[*Translating and the Computer* 29, November 2007]

A Free Terminology Extraction Suite

Antoni Oliver and Mercè Vàzquez
Universitat Oberta de Catalunya
{aoliverg, mvazquezga}@uoc.edu

November 5, 2007

Abstract

In this paper we will present a set of terminology extraction tools that are distributed under a Free Software License, so that users can freely download, use, distribute and modify them to meet their needs. The tools are mainly programmed in Perl and they will work under different platforms, such as Windows or Linux. These terminology extraction tools will help freelance translators, translation agencies and companies to find the best translation of a term or to build monolingual or multilingual terminological glossaries. Moreover, translators, correctors and terminologists can use The Free Terminology Extraction Suite to create a terminological database for a specialist domain so as to automatically obtain a list of domain-specific lexical units (potential terms) with their equivalent translations from bilingual corpora of domain-specific documents.

1 Introduction

Terminology is a key component of many specialist documentation generation processes and the translation processes of this documentation. Having good terminological compilations, be they monolingual or multilingual, is absolutely essential for many organisations.

The usual consultation sources, such as dictionaries, specialist websites or own compilations do not always compile all the terms that we need. For this reason, it is very interesting to have tools that can help create terminological compilations based on a set of texts. If this set of texts is monolingual, we can construct lists of candidate terms in one language. Following a manual revision by a specialist, some of these candidates may be added to our terminological compilations. If we have texts in more than one language, we can also construct bilingual or multilingual terminological resources as the tools enable the automatic detection of the possible translation equivalents of a candidate. The automatic translation equivalents search is highly effective if we have aligned parallel corpora.

In this article, we present a set of free terminology extraction tools developed by the LPG research group *(Language Processing Group)* and by the Language Service of the Universitat Oberta de Catalunya. These tools are published with a GNU/GPL licence[1], which permits their free and gratis use by any translator, business or organisation of any kind. The tools can be downloaded at www.linguoc.cat.

2 Automatic terminology extraction: theoretical hypotheses

The theoretical hypotheses that we are presenting in this article have their basis in terminology extraction methods, information retrieval and information management.

Firstly, we focus on the linguistic aspect of terminology, specifically proposing the basic foundations of the Communicative Theory of Terminology (section 2.1). We then introduce the specialist meaning unit (section 2.2) and the techniques that are currently being used to extract these units in a specialist knowledge field (section 2.3). Finally, considering that the terminology extraction field is closely linked to information retrieval and information management, we describe, on the one hand, general concepts on information retrieval (section 2.4) and, on the other, the close relationship that exists between the use of specialist information and terminology extraction (section 2.5).

2.1 The linguistic aspect of terminology

As a systematic subject and organised practice, modern terminology was created in Vienna in the 1930s thanks to the work of Eugen Wüster. The reasons that led Wüster to explore the field of terminology were purely practical, i.e. to overcome the obstacles to professional communication caused by the inexactness, diversification and polysemy of natural language. Wüster regarded terminology as a work instrument that should serve efficiently to disambiguate scientific and technical communication. Initially, his concern was basically methodological and regulatory rather than theoretical. His interest in the theory would come later as a result of the reflection on his work process in the production of his dictionary. His posthumous work of 1979 features the compendium of his theory, called the General Theory of Terminology (GTT), which would subsequently be developed by members of the Vienna school. Today, Wüster is recognised as the creator of GTT and the founder of modern terminology.

In recent years, criticism has begun to appear regarding the principles of GTT, focusing on its lack of capability to explain globally specialist commu-

[1] http://www.gnu.org/copyleft/gpl.html

nication and its more representative units – the terms – and also to describe the terminological varieties in all their representative and formal complexity. The criticism that has been made of GTT refers to three aspects of the terminology that constitute the foundations of its interdisciplinary nature: the cognitive, linguistic and social aspects.

Following the contribution by Cabré (1999) made in the Communicative Theory of Terminology, which looks for new foundations that shed light on a new theory of terms based on the foundations of language and on its socio-cultural nature and which aims to see terms as both unique units and units resembling other communication units within a global scheme representing reality, accepting conceptual and denominative variation and taking the textual and discursive dimension of the terms into account, we assume a series of theoretical hypotheses in relation to the terminology and its subject of analysis, which we summarise as follows:

1. Terminology is a subject that has an intrinsically interdisciplinary nature, which receives contributions from a theory of language, which includes linguistic, cognitive and social aspects; a theory of communication and a theory of knowledge.

2. The object of study are terminological units per se. The nature of term of these units is activated according to the use that is made of them in a specific context and a specific situation.

3. The terms are lexical units that have form and meaning. The form is constant, but the meaning varies according to the type of situation and field in which they are found.

4. The value of a term is established by the place it occupies in the conceptual structure of a subject. Terms do not belong to a field but are used in a field with a uniquely specific value.

5. The objective of applied terminology is to compile units of terminological value in a subject and establish their characteristics.

6. The applied aim of the compilation and analysis of the units of terminological value used in a field is highly diverse.

2.2 From the term to the specialist meaning unit

The set of specialist words of a specific discipline (or specific activity domain) constitute the terminology of this speciality. The terms, which are the base units of the terminology, are sign, distinctive and meaning units that name the concepts of each specialist discipline.

Without exception, all the terms are associated with a basic grammatical category and just one, which is only nominal in a conception of exclusively denominative terminology. Given all this, the “name” category may apply to other categories of verbal or adjectival origin. If, instead of identifying itself by its denominative capacity, terminology is defined by its meaning specificity (the meaning of the field in which it is used) and pragmatic specificity (communicative situation), it extends its identification beyond the nominal units and overlaps with other types of units, such as phraseology or specialist expressions (Cabré, 1999).

In our approach, apart from the denominative conception, we consider the meaning and pragmatic aspects of a unit essential to determine whether it is terminological or not, as often the use to which a unit is put in a specialist context reveals its terminological nature. Consequently, as well as the terms in their denominative aspect, we consider that there are units which, in terms of the meaning and the use that they have in a given context, also have a specialist meaning. This is because specialist meaning units (Estopà, 1999) go beyond the term understood as a classical concept in the sense that they convey the specialist knowledge of a specific speciality and may refer to both linguistic units and non-linguistic units. Linguistic units can be lexical -- nominal, adjectival, verbal, adverbial – or non-lexical – specialist phraseology units (verbal, nominal, adjectival, adverbial) or recurring combinations (descriptive) – and the non-linguistic units may be symbols or formulas. In this classification of specialist meaning units, the terms in the classical sense are situated within the framework of nominal lexical linguistic units, in other words, the terms are considered a sub-set of the specialist meaning units.

2.3 Specialist meaning unit extraction techniques

Today, specialist meaning unit extraction techniques use different methods to achieve the objective of obtaining the most representative units from a speciality corpus. We will now briefly describe what these methods are, classed according to whether they use information from the term itself to extract the units *(endogenous methods)* or external information to the term *(exogenous methods).*

2.3.1 Endogenous methods for specialist meaning unit extraction

The endogenous methods that use the information from the speciality corpus to extract the specialist meaning units are the statistical methods, linguistic methods and hybrid methods.

Statistical methods

Statistical methods recognise terminological units on the basis of their frequency in a specialist corpus. Despite being a very simple calculation, the problem presented is that it does not allow terms that appear only a few times in a speciality corpus to be retrieved. This shortfall can be resolved by creating linguistic filters or by statistical measures (Daille, 1995).

Other techniques that use statistical methods focus on measuring the degree of association that there is between some of their components. To find out the degree of association that there is between the components of a candidate term, statistical calculations are taken, which vary between simple frequencies to more complex measures.

Thus, to obtain better results in specialist meaning unit extraction, the statistical methods allow the use of lists of functional or empty words (stop-words) – articles, pronouns, prepositions, conjunctions, etc. – to prevent there being a non content word at the start or end of the candidate term, and also the use of measurements of association between the elements of a multiword unit to be able to extract only the candidates that have the greatest probability of being candidate terms by degree of association, such as the Log-likelihood ratio, Pearson's Chi-square test, the Odds ratio, the PHI coefficient, the T-score measurement, the Dice coefficient, the Mutual information measurement, et al.

If one is working with a small corpus, this type of method generates a lot of silence or a number of unrecognised terms out of the total terms present in a text. If the corpus is large, there is always a number of terms that, due to their low frequency, cannot be retrieved. They also generate noise, i.e. they retrieve candidate terms that have no terminological value. This is due to the fact that in specialist texts there are also words with non-specialist meaning that form part of the general language and that appear there with a high frequency.

Linguistic methods

Linguistic methods use linguistic knowledge to recognise terms: lexicographical resources, such as dictionaries of terms or dictionaries of auxiliary words -- Fastr (Jacquemin, 1999), Ana (Enguehard and Pantera, 1994); morphological resources, such as internal structure models of the word -Terms (Justeson and Katz, 1995); morphosyntactic resources, such as morphosyntactic models – Term (David and Plante, 1991) –, elements that mark the border of the terminological unit – Lexter (Bourigault, 1994) – or syntactic functions -- Nodalida (Arppe, 1995). And, sporadically, semantic resources, such as semantic classification, and pragmatic resources, such as typographical representations or information of term disposition in the text – Drouin (Drouin, 1997) –, among others.

Generally speaking, these methods generate a lot of noise, i.e. they propose many term candidates that then have to be revised manually, and they also generate silence as they do not detect all the term candidate units, either because these correspond to morphological models that have not been gathered due to problems in the disambiguation process or due to deficiencies in the system itself. In addition, due to the type of knowledge that they use, these methods are only applicable to one language. To transfer them to another language, a prior linguistic study needs to be conducted.

Hybrid methods

The use of methods that combine linguistic and statistical techniques allow the confirmation or rejection of the status of term of a linguistic unit. Statistical techniques provide information in relation to the use of the words, which supplements the pragmatic competence that a specialist has of a term.

In this type of method, the order of application of the type of knowledge is important as the results that are obtained are different. The methods that apply statistical knowledge first and then linguistic knowledge have problems of silence, as occurs with linguistic methods – Drouin, (Drouin, 1997). By contrast, if statistical knowledge is used as a complement to linguistic knowledge, the final result is better – Acabit (Daille, 1995), Clarit (Evans and Zhai, 1996).

Some of the systems that are based on the combination of these techniques are Naulleau (Naulleau, 1998), which employs user profiles to be able to extract the candidates that meet each user's needs and incorporates semantic information, and Trucks (Maynard, 1999), which combines statistical measurements with linguistic information (morphological and semantic) and uses contextual information.

2.3.2 Exogenous methods for specialist meaning unit extraction

The role of the exogenous properties of the term is key to identifying the degree that a word has to be a term candidate, especially when combined with the frequency. Exogenous methods can be used for information about the term, semantic strategies or even a contrast corpus for extracting the specialist meaning units.

Methods that search for information about the term

The measurement of similarity is used to observe what the exogenous properties of the term are in the syntactical structure framework in which it is found, so that the list of terms of a specialist corpus can be classified more exactly.

Experiments conducted show that the classification of terms carried out based on syntactical information improves the results that are obtained if they are classed taking only the frequency into account (Basili et al., 2001). Therefore, by using syntactical information, the temporary expressions are situated in lower positions and the terms representing the speciality corpus are situated in the leading positions.

Semantic methods

Semantic strategies are used to hone the results obtained using the statistical and linguistic term extraction methods. There are basically two types of these strategies: on the one hand, the strategies that use semantic categories of a lexical source outside the work corpus, such as WordNet[2] (Miller et al., 1990), EuroWordNet[3] (Vossen, 1999) and AlethDic (Naulleau, 1998), which organise the lexicon on the basis of the meaning of the words and which can be integrated into a term candidate extraction tool; on the other, those that extract the semantic categories from the words of the same corpus through contextual elements which refer to the syntactic-semantic combination of the words, such as the Fabre model (Fabre, 1996).

Methods that use the contrast corpus

Some measures from the information retrieval field are used into the terminology extraction field to provide a list of candidate terms that represent an specific domain.

A measurement that is widely used in information retrieval and which has been incorporated into the terminology extraction task is the *tf-idf* (term frequency - inverse document frequency) measurement, the aim of which is to filter the terms present in many documents. In this approach, it is necessary to quantify the appearance frequency of a term in a document. This parameter is usually known as a term frequency factor (if, local concept) and it is considered to provide an idea as to what extent this term describes the content of the document, in other words, the more a term appears in a document, the more semantic weight it has. However, the most common terms rarely have the ability to distinguish whether a document is relevant or not for a specific search. For this reason, a factor is entered that is calculated on the basis of an inverse relationship regarding the frequency with which the term appears in a set of documents (inverse documents frequency, *idf*), in other words, the appearance of the term in a set of documents decreases the more documents speak of it; a concept based on the corpus. And the more frequent a term is in a set of documents, the less important and less ability to discriminate it will have and, therefore, it will be less representative of the set of documents. By contrast, terms that rarely appear in

[2] http://wordnet. princeton.edu/

[3] http://www.illc.uva.nl/EuroWordNet/

the set of documents a re the ones that will be more important in the *tf-idf* measurement and, therefore, will better represent all the documents.

In the field of terminology extraction, the *tf-idf* measurement is very productive in determining the relevant terms of a speciality corpus. However, unlike in the information retrieval field, the selection of specialist meaning units is carried out using a general language corpus which will be used to compare the units that appear in this corpus with the ones of a specialist corpus.

The pattern that the *tf-idf* measurement follows is shown below:
Given a collection of documents D, a word *w* and an individual document *d* which belongs to *D,* the following calculation is made:

$$w_d = f_{w,d} log(|D|/f_{w,D})$$

Where $f_{w,d}$ is equal to the number of times that *w* appears in *d* (frequency of the term or tf), $|D|$ is the size of the corpus (total number of documents) and $f_{w,D}$ is equal to the number of documents in which *w* appears in *D.* Finally, $log(|D|/f_{w,D})$ corresponds to the inverse frequency of the document (idf) (Salton and Buckley, 1988; Berger et al., 2000).

2.4 Information retrieval

An important application of terminology detection and extraction techniques is found in the area of information retrieval, in both the indexation and consultation phases. The use made of the terms in this area is also key, as it will be used to class documents and index information.

In general terms, information retrieval is the representation, storing, organisation and retrieval of informational objects that we will call “documents”. This representation and organisation must provide the user with easy access to the information in which they are interested. Unfortunately, the task of characterising the user’s information needs is a complex problem. This need first has to be translated into a search or consultation equation that can be processed by the search engine. Generally speaking, this translation achieves a set of key words or indexation terms that, theoretically, represent the description of the user's information needs. Given the search, the primary aim of the information retrieval system is to retrieve information relevant to the user, which does not necessarily mean those documents that contain all the consultation terms.

For many years, interest in these questions has been limited to documentalist and information experts, despite the fast dissemination of information retrieval tools. However, at the start of the 1990s, an event completely changed the situation, the introduction of the internet. The web has become the repository of knowledge and human culture, which has enabled us to share ideas and information at a speed and to an extent never seen until that time (Baeza-Yates and Ribero-Neto, 1999).

Today, information retrieval from the content published on the internet is characterised by the constant changes that occur in databases and in the variation of the cover of search engines. For this reason, it has been decided to create a static website corpus, whereby the functioning of the search engines and the different information retrieval techniques can be better assessed. And it is the dynamism of the web that marks the differences with traditional information retrieval. If previously in the assessment of information retrieval systems the precision and cover were taken into account, if we now talk about the web only precision is measured, because cover is difficult to measure.

2.5 Information management

Information management is a field that is closely linked to information retrieval and terminology extraction. Information and the strategic use made of it represent a competitive edge for an organisation (Nonaka and Takeuchi, 1995). In this sense, the ability to create, use, retain and transfer information in the constant and dynamic knowledge creation process is key.

If knowledge is the competitive edge of an organisation, access to information to create knowledge and the processes carried out to retain and transfer this knowledge become the nerve centre of the organisation.

Competitive organisations have access to the same intelligence, but the key lies in using them and not only having access. To do this, resources are needed which enable knowledge to be created and information to be used strategically. In this sense, the application of terminology extraction techniques makes it possible to reach key information that organisations have more quickly, which gives them a competitive edge over other organisations and, at the same time, improves their knowledge management. There are two determining points in the strategic response of an organisation: the capacity to create knowledge and the ability to process strategic information.

Information theory focuses basically on the processing of human information: people use the cognitive process to interact with information. Bertran Brookes uses an equation to explain the notion of “cognitive interactions”:

$$K[S] + \Delta I = K[S + \Delta \mathrm{S}]$$

Where $K[S]$ corresponds to “knowledge structure”, ΔI corresponds to “information increase” and $K[S + \Delta S]$ corresponds to “change to the knowledge structure2.

In Brookes’ formulation there are two key ideas that should be stressed: on the one hand, the change in the structure of knowledge is due to the new needs and situations that may occur, i.e. it is due to the new information uses that there may be in any field of knowledge; on the other hand, the increase in information is directly proportional to the increase in knowledge in an organisation. Therefore, we observe that in the field of knowledge management, information use and, consequently, the concepts that articulate these new uses are key, which will mean that the information has a specific structure.

3 Automatic translation equivalents search

Automatic equivalents search works as follows and can be conducted completely statistically. We take a candidate term in language A and select a subset of the parallel corpus consisting of all the segments that contain that term. We only keep the segments corresponding to language B and run a statistical terminology extraction on these sentences. The most frequent candidate term in these sentences will be the most probable translation equivalent of the term selected. This is so because it is expected that in all, or most, of the sentences on language B of this subset the translation equivalent of the term selected will appear.

Translation equivalents search can also be supported by linguistic information. In this case, knowing the category structure of the original term, a similar search to the statistical one can be run, but only looking for equivalents that have a specific category structure. In this case, we first need to know what category structures in the target language may be a translation of a specific structure of the source language.

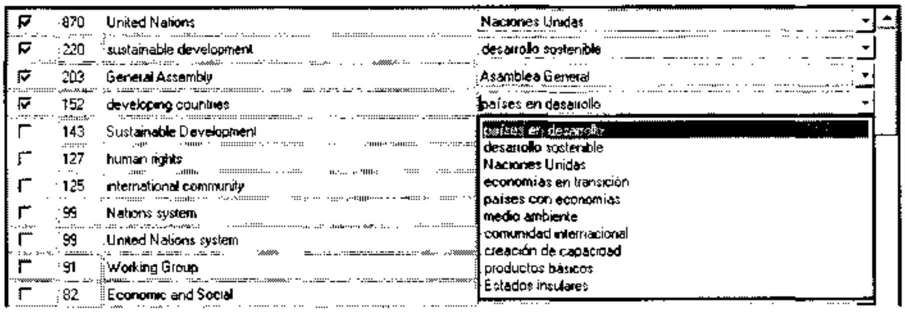
4 Linguoc LexTerm: our first terminology extraction tool

The LexTerm program (Oliver et al., 2007) is a free statistical automatic terminology extraction program, distributed as a free and open source program. LexTerm offers any translator, terminologist, business or institution the possibility of creating terminological glossaries quickly and efficiently.

LexTerm is developed entirely in Perl and is, therefore, a multi-platform program. It is distributed with the Perl code, which can be executed on any computer that has a Perl interpreter installed, and also in executable version for Windows. The executable version for Windows will work on any computer that incorporates this operating system, whether the Perl interpreter is installed or not.

LexTerm permits automatic terminology extraction and automatic translation equivalents search. Automatic terminology extraction can be conducted on the basis of a text document or set of text documents or from a parallel corpus (in text format separated by tabulators). The translation equivalents search is run using a parallel corpus.

Figure 1 shows the result of an automatic terminology extraction. On the left of the original term, we can see the term’s frequency of appearance and a selection box that allows us to indicate the terms that we want to export. On the right is the translation equivalent of the term selected, automatically calculated statistically. This process determines a series of possible translation equivalents with an associated probability. The first one that is presented is the one with the highest probability, but we can scrolldown a list of the other candidates.

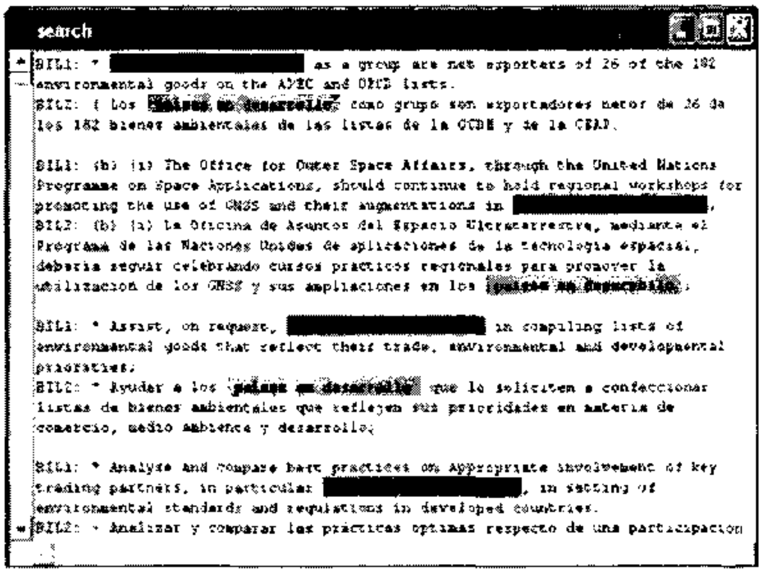


☑	870	United Nations	Naciones Unidas
☑	220	sustainable development	desarrollo sostenible
☑	203	General Assembly	Asamblea General
☑	152	developing countries	países en desarrollo
□	143	Sustainable Development	
□	127	human rights	
□	125	international community	
□	99	Nations system	
□	99	United Nations system	
□	91	Working Group	
□	82	Economic and Social	

países en desarrollo
desarrollo sostenible
Naciones Unidas
economías en transición
países con economías
medio ambiente
comunidad internacional
creación de capacidad
productos básicos
Estados insulares

Figure 1: Result of an automatic terminology extraction and of the translation equivalents search

To help the user determine whether a candidate term is really interesting, or whether the calculated translation equivalent is the correct one, it is possible to obtain the appearance contexts of the candidate term. Figure 2 shows the contexts screen.



```
BIL1: * [redacted] as a group are net exporters of 26 of the 182
environmental goods on the APEC and OECD lists.
BIL2: { Los países en desarrollo como grupo son exportadores netos de 26 de
los 182 bienes ambientales de las listas de la OCDE y de la CEAP.

BIL1: (b) (i) The Office for Outer Space Affairs, through the United Nations
Programme on Space Applications, should continue to hold regional workshops for
promoting the use of GNSS and their augmentations in [redacted],
BIL2: (b) (i) La Oficina de Asuntos del Espacio Ultraterrestre, mediante el
Programa de las Naciones Unidas de aplicaciones de la tecnología espacial,
debería seguir celebrando cursos prácticos regionales para promover la
utilización de los GNSS y sus ampliaciones en los países en desarrollo;

BIL1: * Assist, on request, [redacted] in compiling lists of
environmental goods that reflect their trade, environmental and developmental
priorities;
BIL2: * Ayudar a los países en desarrollo que lo soliciten a confeccionar
listas de bienes ambientales que reflejen sus prioridades en materia de
comercio, medio ambiente y desarrollo;

BIL1: * Analyse and compare best practices on appropriate involvement of key
trading partners, in particular [redacted], in setting of
environmental standards and regulations in developed countries.
BIL2: * Analizar y comparar las prácticas óptimas respecto de una participación
```

Figure 2: Sample screen that shows the contexts of the candidate terms

Linguoc Lexterm uses a statistical strategy for both the extraction of candidate terms and for determining translation equivalents. This methodology requires a list of stopwords. The distribution includes a list of stopwords for English, Spanish and Catalan. However, stopword lists can be created easily for other languages or can be found published on the internet for free download.

5 Design of the Free Terminology Extraction Suite

The tool presented in the above section is fully operational and can be useful in a wide range of situations. Nonetheless, the LPG research group and the UOC's Language Service are developing a series of tools for terminological extraction with the following characteristics:

- Tools distributed as free software, under the GPL licence[4].
- A fully modular development, allowing advanced users to adapt these modules to their specific needs.
- As well as the modules, a complete tool will be distributed with an easy-to-use visual interface.
- It is developed in its entirety in Perl, which means the resulting programs are multi-platform.
- Users can choose between linguistic and statistical methodologies for both the extraction of term candidates and for the calculation of translation candidates. Use of the linguistic methodology is subject to the availability of a tagger for the languages in question.
- The modules can be used easily with any tagger and are fully adapted for use with TreeTagger[5] (Schmid, 1994) and FreeLing[6].(Carreras et al., 2004)

The following sections briefly describe each of the modules currently available. These modules can be downloaded from www.linguoc.cat.

5.1 Modules for statistical terminology extraction

The modules described here can be used to automatically extract term candidates employing statistical methodology.

5.1.1 ngrams.pm

This module calculates the ngrams. The input parameters are the lower order n, the upper order *n* and an array containing the tokens from the text(s) for calculation of the ngrams. The module’s output is a hash containing the different ngrams as the key and the frequency as the value.

[4] http://www.gnu.org/licenses/gpl-3.0.txt

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://garraf.epsevg.upc.es/freeling/

5.1.2 outerfilter.pm

This module removes ngrams whose extremes, ie, the first and the last, are words from a list of stopwords. The input parameters are the file containing the list of stopwords and a hash containing the ngrams with their frequencies (calculated using ngrams.pm). The output is a hash containing the filtered ngrams.

5.1.3 innerfilter.pm

This module removes the ngrams containing words from a list of stopwords in inner positions, ie, from any position except the first and the last. The input parameters are the file containing the list of stopwords and a hash containing the ngrams with their frequencies (calculated using ngrams.pm). The output is a hash containing the filtered ngrams.

5.1.4 ngramstxt.pm

This module allows for the calculation of the ngrams in a text string passed as a parameter. The input parameters are the text string, the lower order *n* and the upper order *n*. The module's output is a hash containing the different ngrams as the key and the frequency as the value.

5.1.5 ngramsfiletxt.pm

This module allows for the calculation of the ngrams in a text document. The input parameters are the path to the text file to be processed, the lower order *n* and the upper order *n*. The module's output is a hash containing the ngrams as the key and the frequency as the value.

5.1.6 ngramsdirtxt.pm

This module allows for the calculation of the ngrams for all the text documents held in a directory. The input parameters are the directory containing the text documents to be processed, the lower order *n* and the upper order *n*. The module's output is a hash containing the different ngrams as the key and the frequency as the value.

5.1.7 ngramsrecdirtxt.pm

This module allows for the calculation of the ngrams in all the text documents held in a directory, processing this directory recursively, ie, a recursive process through all the sub-directories it might contain. The input parameters are the directory containing the text documents to be processed, the lower order *n* and the upper order *n*. The module's output is a hash containing the different ngrams as the key and the frequency as the value.

5.1.8 Sample program

Below, we present a simple sample program that uses some of the modules described above. The program calculates the term candidates for all the files held in the "texts" directory. First, it calculates all the ngrams between order 2 and 3. Then, all the candidates with stopwords from the stopwords.txt file in the first or last position are removed. Lastly, all the candidates are written to the candidates.txt file with their frequency, ordered from most to least frequent.

```
#! /usr/bin/perl
use outerfliter;
use ngramsdirtxt;
%ngrams=ngramsdirtxt ("./texts",2,3);
%ngrams=outerfilter("stop-eng.txt",%ngrams);
open(OUT,">candidates.txt");
foreach $key (sort{$ngrams{$b}<=>$ngrams{$a}}(keys(%ngrams)))
{
   print OUT "$ngrams-[$key}\t$key\n";
}
```

As you can see, using these packages is very simple for users with a basic knowledge of Perl. The suite's documentation explains how to use the different components in detail. A graphic interface will be provided for those users with no programming experience so as to make it intuitive and easy to use.

Below is a sample result from this program. In this example, we use English text from a 10.000 sentence fragment of the Crater corpus[7]. Here we can see the first 20 candidates and their associated frequency:

```
478 data link
310 link layer
302 data link layer
249 Recommendation Q
213 supplementary service
170 state exists
160 information element
130 information transfer
127 supplementary services
120 access connection
120 link connection
106 call reference
105 mobile station
104 layer management
102 data link connection
94 location register
90 network interface
84 layer entity
```

[7] Crater. Multilingual Aligned Annotated Corpus, http://www.comp.lancs.ac.uk/ linguistics/ crater/ corpus.html

```
80 link layer entity
78 call establishment
```

As can be seen, many of these candidates are significant from a terminological point of view. Given that the system does not use any linguistic information, the candidates may include variations of the same term *(Signalling System - signalling system* or *signal unit - signal units).* Terminologists have to revise the list and remove candidates that are not of interest or unify variations of the same term.

5.2 Modules for linguistic terminology extraction

As we have already mentioned, the linguistic methodology is based on recognising certain morphosyntactic patterns that may be typical from a terminological point of view. The prior step required before linguistic extraction is morphosyntactic tagging of the texts to be used as the basis of the extraction. The Free Terminology Extraction Suite does not provide a tagger, but it can work, with slight modifications to the output, with virtually all taggers.

The output modules are designed to work with two taggers: TreeTagger and Preeling. Below is a fragment of the Crater corpus tagged using TreeTagger:

```
The DT the
location NN location
register NN register
should MD should
as RB as
a DT a
minimum NN minimum
contain VVP contain
the DT the
following VVG follow
information NN information
about IN about
a DT a
mobile JJ mobile
station NN station
:  :  :
. . .
```

Here is the same output tagged using Freeling:

```
The the NP
location location NN
register register NN
should should MD
as as IN
a a DT
```

```
minimum minimum NN
contain contain VBP
the the DT
following follow VBG
information information NN
about about IN
a a DT
mobile mobile NN
station station NN
: : Fd
. . .
```

Both outputs are very similar. The output from TreeTagger is organised by form tab tag tab lemma; whereas the output from Freeling is organised by form tab lemma tab tag. When using taggers, it should be taken into account that errors may occur. These errors may affect the quality of the terminology extraction.

The terminology extraction process consists of searching for a series of typical terminological patterns. To do so, a file is required to define these patterns.

5.2.1 lingextrac.pm

The lingextract module's input parameters are the tagger used ("tt" for TreeTagger and "fl" for Freeling), the tagged file and the file with the patterns to be searched for. The module's output is a text file containing the term candidates and their frequency.

For example, if the pattern file defines the following patterns:

```
NN NN
JJ NN
NN NN NN
JJ NN NN
JJ JJ NN
```

The following candidates are obtained for a 10.000 sentence fragment of the Crater corpus analysed using Freeling (the first 20 are shown):

```
280 link layer
260 data link
198 data link layer
160 information element
130 information transfer
120 access connection
112 link connection
106 call reference
104 layer management
90 network interface
```

```
84 layer entity
80 link layer entity
78 location register
74 call establishment
70 system management
68 management entity
64 call control
53 mobile station
52 physical layer
50 stop element
```

5.2.2 findpatterns.pm

One of the main problems with linguistic extraction is the need to have a file with the typical terminological morphosyntactic patterns. The findpatterns module finds typical patterns from a list of known terms and a tagged corpus that may contain these terms. The program searches for terms in the tagged corpus and stores the patterns found. The input parameters are the tagger used to tag the corpus, the path to the list of reference terms and the path to the tagged corpus. The output is a list of patterns and their frequency. The output when using as reference terms the terms in English from Termcat's Dictionary of Telecommunications[8] is as follows:

```
11314 NN
1260 VBG
669 NN NN
274 JJ
172 JJR
161 VBP
77 JJ NN
32 JJR NN
30 VBP NN
22 NN NN NN
14 VBN NN
12 NNS NN
10 NN VBG
6 NNS NN NN
6 VBP VBG
4 NN NN JJ
4 NP NN
2 NN IN VBP
2 VBN
2 JJ NN NN
2 NNS
2 JJ NNS
2 VBG NN
2 JJ JJR NN
```

[8] www.termcat.cat

```
2 VBD NN
```

5.3 Modules to automatically search for translation equivalents

Currently, we have only developed a module that allows for an automatic search of translation equivalents using a statistical methodology and parallel corpora. The module is called steqfind.pm and is described in more detail below.

5.3.1 steqfind.pm

The steqfind.pm module searches for a series of translation equivalent candidates for a given term using a parallel corpus. The module needs the following input parameters: the original term that we want to find a translation equivalent for, the path to the stopwords file corresponding to the original language, the path to the stopwords file corresponding to the target language and a hash containing the parallel corpus, where the key is the original segment and the value the translated segment(s). The module returns a hash with the translation equivalent candidates and their estimated probability.

5.4 Modules to be developed in the future

5.4.1 Module to aid standardisation of terms

Currently, using our statistical or linguistic terminology extraction modules provides a list of term candidates that may contain a number of variants of the same term (inflected forms, differences in lower and upper cases, etc.). We plan to develop a module to aid standardisation of the terms, selecting only a base form. The module will propose a group of different term candidates and a proposed base form for each of them. The user will have to validate the proposal.

5.4.2 Automatic search for translation equivalents using a linguistic methodology

The module presented in section 5.3.1 functions using a statistical methodology. The results are satisfactory, but we feel that the methodology for searching for translation equivalents in parallel corpora can be improved upon by using linguistic information. In the same way that we define a series of morphosyntactic patterns for extracting term candidates, we define a series of morphosyntactic patterns in the target language, which are typically those for the translation equivalent. For example, the morphosyntactic pattern in English "NN NN" is very productive (*access control, access point,* *aspect ratio).* The translation equivalents for these terms in Spanish (*control de acceso, punto de acceso, relación de aspecto*), typically follow the “N de N” pattern. Knowing this can be of use in finding translation equivalents for new terms.

This module is to be accompanied by a module that learns these patterns from a bilingual terminological dictionary and a tagged parallel corpus.

5.4.3 Automatic search for translation equivalents in comparable corpora

The modules presented in section 5.3.1 and 5.4.2 use aligned parallel corpora to search for translation equivalent candidates for a given term. These methodologies work reasonably well, but there is the drawback that they need parallel corpora. Parallel corpora are more difficult to obtain than comparable corpora. There are a number of methodologies for searching for translation equivalents in comparable corpora (Rapp, 1995) and (Gamallo and Pichel, 2007). We plan to implement one of these methodologies in the suite.

5.4.4 Implementing statistical measures for candidate reordering

In the extraction of terminology, whether using linguistic or statistical methodology, we only use the frequency of appearance of candidates in our corpus to determine the ordering of term candidates. We are exploring a series of statistical measures to improve this ordering. In section 6, there is a detailed explanation of the research being carried out. Once we have determined the measure or measures that work best for this task, we will implement them in our suite.

6 Statistical measures for term candidate reordering

The use of statistical measures to extract term candidates can improve the results obtained by terminology extraction tools that use statistical methods, as shown in section 2.3. This is the case because statistical measures reorder the position in which term candidates are found in terms of their frequency, ie, they place a greater number of terminological units in the first places.

The study we have carried out on eleven statistical measures based on bi-gram combinations from the specialist Crater corpus has shown us which measure performs best in terms of reordering term candidates. Specifically, we have studied the Dice coefficient, two-tailed Fisher test, Jaccard coefficient, Log-likelihood ratio, True mutual information measure, Pointwise mutual information measure, Odds ratio, Pearson’s Chi-square test, T-score test, Poisson Stirling measure and PHI coefficient.

The resources used for this study include a statistical analysis tool[9], the specialist Crater corpus from the field of telecommunications, a corpus of reference terms also linked to the field of telecommunications[10] and a list of stopwords[11].

The results obtained from this study show, on the one hand, which statistical measure retrieves the greatest number of reference terms and which best orders them; and, on the other, which measure retrieves the greatest number of specialised meaning units, including the reference terms. The latter results are based on the manual review of a sample of two hundred term candidates carried out by five informants.

6.1 Results obtained for the number of reference terms

In the first series of results, we want to establish the number of reference terms present in our specialist corpus and the measures that best retrieve them in the highest places in the list of results. This is achieved by filtering the results with the stopwords list available.

To gain these results, we first automatically ascertained the number of terms from the reference term corpus which were to be found in the specialist corpus. The result returned was 1,170 reference terms from a total of 4,000. Then, we extracted terms from the specialist corpus using the statistical program and produced a list of filtered term candidates, in descending order of frequency. These initial results correspond to the statistical frequency calculation and return a total of 44,498 term candidates from the specialist corpus. Finally, we have prepared the results corresponding to the eleven statistical measures. The results obtained for each measure reorder the position of each of the term candidates returned by the statistical frequency calculation using different statistical calculations. Thus, the fact that there are eleven different measures to reorder the same list of term candidates has allowed us to see which measure performs best with respect to term recognition.

In figure 3, we can see the number of reference terms recognised by the five statistical measures that offer the best results: the frequency calculation, Poisson Stirling measure, True mutual information measure, Log-likelihood ratio and T-score measure. These results have been gained by assessing all the term candidates (44,498). Specifically, the category axis (X) shows the different positions in which the term candidates can be found (the order in which the measures place the candidates) and the value axis (Y) shows the number of reference terms that the measures retrieve in each position.

[9] The Ngram Statistics Package tool can be found on SourceForge (http://sourceforge.net/projects/ngram).

[10] *Diccionari encidopèdic de telecomunicacions,* Enciclopèdia Catalana, Polytechnic University of Catalonia and Termcat, Centre de Terminologia, 2007

[11] NLTK (Natural Language Toolkit): http://nltk.sourceforge.net

As our study only looked at bi-grams, a total of 355 reference terms were retrieved.

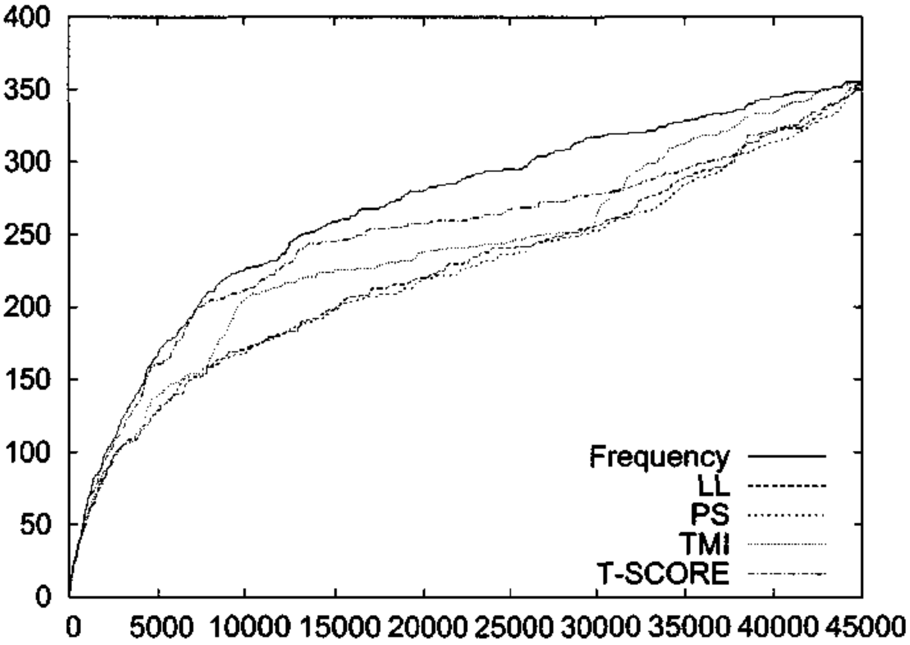


Figure 3: Number of reference terms retrieved using five statistical measures.

The results that are of most interest to us in our study are those on the left-hand side of the figure, as these correspond to the terms that appear most frequently in the corpus and, thus, are those that are of most interest from a terminological point of view. Consequently, the measures that retrieve the most reference terms in this initial stage of results are those that recognise them most easily and that save time when it comes to their manual review. Thus, we can see in figure 4 that up to, approximately, position 1,000, the T-score measure is that which retrieves most reference terms, followed by the frequency calculation, True mutual information, Log-likelihood, Poisson Stirling and the other measures.

Focusing on the number of reference terms retrieved by the different measures up to position 1,000, the table 1 shows that the T-score measure is that which recognises the most reference terms, followed by the frequency calculation, True mutual information, Poisson Stirling and Log-likelihood. Similarly, if we look at the results obtained for positions 250, 500 and 750, we can see that T-score is again the measure that places most reference terms in the highest places on the list of results, alongside the frequency calculation.

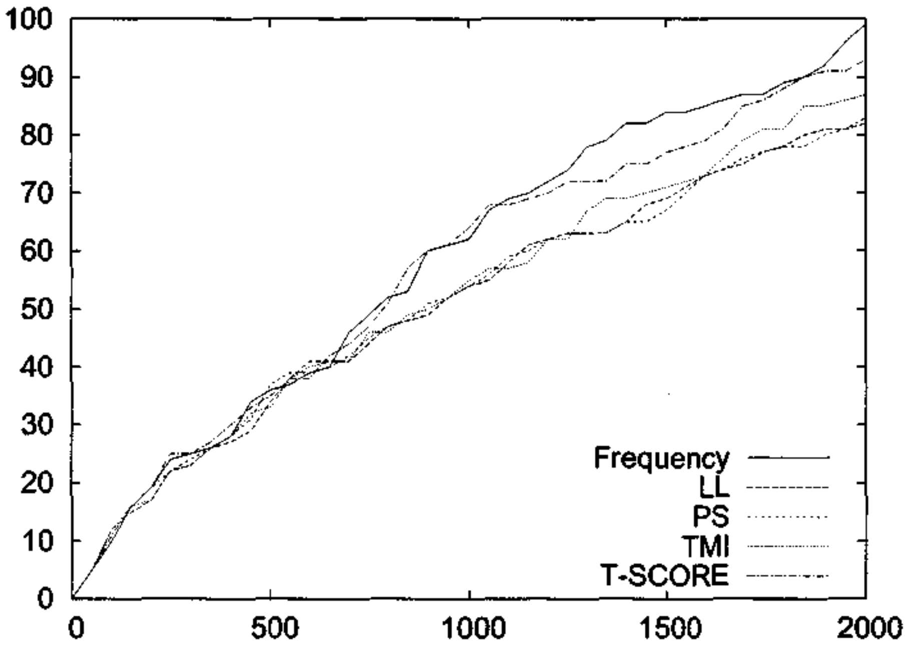


Figure 4: A closer look at the number of reference terms retrieved by each statistical measure.

Thus, the measure that retrieves the greatest number of reference terms coincides with that which orders or places them in the highest places on the list.

6.2 Results obtained for the number of specialist meaning units

The second series of results required manual assessment by five informants of the first two hundred term candidates from the five statistical measures that were evaluated in the first series of results; ie, the frequency calculation, Poisson Stirling measure, True mutual information measure, Log-likelihood ratio and T-score measure.

These term candidates have been filtered using the stopwords list and ordered by frequency in descending order. In this case, then, we are using a term acquisition approach which focuses on finding new terms in a specialist corpus. In our case, however, instead of doing so automatically, the informants found the terms manually.

Statistical measures	250	500	750	1000
Frequency calculation	24	36	49	62
Dice coefficient	1	1	2	2
Two-tailed Fisher test	3	6	10	13
Jaccard coefficient1	1	1	2	2
Log-likelihood r atio	22	34	44	54
True mutual information measure	22	33	46	55
Pointwise mutual information measure	0	1	1	2
Odds ratio	1	1	1	2
Pearson's Chi-square test	1	1	1	2
T-score measure	25	35	47	64
Poisson Stirling measure	22	37	45	54
PHI coefficient	1	1	2	2

Table 1: Number of reference terms in the specialist corpus

6.2.1 Informants' selection of specialist meaning units

The five informants chosen to select the terms from the specialist corpus we worked with had the following profiles: two are experts in the field of telecommunications and three are experts in terminology.

The sample chosen for the assessment of the results included two hundred term candidates, as we felt that this number of units would offer sufficiently representative data for the behaviour of the different statistical measures in terms of the placement of terms in the highest positions in the list of results and would also allow us to assess the data with informants.

The people evaluating the results have done so bearing in mind the terminological pertinence of the two hundred term candidates from the five aforementioned measures so as to be able to compare and contrast the results; in total, 263 candidates have had to be supervised. Likewise, assessment has had to be made of the units that are not specifically linked to the field of telecommunications, but which are of vital importance in this specialist area. Words from general language that appear in a specialist context and which are deemed terms for the use made of them have also been taken into account. In short, the informants have had to handle three groups of term candidates:

1. Terms that are specific to the field of telecommunications.
2. Terms that belong to other fields, but which also have a specialist character in terms of telecommunications.
3. Terms from general language, which, due to the fact that they are used in the corpus, are deemed terms.

Statistical measures	Reference terms	Manual selection
Frequency calculation	19	126
T-score measure	19	124
Poisson Stirling measure	19	115
True mutual information measure	17	113
Log-likelihood ratio	17	112

Table 2: Comparison of the number of reference terms and manual extraction

6.2.2 Evaluation of the results

With regard to the results obtained from the assessment made by the informants, and in terms of the level of agreement that there has been in their selection of the specialist meaning units, it should be pointed out that all five informants chose 96 of the 200 term candidates; ie, there was unanimous agreement in 48% of cases among the informants. This result has to be added to the fact that four of the five informants chose a further 56 of the 200 term candidates, which corresponds to 28% of the total. Thus, the sum of these two sets of results means that a total of 76% of the terms were chosen in almost unanimous agreement by the informants.

This number of terms chosen by the informants (a total of 152) offers a series of new reference terms that can be used to expand the corpus of reference terms that was used as the basis of our study. Likewise, these new reference terms serve to show the number of specialist meaning units, including the initial reference terms, in the sample of two hundred term candidates from the five measures analysed. The results obtained are as follows: the frequency calculation found 126 terms, the T-score measure 124, the Poisson Stirling measure 115, the True mutual information measure 113 and the Log-likelihood measure 112. Thus, we can see that none of the measures can improve on the results obtained using the frequency calculation when it comes to finding new terms, despite the fact that one of the results comes close: the T-score measure.

Table 2 uses a sample of two hundred candidates to compare the results of reference term extraction obtained using the frequency calculation, Poisson Stirling measure, True mutual information measure, Log-likelihood ratio and T-score measure with the results revised manually by the five informants.

The results obtained following the manual review of the data only go to confirm that the best measure for retrieving specialist meaning units from a specialist corpus is the statistical frequency calculation, alongside the T-score measure, which comes second in terms of results.

Thus, we can see that the two working methods that we have used to assess the behaviour of the different statistical measure, with reference terms and informants, return the same ranking of measures with respect to the number of units retrieved. Likewise, the manual search for terms highlights even more clearly the position in which each of the five statistical measures can be found in this ranking.

7 Conclusions

This article reviews the state of the question with regard to terminology extraction, and information retrieval and management. Likewise, we have described a series of freely distributed open-source tools that allow for the automatic building of monolingual, bilingual or multilingual terminologies, in order to make the manual review tasks of specialists much easier. These terminologies are designed for proofreaders, translators, terminologists, documentalists, information managers and specialists in general.

With respect to the tools presented, it should be pointed out that they work, fundamentally, with endogenous methods; ie, they use statistical and linguistic methods to extract units that are term candidates. On the one hand, the Lexterm tool allows for terminology extraction using statistical methods and automatic searches for translation equivalents. And on the other, the series of tools that we have started to develop allow for terminology extraction using statistical and linguistic methods, as well as automatic searches for translation equivalents.

The use of endogenous methods in the tools developed allows for the production of a relatively representative list of term candidates for the specialist field they are taken from. In short, the comparative study that we have carried out on eleven statistical measures in the terminology extraction process allows for a greater number of units returned corresponding to term candidates.

8 Future work

In terms of future work, we plan to incorporate exogenous methods into the series of tools developed so as to produce an exhaustive list of term candidates. Thus, we want to introduce the use of methods that use corpora to contrast the terminology extraction, and methods that focus on the information surrounding the term.

References

A. Arppe. Term extraction from unresticted text. In *Proceedings of the X Nordic Conference of Computational Linguistics*

(NODALIA 1995), 1995. URL `http://www2.lingsoft.fi/ doc/ nptool/ term-extraction.html`.

R. Baeza-Yates and B. Ribero-Neto. *Modern information retrieval.* ACM Press, 1999.

R. Basili, M.T. Pazienza, and F.M. Zanzoto. Modelling syntactic context in automatic term extraction. In *Proceedings of the III Conference on Recent Advances in Natural Language Processing (RANLP),* 2001. URL `http://citeseer.ist.psu.edu/ basili01modelling.html`.

A.Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: Statistical approaches to answer finding. In *Proceedings of the XXIII Annual Internationa ACM SIGIR on Research and Development in Information Retrieval,* pages 192-199, 2000. URL `http://citeseer.ist.psu.edu/ article/ berger00bridging.html`.

D. Bourigault. *Lexter, un logiciel d'Extractions de Terminologie.* PhD thesis, École de Hautes Études en Sciences Sociales, 1994.

M.T. Cabré. *Una nueva teoría de la terminologia: de la denominación a la comunicación.* La terminología: representatión y comunicación. Una teoría de base comunicativa y otros propósitos. Institut Universitari de Lingüística Aplicada, 1999.

X. Carreras, I. Chao, L. Padro, and M. Padro. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04),* 2004. URL `http://www.lsi.upc.es/ nlp/ papers/ 2004/ lrec04-ccpp.pdf`.

B.Daille. Combined approach for terminology extraction: lexical statistics and linguistic filtering. *UCREL,* 5, 1995. URL `http://www.comp.lancs.ac.uk/ ucrel/ papers/ techpaper/ vol5.pdf`.

S. David and P. Plante. Le progiciel termino: de la néceessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes. In *Actes du Colloque international des industries de la langue: perspectives des années,* pages 71-88, 1991.

P. Drouin. Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme. *Meta,* 42(1):45-54, 1997. URL `http://www.erudit.org/ revue/ meta/ 1997/ v42/ n1/ 002593ar.pdf`.

C.Enguehard and L. Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics,* 2(l):27-32, 1994.

R. Estopà. *Extracció de terminologia: elements per a la construcció d'un sistema d'extracció automàtica de candidats a unitats de significació especialitzada.* PhD thesis, Universitat Pompeu Fabra, 1999.

D. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of XXXIV Annual Meeting of the Association for Computational Linguistics (ACL 1996),* pages 17-24, 1996. URL `http://citeseer.ist.psu.edu/ evans96nounphrase.html`.

C. Fabre. *Interprétation automatique des séquences binominales en anglais et en francais. Aplication à la recherche d'informations.* PhD thesis, Université de Rennes I, 1996. URL `http://www.inria.fr/ rrrt/ tu-0909.html.`

P. Gamallo and J.R. Pichel. Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. In *Proceedings del XXIII Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural,* pages 241-248, 2007.

C.Jacquemin. Syntacmatic and paradigmatic representations of term variation. In *Proceedings of the XXXVII Annual Meeting of the Association for Computational Linguistics (ACL 1999),* pages 341-348, 1999. URL `http://citeseer.ist.psu.edu/ jacquemin99syntagmatic`.

J. Justeson and S. Katz. Technical terminology: some linguistic properties and algorithms for identification in text. *Natural Language Engineering,* 1(l):9-27, 1995.

D. Maynard. *Term recognition using combined sources.* PhD thesis, Universitat Metropolitana de Manchester, 1999.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography,* 3(4):235-244, 1990. URL `ftp://ftp.cogsci.princeton.edu/ pub/ wordnet/ 5papers.ps.`

E.Naulleau. *Apprentissage et filtrage syntaxico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire.* PhD thesis, Université Paris XIII, 1998. URL `http://smiosys.free.fr/ Semio.Sys/ past/ THESEEN98.pdf.`

I. Nonaka and H. Takeuchi. *The knowledge-creating company: How Japanese Companies Create the Dynamics of Innovation.* Oxford University Press, New York, 1995.

A. Oliver, M. Vàzquez, and J. Moré. Linguoc lexterm: una herramienta de extracción automática de terminología gratuita. *Translation Jour-*

nal, 11(4), October 2007. URL http://accurapid.com/ journal/ 421inguoc.htm.

R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33th Conference of the ACL'95,* pages 320-322, 1995.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* 5(24):513-523, 1988.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing,* 1994. URL http://www.ims.uni-stuttgart.de/ftp/pub/corpora/ tree-taggerl.pdf.

P. Vossen. *EuroWordNet as a multilingual database.* Wolfgang Teubert (ed) TWC, 1999.