI comprises six main parallel corpora belonging to four specialized 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) [4]. The format chosen for storing the aligned parallel texts is an adaptation of the TMX format (Translation Memory eXchange), as this is the XML encoding standard for translation memories and parallel corpora, regardless of the application used [9]. A translation memory is a database which holds the original and translated version for each of the sentences translated in the framework of a computer-aided translation system, with the aim to reuse 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 [10].

The Galician Technical Corpus (CTG), available since 2006 for free consultation at `http://sli.uvigo.es/CTG`, is an open monolingual corpus of contemporary specialized Galician, in the fields of law, computing, economics, environmental science, sociology and medicine, with a current extension of 12 million words. The CTG is stored in the XML format, annotated with bibliographic and thematic information, and segmented into sentences. The web application developed in PHP for the searching and browsing of the CTG permits to query words or groups of words, use wildcards looking for complex patterns (regular expressions), and specify the subset of the corpus to which you want to limit the search. At present, the CTG is being annotated with information about the lemma and part-of-speech of words.

The Terminological Databank of the University of Vigo (TUVI) is a terminological database based on the monolingual and parallel specialty texts collected in the corpora of the University of Vigo, namely in the Linguistic Corpus of the University of Vigo (CLUVI) and in the Galician Technical Corpus (CTG). This terminological database is freely accessible on the web at `http://sli.uvigo.es/TUVI`, and currently has 5,625 terms documented in the CLUVI and CTG corpora belonging to the areas of law (1,411 bilingual and monolingual entries), sociology (954 tetralingual and monolingual entries), economy (1,163 monolingual entries) and ecology (1,324 monolingual entries). All terms in the TUVI are documented in the corpora, the terminological inventories in the fields of computer science and medicine being in progress.

In the TUVI, terminological information is structured around concepts. Each TUVI record includes all the information relating to a concept expressed with a Galician term which can be recorded also with variants, both intralinguistic (synonymic 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 matter. Also, the concepts in the database form a navigateable lexical-semantic network where conceptual nodes interact with each other according to the semantic relations (antonymy, hiperonymy, holonymy, etc.) among them.

Both the textual corpora that are used as the source for terminological information, as well as the terminological database we build from this information, are stored and maintained by linguists in XML format, and converted to MySQL format for consultation through a PHP-based web application, which allows for significantly expedite processing. XML to MySQL conversion is done in two ways, depending on the original XML document. For these textual corpora, which have a relatively simple structure, we created an ad hoc database with two related tables, one for the source texts and another for the sentences, and the data is imported from two delimited texts generated through XSL. For the termbase, which has a complex structure branching at various levels, we use Altova XMLSpy in order to export XML as delimited text. Altova XMLSpy converts XML input in ten interrelated tables which are imported from a MySQL database created from the output of XMLSpy converter. Terminological information in the TUVI database is structured according to the following DTD:

<!ELEMENT dic (cc+)> <!−−a dictionary is a set of concepts−−>
<!ELEMENT cc (ic, rs*, def*, ct+, lg+)>
<!ELEMENT ic (♯ PCDATA)> <!−−ic: concept index−−>
<!ELEMENT rs (♯ PCDATA)> <!−−rs: semantic relations−−>
<!ATTLIST rs <!−−tipo-rs: set of semantic relations−−>
tipo-rs (hipo | hiper | ant | mero | holo | eant | epost | tant | tsim | tpost
| axente | prod | caus | efec | instr | fin) ♯ REQUIRED >
<!ELEMENT def (texto_def, fonte_def?)> <!−−def: definition−−>
<!ATTLIST def xml:lang (gl | es | en | fr | pt) ♯ REQUIRED >
<!ELEMENT texto_def (♯ PCDATA)>
<!ELEMENT fonte_def (♯ PCDATA)>
<!ELEMENT ct (♯ PCDATA)> <!−−ct: thematic field−−>
<!ATTLIST ct st CDATA ♯ REQUIRED > <!−−st: standard for classification−−>
<!ELEMENT lg (var+)> <!−−lg: language-specific information−−>
<!ATTLIST lg
xml:lang (gl | es | en | fr | pt) ♯ REQUIRED >
<!ELEMENT var (lema, pms?, cat, ex*, frec+)>
<!ATTLIST var <!−−var: linguistic variant−−>
tipo (com | orto | morf | sigla | acro) ♯ REQUIRED
<!ELEMENT lema (♯ PCDATA)>
<!ELEMENT cat EMPTY> <!−−cat: grammatical category−−>
<!ATTLIST cat
valor (m | f | s | com | adx | lconx | lprep | ladx | lnom | lvb | ladv |

ladvlat | vt | vi | mpl | fpl | spl | compl) ♯ REQUIRED >
<!ELEMENT pms (♯ PCDATA)> <!−−morphosyntactic pattern−−>
<!ELEMENT ex (texto_ex, fonte_ex)> <!−−ex: term in context−−>
<!ELEMENT texto_ex (♯ PCDATA)>
<!ELEMENT fonte_ex (obra, num?)>
<!ELEMENT num (♯ PCDATA)>
<!ELEMENT obra (♯ PCDATA)>
<!ELEMENT frec (fab, vcorpus, palcorpus)> <!−−frec: term relative frequency in the corpus−−>
<!ELEMENT fab (♯ PCDATA)> <!−−fab: absolute frequency−−>
<!ELEMENT vcorpus (♯ PCDATA)> <!−−vcorpus: corpus version−−>
<!ELEMENT palcorpus (♯ PCDATA)> <!−−palcorpus: corpus size−−>

2 Extraction of term candidates

Next, we will explain our methodology for the extraction of term candidates from corpora, focusing on the AUGA corpus, a subset of CTG corpus devoted to language of ecology and environmental sciences, consisting of 2,349,362 words of journalistic, legislative, academic and informative texts. The texts in the AUGA corpus, as a whole, are about different themes on relations between human beings and nature, including the study of environmental problems and models for sustainable development.

2.1 Wordgrams-based extraction

First, we extract the most frequent words of the corpus, as well as the most frequent sequences of n-words (wordgrams), taking into account sequences until 4 words (bigrams, trigrams and tetragrams). Below we include some examples of terms identified in this way, with the frequency with which they occur in the corpus analyzed:

- *ambiental* (5,216 times)
- *ambiente* (3,807 times)
- *especies* (2,464 times)
- *contaminación* (1,658 times)
- *impacto ambiental* (903 times)
- *educación ambiental* (527 times)
- *augas residuais* (516 times)
- *avaliación ambiental* (508 times)
- *residuos perigosos* (499 times)
- *xestión de residuos* (455 times)
- *calidade do aire* (274 times)
- *plan de xestión* (223 times)
- *organismos modificados xeneticamente* (214 times)
- *autorización ambiental integrada* (213 times)

- *avaliación de impacto ambiental* (247 times)
- *gases de efecto invernadoiro* (220 times)
- *estudo de impacto ambiental* (166 times)
- *declaración de impacto ambiental* (160 times)

2.2 Extracting low frequency terms

The wordgrams-based terminology extraction is completed with the consultation of authoritative literature on the subject, which allows us, among other things, to identify key terms for the domain that have a low frequency in the corpus. In the case of environmental terms, the reference work for the Galician language was the Léxico do medio (`http://www.linmiter.net/lexique/_index.html`), elaborated by the União Latina (`http://www.unilat.org`) in the European project "Linmiter" (`http://www.linmiter.net`), a project created as a tool to support the terminology of minority Latin languages. Thus, we have completed our wordgrams-based inventory with the terms of the Léxico do medio which are equally documented in our corpus but which were not included in our initial list. These are new terms which, despite having a low presence and frequency in the body, are important terms in the field.

2.3 Extracting morphosyntactic patterns

As we said earlier, the CTG corpus is being tagged with information about lemma and part-of-speech of words. The tagset used in CTG is based on the tagset proposed by the Eagles group [5] for the annotation of morphosyntactic lexicons and corpora for all European languages. Here follows, by way of example, a fragment extracted from the CTG corpus in its PoS-tagged version and in its untagged version:[2]

<s>Galicia é a primeira Comunidade Autónoma pesqueira do Estado español, o sector pesqueiro representa o 8% do PIB e o 5% da poboación activa, estas cifras a pesar de estar en consonancia coa importancia do litoral a nivel mundial, o 40% da poboación do mundo vive nas zonas costeiras, presenta unhas cifras moi por enriba de calquera dos outros países comunitarios.</s>

<s><t w="Galicia" c="NP00000">Galicia</t> <t w="ser" c="VIP3-S00">é</t> <t w="o" c= "AFS">a</t> <t w="primeiro" c="NO0-FS">primeira</t> <t w="Comunidade" c="NCFS000">Comunida-de</t> <t w="Autónomo" c="A0FS0">Autónoma</t> <t w="pes-queira" c="A0FS0">pesqueira</t> <t w="de" c ="SPS00">do</t> <t w="o" c="AMS">∼</t> <t w ="Estado" c="NCMS000">Esta-do</t> <t w="español" c="A0MS0">español</t> [...] </s>

[2] Textual segment included in the CTG corpus belonging to the doctoral dissertation of Alfredo López Fernández, *Estatus dos pequenos cetáceos da plataforma de Galicia* (University of Santiago de Compostela, 2003), directed by Angel Guerra Sierra and Graham J. Pierce.

From this tagged corpus we can retrieve new term candidates based on the morphosyntactic patterns most frequently found in the terminological database built until that moment. Thus, we first determine which are the most common tag combinations in the terminological inventory done so far, and afterwards we observe the sequences of tokens in the corpus that correspond to these patterns. For example, this is the list of morphosyntactic patterns most frequently found in the inventory of 1,444 terms extracted from the AUGA corpus on environmental sciences (we indicate the number of times the term occurs in the list with the specific morphosyntactic pattern):

- Singular feminine noun (216 times)
- Singular masculine noun (209 times)
- Singular feminine noun + singular feminine adjective (157 times)
- Singular masculine noun + singular masculine adjective (145 times)
- Singular feminine noun + singular common adjective (98 times)
- Singular masculine noun + singular common adjective (97 times)
- Singular masculine noun + preposition + singular feminine noun (66 times)
- Singular masculine adjective (52 times)
- Singular feminine noun + preposition + singular feminine noun (41 times)
- Singular common adjective (34 times)
- Singular masculine noun + preposition + singular masculine noun (33 times)
- Verb (21 times)

Comparing the results of the most common morphosyntactic patterns for environmental terms with the patterns identified in the list of 1,768 legal terms extracted from GALEX subcorpus of legal texts (2,516,846 words from CTG), it can be seen that the frequency of patterns is similar. Thus, the most popular patterns are, on the one hand, nouns and noun+adjective combinations and, on the other hand, noun+preposition+noun combinations.

We apply these results to the search for term candidates in the tagged section of the CTG corpus, grouping subcategories such as gender and number of nouns, adjectives and articles, and discarding monocategory patterns. Thus, we obtain a comprehensive list of term candidates, with their frequency in the corpus, from which we extract the valid terms after a thorough review. The resulting list of candidates contains combinations from 2 to 6 elements, for example:

- noun + adjective: *diversidade biolóxica, catástrofe ecolóxica, auditoría ambiental, política forestal, cambio climático, augas residuais*
- noun + preposition + noun: *planta de tratamento, capa de ozono, dióxido de carbono*
- noun + preposition + article + noun: *calidade da auga*
- noun + preposition + article + noun + adjective: *avaliación do impacto ambiental*
- noun + adjective + adjective: *recursos naturais renovables, autorización ambiental integrada*
- noun + preposition + noun + conjunction + noun: *centro de recollida e descontaminación*

– noun + preposition + article + noun + preposition + noun: *saturación do proceso de cambio*

We are currently working on the process of filtering the data automatically extracted from the corpus to obtain term candidates based on morphosyntactic patterns. For the moment, in this process of filtering we have applied two complementary approaches: human testing of data, and verification of the terms in a corpus other than the corpus which is used as source of the data.

Of course, the most reliable approach is the human revision of data by Galician terminologists and specialists in the field. However, due to limited resources and the difficulty of finding experts willing to collaborate, we have to opt for other less efficient methods.

A variant of the second approach is to verify the presence of the term candidates in other subcorpora of CTG different from the subcorpus used as source. For example, during the identification of environmental terms extracted from the AUGA corpus, if a term candidate does not appear or has a very limited presence in other subcorpora from other thematic areas, we believe it has a good chance of being a significant and necessary word in the environmental sciences. In the opposite case, the term candidate increases the likelihood of not being a specific term of its field, because it would also have visibility in other fields like law, sociology, medicine, economics, etc.

Another different variant of the second approach is to search the term candidates in the internet. If its presence rate is not very high, it will be an index of the specificity of the term and, therefore, an index of its terminological relevance.

3 Extraction of semantic relations and definitions

The work on corpora being done by our research group allows us, as we said, both to extract lexical units with specialized value (terms) and their terminographic treatment which focuses on the description of the name and concept. With respect to that issue, the work undertaken so far has focused on two aspects: the identification of conceptual relations, and the linguistic expression of concepts (definitions). These two aspects are the subject of much attention by applied and theoretical terminology, because they point very clearly to the true role of scientific terms in texts: transmission of knowledge.

The TUVI terminological database was created with an onomasiological approach where concepts are "the door" to enter the term. This approach leads directly to focus our interest in the description of the conceptual relations and definitions, without forgetting the necessary adscription of terms to a specific branch of the conceptual tree.

Apart from the traditional and necessary identification of semantic relations and the development of definitions (through the use of thesaurus and ontologies, the systematic search in specialized dictionaries, and the consultation with specialists), we began the work of the automation process, focused on the search for linguistic and typographical patterns that can be discovered in corpora, both

for semantic relations [3] and definitions [6, 1]. We understand that the semantic relations can be identified by textual segments that function as real anchors, which are also segments that lead to retrieve textual information relevant to the semantic explicitation of a term, and that when an author of a text defines a term, she or he does so through definitory contexts, considering as definitory context any textual fragment from a document which provides specialized information useful to define a term [2]. All of this information can be automatically retrieved from a PoS-tagged corpus (in a faster and clearer way) or from an untagged corpus. The following data was extracted from the untagged version of the AUGA corpus on ecology and environmental sciences.

3.1 Methodology for semantic information extraction

For the identification of semantic relations, we draw on [3], because in her paper 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, 7], Alarcón [1] and especially in the classic work of Pearson [6], which explains that when an author, in a given text, wants to define a term, she or 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.

3.2 Patterns for semantic relations

In a pattern [X p Y] for the automatic extraction of semantic relations, both X and Y are (inflected) terms well defined 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 are a sample of the the linguistics patterns "p" we use to search for semantic relations in the corpus expressed as regular expressions:

- Resemblance:
  - Partial resemblance: *(é parecid[oa]* a | *son parecid[oa]s* a *)*
  - Antonymy: *((é|son) o contrario de* | *é contrario* a | *se opón* a | *oponse* a | *se opoñen* a | *opóñense* a | *distínguen?se de* | *diferéncian?se de* | *se distinguen? de* | *se diferencian? de)*
- Inclusion:
  - Hyponym-hyperonym: *((é|son) un tipo de* | *é (un|unha)* | *considéran?se* | *se consideran?)*
  - Hyperonym-hyponym: *(agrupa* a | *como:* | *tal como:?* | *como [oa]s?)*
- Sequentiality:

  - General space: *(aparecen? en | ocorren? en | realízan?se en | se realizan? en | están? situad[oa]s? en | orixínan?se en | se orixinan? en | (ten|teñen) lugar | localízan?se en | se localizan? en | d[áa]n?se en | se d[áa]n? en | atópan?se en | se atopan? en | (están? |)presentes? en)*
  - Front/posterior space: *(están? (situad[oa]s? |)(antes de|despois de|detrás de|diante de))*
  - Previous/simultaneous/posterior time: *(seguen? | iníciase en | prodúcese antes de | prodúcese despois de | prodúcese durante | se inicia en | se produce antes de | se produce despois de | se produce durante| ten lugar antes de)*
- Instrumentality: *(serven? (de|para|como) | úsan?se (de|en|para|como) | se usan? (de|en|para|como) | emprégan?se (de|en|para|como) | se empregan? (de|en|para|como) | utilízan?se (en|para|como) se utilizan? (en|para|como) | son empregad[oa]s (en|para|como) | son utilizad[oa]s (en|para|como) | é empregad[oa] (en|para|como) | é utilizad[oa] (en|para|como) | realízan?se (con|mediante|por medio de))*
- Causality: *(orixinan? | causan? | é causa de | son causa| é a causa de | son a causa de | provocan? | contribúen? a | d[aá]n? lugar a | implican? | producen? | son provocad[oa]s por | é provocad[oa] por)*
- Meronymy: *(están? compost[oa]s? (por|de) | constan? de | están? constituíd[oa]s? por | abranguen? | engloban? | están? formad[oa]s? por)*

We can see some examples of these semantic relations identified in the AUGA corpus:

- Resemblance:
  - *Quercus suber* [...] O seu aspecto é parecido ó da *aciñeira*, aínda que se diferencia dela pola súa grosa e esponxosa casca de máis de 15 cm de grosor, chamada cortiza, e polas súas follas menos espiñentas cá desta.
- Sequentiality:
  - Ata a actualidade a *avaliación da calidade do aire* realízase puntualmente nos *lugares de medición*, sen que exista un coñecemento preciso da representatividade territorial das medicións obtidas.
- Causality:
  - Nos solos non agrícolas, a *acidificación* dá lugar á *perda de vitalidade* das plantas producindo a perda e deterioración de follas e en último caso a morte das especies vexetais acompañada de cambios nos organismos do solo, ao favorecer a proliferación de especies acidófilas.
- Meronymy:
  - O *biogás* está constituído na súa meirande parte por *dióxido de carbono* e *metano*, ademais tamén posúe pequenas cantidades de hidróxeno e sulfuro de hidróxeno.

3.3 Patterns for definitions

In a pattern [X = Y] for the extraction of term definitions, "X" is a term from the 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 the gender and the difference [X = Y [specific semantic features]]. Currently, the search pattern is based on [X =] to seek any textual segment including Y. Here are a sample of the patterns "p" we use to search for term definitions in the corpus expressed as regular expressions:

- Verbs: *(é | son | (concíbe|enténde|considéra)n?se | se (concibe|entende|considera)n? | poden? (concibir|entender|considerar)se | se poden? (concibir|entender|considerar) | póden?se (concibir|entender|considerar) | poden? ser (concibid|entendid|considerad)[oa]s?)*
- Reformulative markers: *(, isto é|, é dicir)*
- Linguistic expressions: *(,? como:? |,? tal como)*
- Typographical elements: *(:)* // colon sign

Here follow some examples of definitory contexts identified in the AUGA corpus:

- *A acuicultura* defínese como *o conxunto de actividades encamiñadas ao cultivo de especies acuáticas.*
- *Aire ambiente*: *o aire troposférico e exterior.*
- No Regulamento (CE) nº 2792/1999 do Consello, de 17 de decembro de 1999, polo que se definen as modalidades e condicións das intervencións coa finalidade estrutural no sector da pesca, queda recollida a definición de *acuicultura* como: *a cría ou cultivo de organismos acuáticos con técnicas encamiñadas a aumentar, por encima das capacidades naturais do medio, a produción dos organismos en cuestión; estes serán, ao longo de toda a fase de cría ou de cultivo e ata o momento da súa recollida, propiedade dunha persoa física ou xurídica.*
- *O dióxido de carbono* é *o principal responsable da contribución humana ao efecto invernadoiro a través do uso de combustibles fósiles.* [semantic features]
- *A agricultura ecolóxica* é *unha manifestación da recente preocupación da poboación polo medio natural e o consumo de produtos saudables.* [semantic features]
- *A avaliación de impacto ambiental* é *un proceso de análise mediante o que se integra o medio ambiente e o proxecto deseñado, oferecendo unha serie de vantaxes a ambos aínda que en moitas ocasións estas só son evidentes a longo prazo e que poden permitir aforros nos investimentos e os custos das obras, deseños mais aperfeizoados e integrados no entorno e maior aceitación social dos mesmos.*

3.4 Results and further work

Below we include some results on the accuracy of the patterns used in the extraction of definitions, where "DCIP" stands for "number of different definitory con-

texts identified by the pattern"; "precise" means "number of definitory contexts identified by the pattern that presents a definition"; "relevant" means "number of definitory contexts identified by the pattern that presents an information relevant to a possible definition"; and "irrelevant", "error in the identification of the definitory context".

term	**DCIP**	**precise**	**relevant**	**irrelevant**
acuicultura	6	3	3	0
aeroxerador	12	0	1	11
agricultura ecolóxica	4	0	4	0
aire	15	0	5	10
aire ambiente	4	2	0	2
aleta dorsal	4	0	4	0
ambiente	13	0	7	6
amianto	1	0	1	0
amoniaco	6	0	6	0
amonio	2	0	1	1
avaliación de impacto ambiental	6	1	1	4
dióxido de carbono	5	0	3	2
emisión	7	3	1	3
medio acuático	1	0	1	0
medio natural	8	0	5	3

In short, in 94 DCIP, there are 9 (9.5%) precise definitions, 43 (45.7%) contexts with relevant information, and 42 (44.6%) contexts with irrelevant information. After reviewing the data, we can conclude that:

a) the typographical pattern (:) produces an especially large amount of noise due to the peculiarities of use of this punctuation mark. However, it also identifies precise specific information.

b) The number of precise definitions is low (9.5%, which leads us to redefine and complete the patterns used in extraction.

c) Nevertheless, the relevant information (semantic features that contribute to the understanding of the concept and allow the drafting of definitions) found with the patterns used is high (45.7%), which shows that the automatic extraction is fully justified.

d) As a whole, the percentage of the retrieved semantic information is higher than the noise.

From now on, we will work on a very important aspect in the identification of semantic relations and definitions: the elimination of the "noise" that occurs when linguistic patterns are applied to a corpus. In this regard,

a) we must develop rules for exceptions that can rely on the introduction of linguistic elements that deny the validity of a pattern. The presence of adverbs like "non", "nunca" or phrases like "en ningún caso", will serve to adjust the searches.

b) For the extraction of semantic relations, we must create a specific subcorpus in which the two terms necessary for the establishment of the semantic

relations (X and Y) are clearly identified. Currently, the search pattern is based on X to seek any textual segment (including Y). If we limit the contexts to those where X and Y occur, the noise level will necessarily lessen.

4 Conclusions

The CLUVI corpus and the CTG corpus allow retrieval of linguistic information that facilitates studies on pragmatic aspects of the Galician language. With regard to the terminological treatment of units drawn from the CTG, it must be said that the TUVI terminological database allows a full approximation to the term (usage, denominative information and conceptual information). Automated extraction greatly facilitates identification of term candidates, of conceptual relations and of definitions in large amounts of text, and helps the eventual manual extraction. For the moment, the patterns that retrieve conceptual information (as seen for definitions) reflect a fairly high level of irrelevant information, but they are still very useful to describe the concept.

References

1. Alarcón, R.: Extracción automática de contextos definitorios en corpus anotados. Seminaris de l'IUULATERM, Universitat Pompeu Fabra, Barcelona (2006). <http://www.iula.upf.edu/materials/060526alarcon.pdf>
2. Alarcón, R., Sierra, G.: Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios. Encuentro Nacional de Computación (ENC 2006), San Luis Potosí, México (2006). <http://ccc.inaoep.mx/~tec_lenguaje06/articulos/TLH06-paper3.pdf>
3. Feliu, J.: Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona (2004)
4. Gómez Guinovart, X., Sacau Fontenla, E.: Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo. In: Lino, T. et al. (eds.), Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004, pp. 1179-1182. Lisboa (2004)
5. Leech, G., Wilson. A.: Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Guidelines (1996). <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>
6. Pearson, J.: Terms in Context. John Benjamins, Amsterdam (1998)
7. Pinto, A.S.: Neurodemo: um exemplo de extracção semi-automática de definições e relações semânticas usando o Corpógrafo, Linguateca (2006). <http://poloclup.linguateca.pt/Neurodemo.htm>
8. Pinto, A.S., Oliveira, D.: Extracção de definições no Corpógrafo, Linguateca (2004). <http://www.linguateca.pt/documentos/OliveiraPintoOut2004.pdf>
9. Savourel, Y. (ed.): TMX 1.4b Specification. Localisation Industry Standards Association (2005). <http://www.lisa.org/standards/tmx/specification.html>
10. Simões, A., Dias de Almeida, J.J., Gómez Guinovart, X.: Memórias de Tradução Distribuídas. In: Ramalho, J.C., Simões, A. (eds.), XATA2004 - XML, Aplicações e Tecnologias Associadas, pp. 59-68. Universidade do Porto, Porto (2004).