English for Specific Purposes World, ISSN 1682-3257, www.esp-world.info, Issue 46, 2015

Measuring precision in legal term mining: a corpus-based validation of single and multi-word term recognition methods

María José Marín

ABSTRACT. *Legal terminology presents certain traits which may interfere with its automatic detection such as its relevant presence in everyday language. Thus, this research explores the levels of precision achieved by five single and multi-word term recognition methods on a pilot legal corpus of 2.6 million words. A comparison is carried out with the results presented by Marín (2014a). Once the most effective single and multi-word term recognition method is singled out, it is applied to the reference corpus, BLaRC, with the aim of producing a reliable list of legal terms which might be exploited in areas such as English for Specific Purposes (ESP) instruction, Applied Linguistics or Terminology.*

MOTS-CLÉS: *anglais juridique; méthodes ATR; linguistique de corpus; terminologie*

KEYWORDS. *legal English; ATR methods; corpus linguistics; terminology*

1. Introduction

One of the peculiarities of legal English, as stated by scholars (Mellinkoff, 1963; Tiersma, 1999; Borja, 2000; Orts, 2006), is the presence of legal terms in everyday language. They can be employed by non-specialists in the area and can frequently be found in the press or in other non-technical text sources. Automatic term recognition (ATR) methods often compare the frequency data obtained from general and specialised corpora for the identification of specialised terms in the latter. Such terms, owing to their degree of specificity, are expected to appear less often in the general context, however, this is not so for legal English, as already stated.

As explained in greater detail in section 5 of this paper, a list of legal terms was extracted from the *British Law Report Corpus* (*BLaRC*), compiled by Marín (2014a), by applying the most efficient ATR methods singled out from the list offered in table 1. Subsequently, it was compared with three general English vocabulary lists provided by West (1953), Coxhead (2000) and the *British National Corpus* (2007), which are fundamental references in the field. These vocabulary inventories include the most frequent 3,000 word families of English and therefore reflect basic English vocabulary usage.

In spite of the general character of the vocabulary lists above, the results of this comparison yielded that 40,47% of the legal terms identified in the *BLaRC* could also be found in West's (1953) and Coxhead's (2000) lists. In fact, the percentage was slightly higher, 45.41%, if compared with the *British National Corpus*, thus confirming that almost half of the legal terminology extracted from our corpus was shared with general English. However, from the point of view of automatic term recognition, this peculiarity of the legal lexicon might hamper the efficiency of automatic methods based on corpora comparison, owing to the fact that the presence of legal terms in general corpora might be greater than expected.

Further to this, it is not only the high frequency of legal terms in both the general and legal contexts but also their polysemic character which may affect the levels of precision achieved by ATR methods. General English words often specialise into legal terms acquiring new meanings in the legal field. This might be the case of words like "charge", "battery" or "trial", which can be found amongst the most frequent words of English (conveying general concepts), although they specialise when in contact with the legal context, meaning "an accusation by a judge", "the act of beating or pounding" (understood as a crime) or "the process of being tried in a court of law", respectively. This process of acquisition of new meanings might also affect the collocates generated in each context and the contexts themselves, which might once more add to the difficulties already caused by their high frequency in both the general and the legal fields.

All in all, and owing to the fact that, to the best of our knowledge, there are no studies assessing the results obtained by ATR methods in the legal field (with the exception of Marín's (2014), this research presents a proposal for the validation of these techniques with the purpose of providing reliable data for linguists, NLP specialists, translators or ESP instructors interested in studying legal terminology extraction and its potential applications.

To that end, five different ATR methods (capable of extracting both single and multi-word terms) are assessed by implementing them on a 2.6 million-word legal English corpus, the *United Kingdom Supreme Court Corpus* (*UKSCC*), used as reference. Subsequently, the results are compared with the ones obtained by Marín (2014a) in the evaluation of five single-word term (SWT) recognition methods. The present research is therefore a follow-up of Marín's (2014a) study whereby five single and multi-word term (MWT) recognition methods are vaidated, namely, Frantzi

and Ananiadou, 1999; Park *et al.* 2002; Drouin 2003; Sclano and Velardi, 2007; Nazar and Cabré 2012, in order to determine which of them could reach the highest levels of precision in legal term recognition and whether the degree of success of these techniques depended on them focusing on single or multi-word term units.

Having selected the most efficient single and multi-word term recognition method, it is applied to the *BLaRC*, an 8.85 million-word legal English corpus, with the purpose of producing a reliable lists of legal terms which can be employed for different purposes such as English for specific purposes (ESP) instruction, studies on legal terminology or translation. A sample of this vocabulary inventory is provided in appendix 1.

Section 2 of this article is devoted to the study of the literature on ATR methods followed by section 3, where the working procedure followed in this study is described. Section 4 includes the results obtained in the assessment of five single and multi-word term recognition methods with the aim of selecting the most efficient one, whose implementation on the *BLaRC* is presented in section 5. Finally, the last section presents the main findings of this research and the conclusions drawn from them.

2. Literature review

The literature on Automatic Term Recognition (ATR) methods and software tools has been profusely reviewed (Maynard and Ananiadou 2000; Cabré *et al*. 2001; Drouin 2003; Lemay *et al*. 2005; Pazienza *et al*. 2005; Chung 2003; Kit and Liu 2008 or Vivaldi *et al*. 2012, to name but a few) often classifying them according to the type of information used to identify candidate terms automatically. Some of the reviewed methods resort to statistical information, amongst them: Church and Hanks (1990), Ahmad *et al*. (1994), Nakagawa and Mori (2002), Chung (2003), Fahmi *et al*. (2007), Scott (2008) or Kit and Liu (2008). Other authors like Ananiadou (1988), David and Plante (1990), Bourigault (1992) or Dagan and Church (1994) focus on linguistic aspects. The so-called *hybrid methods* rely on both. The work of Daille (1996), Frantzi and Ananiadou (1996; 1999), Justeson and Katz (1995), Jaquemin (2001), Drouin (2003), Barrón Cedeño *et al*. (2009) or Loginova *et al*. (2012) illustrate this trend. As stated by Vivaldi *et al*. (2012), only a few of these methods resort to semantic knowledge, namely, TRUCKS (Maynard and Ananiadou 2000), YATE (Vivaldi, 2001) and MetaMap (Arson and Lang, 2010).

However, the literature on the evaluation of these methods is not so abundant. There are initiatives for the assessment of ATR methods like the one organised by the Quaero program (Mondary *et al*., 2012) which aims at studying the influence of corpus size and type on the results obtained by these methods as well as the way different versions of the same ATR methods have evolved. Some authors also show their concern about the lack of a standard for ATR evaluation which is often carried out manually or employing a list of terms, a gold standard, which is not systematically described (Bernier-Colborne, 2012: 1). Some researchers like Sauron, Vivaldi and Rodríguez, or Nazarenko and Zargayouna (in Bernier-Colborne, 2012) have worked on this area although there is still much to be done in this respect.

3. Methodology

3.1. *Corpus description: the UKSCC*

The *United Kingdom Supreme Court Corpus* (*UKSCC)* is a pilot legal English corpus of 2.6 million words subset of a larger one, the *BLaRC*, of 8.85 million words. The criteria for the design and compilation of the latter are described in detail in Marín and Rea (2012). Its structure and characteristics abide by the main tenets of corpus linguistics as defined by McEnery and Wilson (2001), Sinclair (2005) and McEnery and Xiao (2006) for general corpora and Pearson (1998), Vargas (2005) and Rea (2010) for specialised ones.

The *UKSCC* is a specialised, synchronic and monolingual corpus of 193 judicial decisions issued by the British Supreme Court and House of Lords[1] between 2008 and 2010. The documents included in the *UKSCC* contain authentic texts obtained from the Supreme Court records and transformed into raw text format for their processing. The *UKSCC* is based on one single legal genre, judicial decisions, owing to the pivotal character this genre plays in common-law legal systems such as the UK's.

The UK belongs to the realm of common law, as opposed to civil or continental law, which is the judicial system working in most Western European countries. In purely common law systems, case law stands at their very basis. As a matter of fact, their legal system revolves around the principle of binding precedent, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio decidendi*), and judicial decisions are employed by law practitioners as the basis for their arguments, decisions, and the like, hence their essential character in legal English and the relevance of the terminology employed in them within this English variety.

3.2. *The gold standard*

The results obtained after implementing the ten ATR methods illustrated below on the *UKSCC* were validated automatically against a legal English glossary used as gold standard. Instead of asking specialists to gather a terminology database extracted from the study corpus, given its size, four different British and American legal English glossaries[2] in raw text format as well as a list of terms obtained from *LEGTeXT*, a corpus of legal English textbooks (Marín, 2014b), were merged and filtered resulting into a list of 10,088 items including both single and multi-word legal terms.

Surprisingly and contrary to Nakagawa and Mori's (2002) assumption that 85% of specialised terms are said to be compound (apparently, this statement applies to all sublanguages), it appeared

[1] The *Constitutional Reform Act, 2005* created the Supreme Court which started to work as the court of last resort of the UK in October 2009, until then, it had been the so-called "Law Lords" of the House of Lords who carried out that function. This is the reason why the texts selected from 2008 to 2010 come from both sources.

[2] Available online at:
http://www.legislation.gov.hk/eng/glossary/homeglos.htm
http://www.judiciary.gov.uk/glossary
http://sixthformlaw.info/03_dictionary/index.htm
http://www.nolo.com/dictionary

that only 44.77% terms out of 10,008 were MWTs, being distributed as follows: 3,275 bi-grams (32.46%), 922 tri-grams (9.13%) and 319 MWTs formed by four or more constituents (3.16%).

Once the output candidate term (CT) lists were produced applying each of the selected methods, they were compared with the gold standard using an excel spreadsheet with the aim of determining the overlap percentage existing between both lists. Whenever a CT was found in the glossary it was confirmed to be a true term (TT), therefore, the overlap percentage found between the CT list and the glossary could be understood as the average level of precision achieved by each of evaluated methods.

3.3. *Implementation procedure*

Table 1 illustrates the process of implementation of the five single and multi-word term recognition methods evaluated in the present research. As shown by it, all the methods were implemented automatically. The first two, *C-value* (Frantzi and Ananiadou, 1999) and *Textract* (Park *et al*., 2002), were applied on the corpus using Zang *et al.*'s *JATE* tools (2008), a java tool set which allows the user to implement a set of ATR methods on any corpus. Zhang's *JATE* lemmatises and POS[3]-tags corpora using Schmid's *Tree Tagger* (1994) and resorts to the *BNC* (2007) lemmatised frequency lists as reference for comparison[4]. However, this tool does not employ any filter prior to the actual implementation of the ATR methods, but rather processes the corpus directly. As a result, the output lists produced by *Textract* and *C-value* were filtered employing the function word list and base word list 15 of proper names provided with Heatley and Nation's (1996) *Range* software. There were hundreds of proper names in the candidate term lists owing to their little frequency in general corpora. Judicial decisions are full of them although they are not legal terms, thus, they had to be purged manually before assessing the effectiveness of these methods.

The remaining three techniques could be implemented using the free software provided by the authors online, not requiring any pre-processing steps except for *Terminus* (Nazar and Cabré, 2012). Nazar and Cabré's method required to go through the training phase before being applied. As shown in the method description section, Terminus offers the possibility of training the system so that it can learn what specialised terms are like in every sublanguage. In order to do so, a list of both SWTs and MWTs was uploaded to the server so that *Terminus* applied the learning algorithm on this data set to improve the term extraction results. This is precisely one of the most outstanding features of this system, since the training phase allows it to store a statistical model that it will apply later in the term extraction phase. This information is saved and made freely available so that any other users willing to process a corpus belonging to the same domain can apply it without any difficulty.

METHOD	IMPLEMENTATION
C-value (Frantzi and Ananiadou, 1999)	Automatic: *JATE* tools (Zang *et al*., 2008)

[3] The words in the texts are labelled according to the morphological category they belong in.
[4] Available online at: http://www.kilgarriff.co.uk/BNClists/lemma.al

Textract (Park *et al*., 2002)	Automatic: *JATE* tools (Zang *et al*., 2008)
TermoStat (Drouin, 2003)	Automatic. Free online access[5]
Termextractor (Sclano and Velardi, 2007)	Automatic. Free online access [6]
Terminus (Nazar and Cabré, 2012)	Automatic. Free online access[7]

Table 1. *ATR method implementation procedures*

3.4. *ATR Method description*

3.4.1. *C-value (Frantzi et. al., 1999)*

This ATR method does not resort to corpus comparison but rather stands as a domain-independent one, only based on a specialised corpus. It is a hybrid method which employs both linguistic and statistical data to produce a list of candidate terms (CTs) ranked according to their termhood score. A term's c-value can be calculated with respect to its frequency and the frequency of its sub-terms:

$$CValue(a) = \log_2|a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)\right) \quad [1]$$

where, f (a) is the frequency of term (a) with $|a|$words, T_a is the set of CTs recognised by the method that contain (a) and $P(T_a)$ is the total number of longer CTs that contain (a).

The linguistic part of the method is articulated into different steps which go as follows:

— The corpus is POS tagged.

— A linguistic filter is applied so as to discard certain patterns and keep a balance between precision and recall (the use of an open filter could favour recall at the expense of precision). Only those strings containing nouns premodified by other nouns, adjectives or combinations of both are kept.

— A stop list is employed which comprises both function words and high frequency ones from a sample corpus not expected to be terms.

As part of the statistical parameters utilised to select the CTs, the authors take into consideration the frequency of occurrence of the pattern, also the frequency of the pattern as part of other longer structures, the amount of these longer structures and the number of constituents of the pattern.

Frantzi and Ananiadou introduce the concept of *nested terms* as key within the statistical part of their method. With the purpose of trying to discard those patterns which are not true terms (TTs), they decide to select only those which contain strings which also appear by themselves in the corpus

[5] http://olst.ling.umontreal.ca/~drouinp/termostat_web/index.php
[6] http://lcl.uniroma1.it/sso/index.jsp?returnURL=%2Ftermextractor%2F
[7] http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl

displaying relatively high frequency. A frequency threshold of >3 is applied to avoid producing a too long list that might become a hindrance for the experts evaluating the output.

For the assessment of their method, the authors highlight the fact that there is no agreement amongst experts and that such subjectivity necessarily leads to the introduction of the concept of 'relative' precision and recall. Instead of asking an expert to extract all the terms in a corpus, which is time-consuming and hard to attain, recall figures are obtained "with respect to frequency of occurrence, which we use as the baseline method" (Frantzi and Ananiadou, 1999: 8).

The authors also assess precision at three stages: first, evaluating those candidates which have appeared as nested; second, evaluating only those appearing as nested and third, evaluating all the CTs. As a result, the authors realise that, in general, the use of a more open linguistic filter does not affect precision significantly. Moreover, using other statistical data "apart from the pure frequencyof occurrence of CTs, improves the precision of the extracted nested multi-word terms, with a slight only loss on recall" (Frantzi et *al.*, 13).

3.4.2. *Textract (Park et al. 2002)*

Park *et al*. (2002) design a term recognition tool, *Textract*, capable of identifying specialised terms which, in their view, convey a major part of the technical knowledge contained in specialised document collections. Moreover, these terms are of great relevance since they can be employed by different applications providing information on syntactic patterns, definitions of concepts or even "relationships that link concepts" (Park et *al.*, 2002: 1).

Term lists can be organised in specialised glossaries, which is the authors' main objective. Glossary formation follows different steps. On the one hand, the identification of CTs (this is the method that will be evaluated herein), on the other hand, the validation of the list of CTs by an expert through its presentation employing a glossary administration system. After that, the validated glossary "is made available, through suitable APIs[8], to the application system" (Park et *al.*, 2002: 1).

Let us then concentrate on *Textract*, the ATR method presented by the authors to single out the most relevant terms in a specialised corpus. *Textract* is part of a set of text analysis tools, *TALENT* (*Text Analysis and Language Engineering Technology*), designed by the Information Retrieval and Analysis Group at IBM.

This tool identifies both single and multi-word terms (both noun and verb phrases). The authors apply several filters. To start with, patterns with more than six units are eliminated, proper nouns (person and place names) are removed as well as special tokens such as URLs, words with special characters, etc. Generic premodifiers are also detected, by automatically identifying their level of specificity within a given domain, and purged.

Subsequently, CTs are ranked according to their goodness, that is, their termhood level. Goodness is measured on the basis of a candidate's domain-specificity and the level of cohesion amongst its constituents. The level of specificity of a term (labelled as *confidence* by the authors) is defined as:

[8] Application Programming Interfaces

$$C(T) = \alpha * TD(T) + \beta * TC(T)^{9} \qquad [2]$$

where *TD* stands for term domain-specificity, *TC* for the term's cohesion, and α and β "are constant values which decide the relative contributions of *TD* and *TD* respectively" (Park et *al.*, 2002: 5).

Concerning the evaluation of *Textract*, it is carried out both mechanically and resorting to the help of three judges. Human validation turns out to be more successful as the specialists confirm that 216 (72%) amongst the top 300 candidates extracted are TTs. As for automatic validation, the authors establish the level of overlap between the CTs extracted by their tool and two well-known measures: Church and Hank's (1990) mutual information and Dunning's (1993) log-likelihood. The results of this comparison yield 17.55% overlap for the former as opposed to 55.33% for the latter.

3.4.3. *TermoStat (Drouin, 2003)*

Drouin designs *TermoStat*, a free online software, for automatic term extraction in French, English, Spanish, Italian and Portuguese which can process raw text files up to 30 Mb. He employs a hybrid technique to detect both single and multi-word CTs and rank them according to their level of specialisation. Its main aim is to reduce the amount of noise produced by other automatic methods by cutting down on the number of items included in the lists generated by the system. With this purpose, the author establishes a test-value threshold of +3.09 "which means that probability of finding the observed frequency is less than 1/1000" (Drouin, 2003: 101) acting as a cut-off point between terms and non-terms.

TermoStat also employs Schmid's *Tree Tagger* as lemmatiser and POS tagger, thus producing a list where not only is the term's specificity value recorded but also its frequency as lemma, its variants, and its POS tag. The lexical categories identified by *TermoStat* are: nouns, adjectives, adverbs and verbs. It also detects MWTs having nouns and adjectives as phrase heads.

Based on previous work on lexicon specificity such as Muller's, Lafon's, or Lebart and Salem's (in Drouin, 2003), Drouin claims that the frequency of technical terms in a specialised context differs, in one way or other, from the same value in a general environment and that "focusing on the context surrounding the lexical items that adopt a highly specific behaviour ... can help us identify terms" (Drouin, 2003: 100).

The author uses a corpus comparison approach which provides information on a CT's standard normal distribution giving "access to two criteria to quantify the specifity of the items in the set … because the probability values declined rapidly, we decided to use the test-value since it provides much more granularity in the results" (Drouin, 2003: 101).

He applies human and automatic validation methods to evaluate the levels of precision and recall of his software. The author resorts to three specialists who identify the true terms (TT) from the list generated by *TermoStat* noticing that subjectivity played a relevant role in this evaluation phase and that it might also be interesting to study human influence on validation processes. Regarding automatic validation, he compares the lists of CTs with a telecommunications terminology database. *TermoStat* reaches 86% precision in the extraction of SWTs. The author insists on the importance of complementing these methods with others that help identify the meanings of those words which activate a specialised sense in a specific context.

[9] For more details on the calculation of *TD* and *TC*, see Part et *al.*, 2002: 5.

Drouin's method can be configured to recognise both SWTs and MWTs. For this section, the parameters were adjusted so that it only extracted noun and adjective phrases, which is the type of MWT pattern this method concentrates on. The results obtained by the author after evaluating MWT extraction are poorer than those obtained in SWT recognition. While *Termostat* manages to detect 81% SWTs on average, it fails to detect 35% MWTs. As a solution to solve this problem, Drouin points at the possibility of resorting to other types of statistical measures like mutual information (Church & Hanks, 1990) or termhood-weighting factor (Frantzi & Ananiadou's 1997; Nagakawa & Mori's, 2002).

3.4.4. *TermExtractor (Sclano & Velardi, 2007)*

Sclano and Velardi's method introduces an evaluation process different from other ATR methods. The results obtained by *TermExtractor*, the free online tool developed by the authors, are assessed "by web communities and individual users on different domains" (Sclano & Velardi, 2007: 6). The online software interface allows the creation of a team of judges who will validate the results obtained once a given corpus has been processed and a list of CTs produced. The average precision attained having consulted both private and public institutions (such as Stockholm University, the University of Ottawa or the Institute of Systems Analysis and Computer Science in Rome, amongst many others), as well as private users, was 80%, reaching a peak of 99.4% for a group of texts (7680) belonging to the field of anatomy and medicine.

TermExtractor manages to identify terms based on two distinct phases. The first one, linguistic, consisting in the extraction of typical patterns from a collection of specialised texts, basically noun-noun, adjective-noun or noun-preposition-noun after automatically parsing the text. The parsing process gives greater relevance to those elements which are highlighted by any means (underlining, bold types, etc.).

The second one consists in the application of several filters. Domain relevance is one of them. It is an entropy-based[10] measure which takes into account a candidate's frequency in the specialised domain by comparison with other domains. Domain Consensus (introduced by the authors in Navigli and Velardi, 2002) is also entropy-related and "simulates the consensus that a term must gain in a community before being considered a relevant domain term". Lexical cohesion is another parameter affecting term extraction. The authors follow Park *et al.*'s model (in Sclano and Velardi, 2007: 3), which measures the degree of unithood amongst the constituents of a given pattern. Finally, they employ a set of measures to filter the results with the aim of minimising noise levels (removal of generic modifiers and proper nouns, mispelling detection, etc.).

Based on all these steps, a word's weight is defined according to Sclano and Velardi "as a linear combination of the three main filters" (Sclano & Velardi, 2007: 3). Let t be the CT in question, D_i the domain of interest, *DR* the domain relevance filter, *DC* domain consensus and *LC* lexical cohesion. "The coefficients are user-adjustable, but the default is $\propto = \beta = \gamma = 1/3$" (Sclano & Velardi, 2007: 3).

[10] Shannon's Entropy is the key element of Information Theory and represents a way to measure the information in a message. In Statistics, Entropy measures the disorder of a distribution.

$$w(t, D_i) = \propto \cdot DR + \beta \cdot DC + \gamma \cdot LC^{11} \qquad [3]$$

3.4.5. *Terminus 2.0 (Nazar & Cabré, 2012)*

Nazar and Cabré propose an ATR method, freely available online, whereby term extraction becomes a fast and easy task. *Terminus 2.0* offers different possibilities for the researcher working on specialised terminology. As indicated on the website guide, it has various functions such as textual corpus search, compilation and analysis; term extraction; glossary and project management; database creation and maintenance and dictionary edition.

Their ATR method is based on the assumption that the system can learn how to recognise terms based on the language samples provided by the user. The expert does not need to formulate rules to help the system work but rather let it learn from the real samples provided of both specialised terms and general language using the latter for comparison.

The program "develops a statistical model with an abstraction of the main characteristics of both samples" (Nazar & Cabré, 2012: 210). As it is open to any user who can upload glossaries and corpora to help the system learn to identify terms in different domains, the more users employ it, the greater its capability will become to identify terminological units. As stated by the authors, the greatest innovation of this method is its collaborative character since it "allows a community of terminologists to share knowledge acquired by the program in each training phase" (Nazar & Cabré, 2012: 212).

The method applied by the system is structured into three distinct phases: syntactic, lexical and morphological. To begin with, using Schmid's *Tree Tagger* (1995), the texts are POS tagged and a syntactic model is developed based on the frequency of distribution of the syntactic patterns identified. After doing so, the frequency of the lexical units displaying those patterns is measured. Finally, it extracts initial and final character *n*-grams. The termhood score is obtained by assigning a higher value to those units which have a "significant frequency in the LSP[12] training material with respect to the general language corpus" (Nazar & Cabré, 2012: 212). This process is followed for all levels of training.

The authors act as judges to validate their method by confirming the candidates extracted as TTs and discarding those which do not qualify as such. The corpus employed as the training set is a 300,000 word collection of papers on corpus linguistics published in 2010. The test corpus is also a collection of papers on the same topic of similar size (340,000 words). Both sets of texts were taken from the scientific journal *Computational Linguistics.* The reference corpus consists in a 2 million-word collection of press articles from the Leipzig Corpora Collection. In the evaluation process the algorithm is trained also using *n*-gram frequency lists and word association measures.

As part of this training, the authors validate 800 terminological units and train the algorithm using this list of terms (both SWTs and MWTs). Once the training phase is accomplished, the study corpus is processed employing the information derived from the training. For the validation of the results obtained after processing the study corpus of 340,000 words, the authors resort to three

[11] For more information on how to calculate the value for each filter see Park et *al*., 2002: 2-3.
[12] Language for Specific Purposes

different classical measures, namely, chi-square test, mutual information and frequency (the most frequent 1500 bigrams are extracted). They also employ a stop word list to filter the results.

As a result, the precision levels achieved are considerably better than those attained by the three methods used for comparison. *Terminus* reaches 85% precision for the top 200 candidates and 75% for the top 400.

4. Results and discussion

The lists obtained by applying the evaluated methods varied in size depending on the configuration of the different parameters available for each of them. Owing to the need to establish a similar method of comparison, the number of candidates evaluated in all cases was 1400 due to the fact that *Termextractor* by Sclano and Velardi (2007) established a cut-off point producing a maximum amount of 1400 CTs. In spite of the bigger size of the output lists generated by *Terminus* or *C-value*, this was the limit set for the validation of the five methods assessed in this section.

It must be highlighted that the five lists had to be supervised manually once the automatic comparison made with the specialised glossary was finished with the purpose of minimising silence throughout this evaluation process. Two specialised dictionaries (Alcaraz & Hughes, 2000; Saint Dahl, 1999) were employed for such supervision As a matter of fact, those MWTs not present in the glossary were added to it for the validation process to be as reliable as possible. This manual supervision of the CT lists increased the number of items in the glossary by 3.4%.

ATR METHOD	**AVERAGE PRECISION 2,000 CTs**	**PRECISION TOP 200 CTs**
TermoStat (Drouin, 2003)	73%	88%
Kit and Liu (2008)	64%	84%
Keywords (Scott, 2008)	62%	85%
TF/IDF (Sparck Jones, 1972)	57.35%	74.5%
Chung (2003)	42.45%	48.5%

Table 2. *Average precision reached by SWT recognition methods*

Table 2 presents the results obtained by Marín (2014a) in the evaluation of SWT recognition methods. It was Drouin's (2003) *TermoStat* which exceled the other four methods in identifying the highest number of true legal terms, not only on average but also for the top 200 CTs, managing to extract 73% TTs from the legal corpus used in her study. On average, Kit and Liu's (2008) and Scott's methods stand 10 points below Drouin's, being closer to it for the top 200 CTs in their

output lists, since the three of them obtain considerably high rates of success in legal term identification reaching 88%, 85% and 84% precision respectively within that frame.

The worst performing methods evaluated in this experiment are Sparck Jones's (1972) and Chung's (2003), although the former achieves higher precision levels at the top of the CT output list. However, Chung's method appears to be particularly ineffective, only managing to identify 48.5% TTs within the top 200 CTs. Marín (2014a) argues that such low rates might be due to domain-dependence owing to the obvious formal differences existing between the English terminology used in Anatomy and Law (Chung's experiment is carried out using an anatomy corpus).

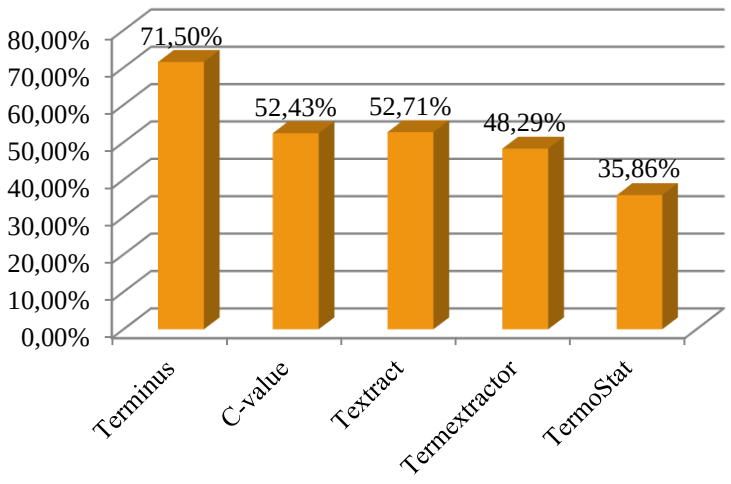


Figure 1. *Average precision of MWT recognition methods*

As far as average precision in MWT recognition is concerned, figure 1 clearly shows how *Terminus* (Nazar and Cabré, 2012) is the best performing method. It stands 19 points above the second ranking method, *Textract* (Park *et al*., 2002), which reaches 52.71% precision. *C-value* (Frantzi and Ananiadou, 1999) is in third position at 52.43%, being closely followed by *Termextractor* (Sclano and Velardi, 2007), at 4 points below. *Termostat* (Druin, 2003) is the worst performing MWT recognition method, which only manages to identify 35.86% MWTs in the corpus. As acknowledged by the author, the efficiency of his method (the most efficient one in SWT recognition according to Marín (2014a), as shown in table 2) when configured to also detect MWTs is much lower, decreasing from 88% (its average precision rate in SWT recognition) to 65% (when including both types of lexical units). The results obtained in this research confirm this fact, as Drouin's method was configured to identify MWTs exclusively, hence its lower precision.

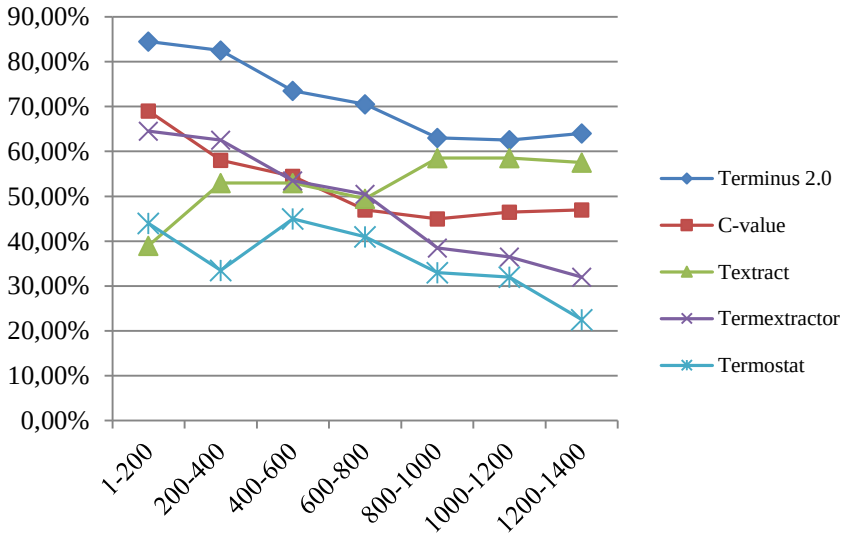


Figure 2. *Cumulative precision of MWT recognition methods for top 1400 CTs*

As revealed by figure 2, the five methods considered for evaluation behave differently as regards cumulative precision and, although their efficiency is, in general, lower than the one achieved in SWT identification (see table 2), they do not seem to reduce it as sharply as SWT recognition methods. While the precision levels reached by the latter go down 28 points on average from candidates 1 to 1400 (except for Drouin's method, which varies slightly from 84% to 82%, and Chung's whose performance is really poor from the beginning of the CT list), those identifying single and MWTs do it in a smoother manner. On average, except for *Terminus* (which goes down only 20 points), single and MWT recognition methods decrease their efficiency by 24 points within the same range (from CT 1 to 1400). Conversely, in the group of CTs 1 to 200, *Textract* improves its performance increasing its precision by 18 points although it does not manage to identify more than 57% TTs on average. Precisely due to the excellent results obtained on the top 400 CTs (83.5% precision), *Terminus* falls down by 9% from CTs 400 to 600 and continues to descend progressively from that point to the end of the graph still remaining in the first position at the end of it (64% precision).

From CTs 1 to 500, the best ranking methods are *Terminus* and *C-value*, although the latter stands 24 points below the former within this range. *Termextractor* remains in third position from

CTs 1 to 700 decreasing its effectiveness from that point to the end of the graph and moving to fourth position from that point on. Finally, *TermoStat* is the worst performing of the five methods evaluated owing to its initial configuration, which excludes SWT detection.

Taking into consideration the results obtained in this comparison between SWT and single and MWT recognition methods, it has been proved that the former are more efficient than the latter. As a matter of fact, except for *Terminus*, which behaves similarly to *Termostat* within the top 600 candidates in the list, the rest of them are far below SWT extraction methods.

Secondly, as far as corpus comparison is concerned, while it yields better results in SWT recognition, it cannot be concluded that it affects MWT recognition positively as three of the five methods assessed above which resort to it rank first, third and fifth respectively. Moreover, it cannot be affirmed either that the greater rate of success of *Terminus* is directly related to the comparison of a general and a specialised corpus but rather to the fact that the system learns about specialised terms when trained by the user, being much more efficient in their identification than other methods which do not implement any learning algorithm.

Thirdly, all the single and MWT recognition methods examined above employ lemmatisation and POS tagging techniques, due to the fact that grammatical patterns need to be identified prior to MWT recognition. Therefore, unlike SWT recognition methods, where lemmatisation yields better results (Atuhor, 2014), it cannot be considered as a relevant factor affecting precision for obvious reasons.

From the evidence provided above, it could certainly be argued that single and MWT recognition presents greater difficulties than the identification of SWTs solely, as Drouin (2003) already affirms. Given his poor results in this area, he suggests applying other measures like mutual information (Church and Hanks, 1990) to improve the performance of his method.

This point is further supported by the fact that the output lists produced by *C-value* (Frantzi and Ananiadou, 1999) and *Textract*, (Park *et al*., 2002), the second and third ranking methods as far as average precision is concerned, had to be purged manually owing to their identification of a large percentage of proper names as CTs (Park *et al*. also had to remove these elements manually before validating their method). This finding could be attributed to the English variety and the genre the corpus is based on. Judicial decisions are characterised by the constant reference to the parties' names and cases are cited basically by using the proper names of the defendants and claimants, appellants and respondents, and the like. Therefore, those methods based on corpus comparison which use frequency in the general and specialised fields as one of the key elements in term identification might identify these lexical items wrongly. Probably, those ATR methods giving greater prominence to other parameters such as the distribution of certain patterns and word types across corpora may not come up with such a high percentage of these items in their CT lists, such as Drouin's (2003) or Nazar and Cabré's (2012).

Another element which could possibly explain the lower rate of success of ATR methods in the legal field could be related to the polysemic character of some of these terms, the so-called *sub-technical* terms (either compound or simple), which could be defined as those specialised vocabulary items whose presence in the specialised and general fields is reasonably relevant, often acquiring a technical meaning when in contact with the specific context. Their high frequency in both contexts as well as their change of meaning in the specialised field might be a hindrance for their automatic detection.

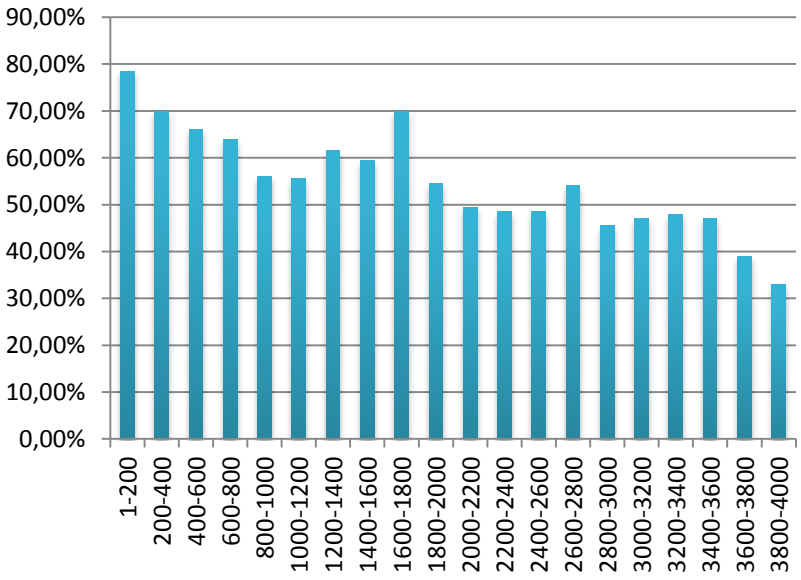
A word like "battery", which frequently occurs in both general and legal English[13], would be associated to completely different words in both contexts. Whereas "assault", "claim" or "damages" could be found amongst its most frequent legal collocates, "operate", "charged" or "powered" are at the top of its general collocate list. Probably, those ATR methods which give preference to the comparison of collocates in both contexts might also find it harder to spot legal terms which prove to be polysemic, since they do not change in form, only in meaning. Nevertheless, further research should be done in this respect to confirm this perception.

5. Processing of the *BLaRC*: identification of single and multi-word terms

After having evaluated the multi-word term recognition methods above, *Terminus*, the ATR method designed by Nazar and Cabré (2012), has proved to be the best performing one which manages to identify 71.5% terms in the *UKSCC*, achieving 83.5% precision on the top 400 candidates. Therefore, it was employed to analyse the *BLaRC*, an 8.85 million-word legal corpus which the *UKSCC* was obtained from, as already stated.

In order to minimise the amount of noise generated by the method, the output list of 5000 CTs was manually supervised to ensure that the automatic validation process had worked properly, two specialised dictionaries (Alcaraz & Hughes, 2000; Saint Dahl, 1999) were also employed for this process of supervision. Those CTs which were confirmed as TTs but did not appear in the glossary and had thus been automatically discarded were added to it and therefore confirmed as terms. Consequently, the silence generated by the automatic comparison with the glossary was kept to a minimum.

This manual supervision also led to the elimination of repeated words. Terminus lemmatises types not assigning a given weight to each lemma but to its variants. It includes the different forms of a lemma separately in the output list (indicating the lemma they are associated with) in spite of such forms often belonging to the same morphological category. This might be a problem area for this method which could possibly increase its efficiency if lemmas were considered as single units and their variants were not assigned different weight depending on their forms. An example of this shortcoming is the word "landowner" whose weight in singular is 3576.60925 and 2185.525021 in plural ("landowners"). As a result of this, the variants of the every lemma were removed from the list to assess precision, leading to the elimination of 671 word forms from the original CT list.



eneral English
nd reference

and multi-

Figure 3 shows the results of the validation process after comparing the whole list of CTs with the gold standard automatically and also supervising it manually. The graph illustrates cumulative precision in groups of 200 candidates from items 1 to 4000 ranked according to the weight assigned to each of them by *Terminus*.

Nazar and Cabré's method does not establish a threshold to discriminate terms from non-terms as clearly as other methods like Drouin's (2003) or Chung's (2003). However, it can be configured so that the number of candidates adjusts to the preferences of the user. In this case, it was configured to produce 5000 terms, therefore, the graph above illustrates the evaluation of the first 4000 candidates once the repeated word forms had been eliminated, as already stated.

Terminus remained considerably efficient (especially considering the precision levels achieved by the other MWT recognition methods assessed above) from CTs 1 to 1800, managing to identify 64.5% single and multi-word terms on average within this range and finding its peak at 78.5% for the top 200. Its effectiveness decreases progressively recovering again from CTs 1600 to 1800 at 70% precision. From that point on, it falls sharply to 54.5% and continues to descend smoothly to 48.5% (CTs 2400 to 2600), slightly recovering from candidates 2600 to 2800 (54% precision) and finally falling to 33% by the end of the graph.

Having observed the evolution of this method, it might be interesting to try and establish a cut-off point which would act as a threshold to discriminate terms from non-terms. Judging by the figures, it appears that the method is considerably efficient up to candidates 1600 to 1800 since, after that point, it does not manage to recognise more than 46.77% terms and its precision level decreases rapidly to 33% from candidates 3800 to 4000.

The weight corresponding to CT 1800 is 1030 and could thus be regarded as the threshold value. Applying this threshold, Terminus could extract 1153 TTs reaching 65% precision on average.

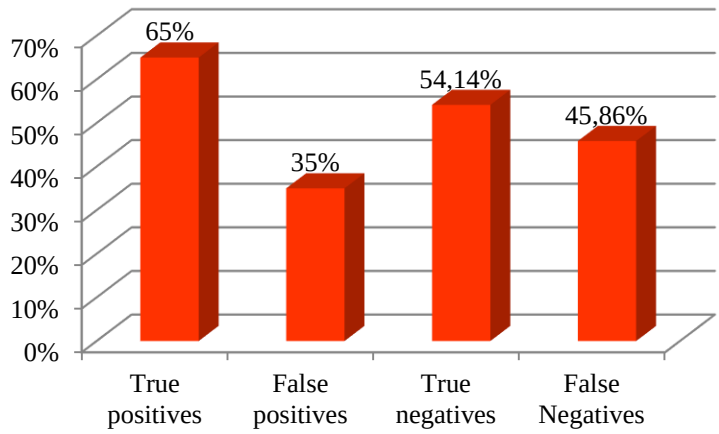


Figure 4. *True and false positives, true and false negatives after applying a +1030 weight threshold*

As illustrated by figure 4, if a +1030 threshold was applied, the percentage of TTs extracted would reach 65%, while noise levels (percentage of false positives generated by the system) would stand at 35%. Conversely, establishing a threshold would affect the amount of false negatives, that is, silence, since it would fail to identify 45.86%.

Finally, owing to the fact that it was not possible to have a definite list of MWTs extracted from the *BLaRC* to use as reference to calculate recall, partial recall could only be established if we considered the whole list of terms generated by Terminus itself, that is, 2312. If the + 1030 weight threshold was applied, Terminus would achieve 49.87% recall with respect to the whole list of terms generated by the system without establishing any cut-off point. Appendix 1 offers a sample of the single and multi-word terms identified by *Terminus.*

6. Conclusion

The present study has been motivated by several factors related to the peculiarities of legal terminology and its automatic recognition. As stated in the introduction, almost half of the SWTs extracted from the *BLaRC*, a legal English corpus of judicial decisions, could also be found in general English vocabulary lists. As a consequence, ATR methods based on frequency data and corpus comparison might find greater difficulties in extracting specialised terms from a legal corpus owing to the relevance of legal terms in everyday language. In spite of that fact and to the best of our knowledge, there are no studies dedicated to the evaluation of ATR efficiency in the legal field except for Marín (2014).

Consequently, this research was conceived as a follow-up of Marín's (2014) work whereby five different SWT recognition methods were assessed with the aim of identifying the most reliable ones in legal term identification. A comparison between the results presented by this author and the ones obtained after evaluating the precision levels achieved by five ATR methods focusing on single and MWT identification was carried out. On average, it was found that SWT recognition methods are more effective than those designed to extract both single and MWTs. Whereas Drouin's *TermoStat* (2003) turned out to be the best performing method in SWT identification (it recognised 73% TTs out of a list of 2,000 CTs), its reliability in extracting multiword units stood at almost 40 points below the precision percentages achieved in the identification of SWT, as acknowledged by the author. On the other hand, Nazar and Cabré's *Terminus* (2012) managed to identify 71.5% TTs in the *UKSCC*, the pilot legal corpus, basically because of its training algorithm, which must be applied prior to the actual implementation of the method itself. By uploading a sample of single and multi-word legal terms to the system, it learnt to identify these terms and exceled by far the other ATR methods tested above.

In relation to the peculiarities of legal English and the possible difficulties that the texts in the *BLaRC*, the legal corpus which *UKSCC* belonged in, might cause in legal term recognition, it was found that the CT lists produced by *C-value* (Frantzi and Ananiadou, 1999) and *Textract* (Park *et al*., 2002) had to be purged manually due to the massive presence of proper nouns in them (Park *et al*. eliminate them manually before evaluating their results). This finding is directly related to the textual genre the corpus is based on, since proper nouns are mentioned repeatedly in case citation and within judgements. Probably, the higher occurrence of these words or their distribution, behaving differently from other vocabulary items owing to their uniqueness (they design one single entity as proper nouns), might have led to their wrong identification.

In the fifth section of this paper, *Terminus* (Nazar and Cabré, 2012), the best performing single and MWT recognition method, was implemented on the *BLaRC* so as to produce a reliable list of both single and multi-word legal terms, which might be exploited in such fields as ESP instruction, legal terminology research or translation. The results obtained with *Terminus* could be improved if

all the members of the same word families under the same headword, the lemma, were not assigned a separate weight but were rather grouped under it, as *TermoStat* (Drouin, 2003) does in SWT identification. If a +1030 weight threshold was establish as a cut-off point in CT recognition, the results would also improve and precision would stand at 65% for this corpus. As regards partial recall, supposing this threshold was applied, *Terminus* would achieve almost 50%.

As further research, it would be worthwhile to investigate the relationship between the so called *sub-technical* terms (which are shared by the general and specialised contexts) and term mining. Their high frequency in both contexts, in spite of their specialised character in the legal field, as well as their polysemic character might pose certain difficulties in their automatic detection. Moreover, the acquisition of new technical meanings when they enter in contact with the legal environment necessarily affects the collocate networks which they may generate in the general and specialised contexts, although they do not vary in form. This fact might also become a problem area for those ATR methods focusing on these parameters.

REFERENCES

Ahmad, K., Davies, A., Fulford, H., Rogers, M., "What is a term? The semi-automatic extraction of terms from text", in Snell-Hornby, M., Pöchhacker, F. and Kaindl, K. (eds.), *Translation Studies: An Interdiscipline*, Amsterdam: John Benjamins, p. 267-278, 1994.

Alcaraz, E. and Hughes, B. *Diccionario de Términos Jurídicos.* Barcelona: Ariel Referencia, 2000.

Ananiadou, S. *A Methodology for Automatic Term Recognition.* PhD Thesis, University of Manchester Institute of Science and Technology: United Kingdom, 1988.

Arson, A., Lang, F. "An overview of MetaMap: historical perspective and recent advances", *Journal of American Medical Informatics Association*, vol. 17, no. 3, p. 229-236, 2010.

Barrón-Cedeño, A., Sierra, G. E., Drouin, P. and Ananiadou, S. "An Improved Automatic Term Recognition Method for Spanish", in Gelbukh, A. (ed.), *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*, Springer, p. 125-136, 2009.

Bernier-Colborne, G. "Defining a Gold standard for the evaluation of Term Extractors", in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Instambul, 2012.

Borja Albí, A. *El texto jurídico en inglés y su traducción.* Barcelona: Ariel, 2000.

Bourigault, D. "Surface grammatical analysis for the extraction of terminological noun phrases", in *Proceedings of the 5th International Conference on Computational Linguistics.* Nantes, p. 977-81, 1992.

British National Corpus, version 3 (*BNC* XML Edition), distributed by Oxford University Computing Services on behalf of the *BNC* Consortium, 2007.

Cabré, M. T., Estopà, R., Vivaldi, J. "Automatic term detection: a review of current systems", in Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds.), *Recent Advances in Computational Terminology*, Amsterdam: John Benjamins/ Natural Language Processing, p. 53-88, 2001.

Chung, T. M. "A corpus comparison approach for terminology extraction", *Terminology*, vol. 9, no.2, p. 221-246, 2003.

Church, K.W., Hanks, P. "Word association norms, mutual information, and lexicography", *Computational Linguistics*, vol. 16, no.1, p. 22-29, 1990.

Dagan , I., Church, K. "TERMIGHT: Identifying and Translating Technical Terminology", *Proceedings of the 4th Conference on Applied Natural Language Processing*, p. 34-41, 1994.

Daille, B. "Study and implementation of combined techniques for automatic extraction of terminology", in Klavans, J.L., Resnik, P. (eds.), *The Balancing act: Combining symbolic and statistical approaches to language*. Cambridge, MA: MIT Press, p. 49-66, 1996.

David, S., Plante, P. *Termino 1.0*. Research Report of Centre d'Analyse de Textes par Ordinateur. Université du Québec, Montréal, 1990.

Drouin, P. "Term extraction using non-technical corpora as a point of leverage" *Terminology*, vol. 9, no.1, p. 99-117, 2003.

Fahmi, I., Bouma, G., van der Plas, L. "Improving statistical method using known terms for automatic term extraction", in *Proceedings of Computational Linguistics in the Netherlands* (*CLIN 17*), p. 1-8, 2007.

Flowerdew, L. "Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, vol. 14, no. 3, p. 393-417, 2009.

Frantzi, K.T., Ananiadou, S.

--------- "Extracting nested collocations", in *Proceedings of the 16th Conference on Computational Linguistics* vol. 1, p. 41-46, 1996.

---------. "The c/nc value domain independent method formulti-word term extraction". *Journal of Natural Language Processing*, vol. 3, no. 2, p. 115-127, 1999.

Heatley, A., Nation, P. *Range* (computer software). Wellington, New Zealand: Victoria University of Wellington, 1996.

Jacquemin, C. *Spotting and discovering terms through NLP*. Massachusetts: MIT Press, 2001.

Justeson, J.S., Katz, S.M. "Technical terminology: some linguistic properties and an algorithm for identification in text", in *Natural Language Engineering,* vol. 1, no. 1, p. 9-27, 1995.

Kit, C., Liu, X. "Measuring mono-word termhood by rank difference via corpus comparison", *Terminology*, vol. 14, no. 2, p. 204-229, 2008.

Lemay, C., LHomme, M.C., Drouin, P., "Two Methods for Extracting 'Specific' Single-word Terms from Specialised Corpora: Experimentation and Evaluation", *International Journal of Corpus Linguistics* vol. 10, no. 2, p. 227-255, 2005.

Loginova, E., Gojun, A., Blancafort, E., Guegan, M., Gornostay, T., Heid, U. "Reference Lists for the Evaluation of Term Extraction Tools", in *TKE 2012: Terminology and Knowledge Engineering*, Madrid, p. 177-192, 2005.

Marín, M.J. and Rea, C. "Structure and design of the BLRC: a legal corpus of judicial decisions from the UK". *Journal of English Studies*, vol. 10. La Rioja: Servicio de Publicaciones de la Universidad de La Rioja, p. 131-145, 2012.

Marín, M.J.

----------"Evaluation of five single-word term recognition methods on a legal corpus". *Corpora, vol. 9, no.1, p. 83-107*, 2014a.

----------"A Proposal to Exploit Legal Term Repertoires Extracted Automatically from a Legal English Corpus". *Miscelánea: A Journal of English and American Studies, vol. 49*, Zaragoza: Universidad de Zaragoza. (in the press, 2014b).

McEnery, T., Wilson, A. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 2001.

McEnery, T., Xiao, R., Tono, Y. *Corpus-based language studies: an advanced resource book*. Routledge Applied Linguistics: New York, 2006.

Maynard, D. and Ananiadou, S. "TRUCKS: A model for automatic multi-word term recognition". *Journal of Natural Language Processing* vol. 8, no. 1, p. 101–125, 2000.

Mellinkoff, D., *The Language of the Law*, Boston: Little, Brown & Co, 1963.

Mondary, T., Nazarenko, A., Zargayouna, H., Berreaux, S. "The Quaero Evaluation Initiative on Term Extraction", in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Instambul, p. 663-669, 2012.

Nakagawa, H., Mori, T. "A simple but powerful automatic term extraction method', in *COLING-02 on COMPUTERM. Proceedings of the Second International Workshop on Computational Terminology*, pp. 1-7, 2002.

Nazar, R., Cabré, M.T. "Supervised Learning Algorithms Applied to Terminology Extraction", in Aguado de Cea, G., Suárez-Figueroa, M.C., García-Castro, R., Montiel-Ponsoda, E. (eds.), *Proceedings of the 10th*

Terminology and Knowledge Engineering Conference (TKE 2012). Ontology Engineering Group, Association for Terminology and Knowledge Transfer. Madrid, p.209-217, 2012.

Orts, M.A. *Aproximación al* discurso jurídico *en inglés. Las pólizas de seguro marítimo de Lloyds*. Madrid: Edisofer, 2006.

Pearson, J. *Terms in Context*. Amsterdam: John Benjamins Publishing Company, 1998.

Panzienza, M.T., Pennacchiotti, M., Zanzotto, F.M. "Terminology extraction: An Analysis of Linguistic and Statistical Approaches", in *Studies in Fuziness and Soft Computing*, vol. 185, p. 225-279, 2005.

Park, Y., Byrd, R.J., Boguraev, B. "Automatic Glossary Extraction: Beyond Terminology Association", in *Proceedings of COLING "02 19th International Conference on Computational Linguistics*, Taipei, p. 1-7, 2002.

Rea, Camino. **"**Getting on with Corpus Compilation: from Theory to Practice", *ESP World*, vol.1, no. 27, p. 1-23, 2010.

Robertson, S. "Understanding Inverse Document Frequency: On theoretical arguments for IDF", in *Journal of Documentation* vol. 60, no. 5, p. 503-520, 2004.

Saint Dahl, H. *Dahls Law Dictionary. Diccionario Jurídico Dahl.* New York: William S. Hein & Co., Inc, 1999.

Schmid, H. "Improvements in Part-of-Speech Tagging with an Application to German", in *Proceedings of the ACL SIGDAT-Workshop*. Dublin, 1994.

Sclano, F., Velardi, P., "A Web Application to Learn the Common Terminology of Interest Groups and Research Communities" in *Proceedings of the Conference TIA-2007*, Sophia Antipolis, 2007.

Scott, M. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software, 2008.

Sinclair, J. "Corpus and Text: Basic Principles", in Wynne, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice, AHDS Literature, Languages and Linguistics: University of Oxford, Chapter 1*, 2005.

Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, vol. 28, no. 1, p. 11-21. 1972.

Tiersma, P. 1999. *Legal Language.* Chicago: The University of Chicago Press.

Vargas, Chelo. *Aproximación terminográfica al lenguaje de la piedra natural: propuesta de sistematización para la elaboración de un diccionario traductológico*, unpublished PhD thesis, Universidad de Alicante, 2005.

Vivaldi, J. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas.* PhD thesis, Universidad Politécnica de Cataluña, 2005.

Vivaldi, J., Cabrera-Diego, L.A., Sierra, G., Pozzi, M. "Using Wikipedia to Validate the Terminology Found in a Corpus of Basic Textbooks", in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Instambul, p. 3820-3827, 2012.

West, M. *A General Service List of English Words*. London: Longman, 1953.

APPENDIX 1

LEMMA	WEIGHT
UNLAWFUL DISCRIMINATION	9782,755472
CONCURRENT	9781,372031
CAUSATIVE	9629,090509
PERSUADE	9584,124739
HEREDITAMENT	9355,16657
DISCRIMINATOR	9169,941848
STATUTORY PROCEDURE	9126,142578
RENUNCIATION	9073,291148
NOTARY	9003,557439
DISCHARGE	8954,219105
TESTATOR	8921,177807
ATTRIBUTION	8870,669195
IMPUTE	8751,069492
DETAINEE	8559,713855
EX TURPI CAUSA	8381,394618
NULLITY	8367,778388
IMMATERIAL	8340,843231
CONTEND	8266,946424
CREDIBILITY	8216,235997
COMPLY	8155,083041
NUPTIAL AGREEMENT	8054,335883
INCONSISTENCY	7893,884107
PERSONAL DATUM	7880,46792
REASONABLE DOUBT	7741,812898
REPUDIATION	7617,533695
CONVICTION	7539,657385
TORTFEASOR	7437,347728
LAWFULLY	7387,042228
PROPORTION	7288,047433
DISALLOWANCE	7264,79791

FIXTURE	7235,630525
PROROGATION	7192,904206
PEREMPTORY	7179,530454
DUE COURSE	7173,686885
UNLAWFUL ACT	7153,623464
CONSIGNEE	7130,001174
ONUS PROBANDI	7124,171469
SIGNIFICANT	7121,412092
MONIES	7054,562892
INDEMNIFY	7007,328045
POSSESSION ORDER	7006,53817
TESTAMENTARY	7005,020567
COGNIZANCE	6952,597602
ANTECEDENT	6926,379656
CONFESSION	6921,594246
EXCLUSIVE JURISDICTION	6840,179787
GRAVITY	6824,93986
STATUTORY PROVISION	6815,511133
INTEGRITY	6800,634167
DUE DILIGENCE	6666,709751
SEISE	6767,15794
PROVISO	6764,398186
REASONABLENESS	6745,153213
MISBEHAVIOUR	6705,429604
WITNESS BOX	6652,736039
CONTRAVENTION	6639,148874
CONSTRUCTIVE TRUST	6538,025433
OUTGOINGS	6537,614609
BREACH	6528,64733
INCOMPATIBILITY	6436,541343
NEW EVIDENCE	6406,640402