(This is a pre-print version of the article published by the journal *Corpora*)

Evaluation of five single-word term recognition methods on a legal English corpus

Abstract

Specialised texts are characterised, amongst other features, by the presence of terminology which conveys domain-specific concepts essential for the specialist interested in analysing such texts. Automatic term recognition methods (ATR) are employed to automatically identify those terms, especially due to the large size of corpora nowadays. However, they tend to concentrate on the identification of multi-word terms (MWTs) neglecting single-word terms (SWTs) to a certain extent. This might be related to the greater number of the former found in fields such as biomedicine. However, as far as legal English is concerned, testing has shown that SWTs represent 65.22% of the items in the specialised glossary employed for the evaluation of the ATR methods examined herein. This article presents the evaluation of five single-word term recognition methods, namely, Chung's (2003), Drouin's (2003), Kit and Liu's (2008), Keywords (2008), *and TF-IDF (term frequency-inverse document frequency) which were tested on the United Kingdom Supreme Court Corpus (UKSCC), a 2.6 million-word legal corpus designed and compiled with such purpose. The results indicate that Drouin's TermoStat software is the best performing one achieving 73.45% precision on the top 2000 candidate terms.*

Keywords: automatic term recognition; specialised corpora; single-word terms; sublanguage; legal English

1. Introduction

The role played by specialised corpora as reliable sources of information to resort to for the teaching and learning of English for specific purposes (ESP) is discussed by scholars like McEnery and Wilson (1996: 121) who underline the fact that they meet the needs of ESP students better than general corpora "including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain more directly than those taken from more general language corpora". They continue to assert their advantages in exposing learners to genuine language samples and acting as reference for scholars to review existing didactic materials. Schmitt (2002: 1) affirms that their use might be beneficial regarding them as a valuable teaching resource as well as a useful tool to assess vocabulary acquisition. In addition, Gilquin and Granger (2010: 359) insist on the importance of ESL (English as a second language) learners' exposure to authentic materials based on corpora which also offer "a large number of authentic instances of a particular linguistic item" thus helping to identify their meanings depending on the context where they occur.

Conversely, Flowerdew (2009: 395) also criticises data-driven teaching methods since they have a predominantly inductive character, tending to offer decontextualised language samples extracted from corpora. She agrees with Swales and Kaltembök & Mehlmauer-Larcher (as cit. in ibid.) that "corpus linguistics techniques encourage a

more bottom-up rather than top-down processing of text in which truncated concordance lines are examined atomistically".

Nonetheless, controversy aside, the use of corpora is widespread in ESP and ESL teaching since they can contribute to a greater or lesser extent to second language acquisition yet, as far as legal English is concerned, the scarceness of didactic materials based on legal corpora is manifest. Legal English is a particularly complex branch of ESP, as underlined by scholars (Mellinkoff 1963; Alcaraz 1994; Tiersma 1999; Borja 2000; Orts 2006), which requires plenty of resources to be taught. However, the number of legal corpora available[1] is relatively small, which creates a methodological gap in the area. As a result, we designed and compiled the United Kingdom Supreme Court Corpus (UKSCC) in order to study the features of legal terminology and have recourse to it to elaborate didactic materials, amongst other possible uses.

Getting to know the terms in a specialised text definitely contributes to our understanding of the text's *aboutness*. Moreover, terminology is used to share domain-specific information amongst the members of a specialised community (Rea 2008: 77). As Kit and Liu (2008: 204) put it, "terms are linguistic representations of domain-specific key concepts in a subject field that crystallise our expert knowledge in that subject", in other words, a term is "a textual realisation of a specialised concept" (Spasic et al. 2005: 240). To Chung (2003a: 221-2), terms display distinctive features both qualitative (e.g. morphology) and quantitative (e.g. their frequency of occurrence). Hence, identifying and extracting the terms in a specialised corpus becomes an essential task when using it as a source of information for ESP teaching and learning. However, handling and processing large amounts of data is a time-consuming task, and the application of effective ATR methods is essential for the terminologist to draw reliable conclusions on the information retrieved by such methods.

This article presents the evaluation of five single-word term extraction methods tested on a 2.6 million-word legal corpus (14,654 KB), UKSCC, designed and compiled for that purpose. The study of single-word term recognition methods is justified in section 2. Section 3 describes UKSCC as a specialised corpus followed by a description of the five methods under evaluation in section 4. Section 5 is devoted to the description of the implementation and evaluation of these five methods. To finish, the major conclusions drawn from this study are offered in section 6.

2. **Why single-word terms?**

ATR methods typically concentrate on multi-word terms (MWTs) exploring the concepts of *termhood* and *unithood* from different perspectives. Nakagawa and Mori (2002: 1) define *termhood* as "the degree that a linguistic unit is related to a domain-specific concept". According to Kit and Liu (2008: 205), *unithood* establishes "how likely a candidate is to be an atomic linguistic unit". Nevertheless, these authors consider that *unithood* only serves as a way of discarding those units not displaying a high level of cohesion amongst their possible constituents but does not provide any information about their degree of specificity.

In the past, the literature on ATR methods and software tools has been profusely reviewed (Maynard and Ananiadou 2000; Cabré et al. 2001; Drouin 2003; Lemay et al. 2005; Panzienza et al. 2005; Chung 2003a, 2003b; 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

[1] See Marín and Rea (2011) for a review on legal corpora.

statistical information, amongst them: Church and Hanks (1990), Ahmad et al. (1994), Nakagawa and Mori (2002), Chung (2003a), Fahmi et al. (2007), Scott (2008a) 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 evaluation 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 (as cit in ibid.) have worked on this area although there is still much to be done in this respect.

In spite of the large number of ATR methods existing to date, very few concentrate on single-word terms (SWTs), which are neglected to a certain extent assuming that they are easily identifiable specially due to the fact that such parameters as unithood do not need to be considered. Neverhteless, as remarked by Lemay et al. (2005), ignoring SWTs implies taking for granted that most specialised terms are multi-word units. Nakagawa and Mori (2002: 1) emphasise this idea by giving concrete data on the percentage of MWTs in specific domains: "The majority of domain specific terms are compound nouns, in other words, uninterrupted collocations. 85% of domain specific terms are said to be compound nouns."

However, this does not seem to be the case of legal English because, having thoroughly studied the legal glossary, which was compiled by merging and filtering four different specialised glossaries of British and American legal English[2], 65.22% of 8715 terms in the list are SWTs.

The term *SWTs* will be used hereinafter to refer to those lexical units which can convey a domain-specific concept by themselves regardless of the lexical category they belong to. As a result, the evaluation of the methods presented below will include the four main lexical categories of the language, namely, nouns, verbs, adjectives and adverbs.

3. UKSCC: The United Kingdom Supreme Court Corpus

UKSCC is a legal corpus of law reports (collections of judicial decisions) which has been compiled according to corpus linguistics standards as stated in Sánchez et al. (1995) and Wynne (2005) for general corpora and their adaptation to specific corpora (Pearson 1998; Rea 2010). It is a 2.6 million-word specialised corpus subset of a larger one: BLaRC (*British Law Report Corpus*), which is still in its compilation phase.

[2] Both English varieties have been included in the glossary although the corpus is formed basically by British texts owing to the fact that, having observed the texts closely before starting any evaluation procedure, some of them, due to the nature of the claim, appeal, etc., included American terminology. As a matter of fact, although there are obvious differences, both BrE and AmE have many legal terms in common as shown in specialised dictionaries and glossaries.

The reasons to focus on this genre to study the linguistic properties of legal terminology are varied. To begin with, 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, the acts passed at their parliaments have gained greater importance being most often cited in case decisions. However, case law stands at the very basis of common law systems which rely on the principle of binding precedent to work, 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 dicendi*), and judicial decisions are employed by law practitioners as the basis for their arguments, decisions, etc.

Another fact that makes law reports an outstanding legal genre is that they not only cover all the branches of law, but might also present full embedded sections of other public and private law genres displaying therefore great lexical richness and variety. Following Sinclair (2005: 5) "the contents of the corpus should be selected … according to their communicative function in the community in which they arise". Consequently, such texts as these have been chosen to form the corpus due to the pivotal role they play in common law legal systems. The Supreme Court was selected as the text source owing to its relevance within the British judicial system (all the decisions made at the Supreme Court set precedent and are cited whenever applicable), and the wide lexical variety of the documents coming from it. It is at the top of the UK judicial pyramid and deals with cases belonging to all branches of law.

As for its structure, UKSCC is a synchronic, monolingual and specialised collection of 193 judicial decisions from the UK Supreme Court and the House of Lords[3] issued between 2008 and 2010. The documents included in UKSCC are authentic judicial decisions as produced by British courts in raw text format.

4. Description of the methods selected for evaluation

Drouin's (2003), Chung's (2003a; 2003b) and Kit and Liu's (2008) methods were singled-out due to the high precision levels reported (over 80%) by their authors. Moreover, except for TF-IDF −term frequency-inverse document frequency− (Sparck Jones, 1972), they all resort to corpus comparison to automatically recognise SWTs[4].

The *Keywords* tool included in the software package *Wordsmith 5* by Scott (2008a), it is not an ATR method *per se*, however, as testing will show below, it can be used as such and it does perform more accurately than others designed specifically to that end. It was chosen due to its popularity and capacity to easily process large amounts of text data providing information on a word's "importance as a content descriptor", in Biber's words (as cit. in Gabrielatos and Marchi, 2011: 5), that is to say, on its keyness. According to Scott (2008b: 184), a word is considered key "if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger word-lists".

Scott's tool was configured to apply Dunning's (1993) log-likelihood calculation (it can also employ the chi-square test to produce a keyword list) since it is a recommended option for long texts such as the ones included in UKSCC, judicial decisions. For the system to calculate a word's keyness it is necessary to resort to a reference corpus in

[3] 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.

[4] Drouin's software can identify both SWTs and MWTs but can be configured so that it only concentrates on SWTs which is the option evaluated below.

order to compare it with the specialised one whose keywords we wish to extract. The reference corpus we employed for corpus comparison to implement *Keywords* and Chung's method is LACELL (*Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía*[5]), a 21 million-word (118,105 KB) general English corpus compiled by the LACELL research team at the University of Murcia comprising mainly texts from the 1990s. It is a balanced synchronic corpus of general English including both written texts from diverse sources such as newspapers, books (academic, fiction, etc.), magazines, brochures, letters and so forth, and also oral language samples from conversation at different social levels and registers, debates and group discussions, TV and radio recordings, phone conversations, everyday life situations, classroom talk, etc. Its geographical scope ranges from USA, to Canada, UK and Ireland, however, those texts not coming from the United Kingdom were removed to avoid skewedness in the results reducing the original size to 14.9 million words (83,400 KB).
The BNC (*British National Corpus*[6]) lemmatised lists provided online by Kilgariff[7] were employed as background for reference to implement Kit and Liu's method owing to the fact that both the SC and RC had to be lemmatised[8]. Therefore, UKSCC was also lemmatised using Schmid's (1995) *Tree Tagger*[9] to apply the calculations on lemmata, not on word types[10].

Drouin designs *TermoStat*, a free online software[11] for automatic term extraction in French, English, Spanish, Italian and Portuguese which can process raw text files up to 30 Mb. They employ a hybrid technique to detect both single and multi-word candidate terms and rank them according to their level of specificity. Their 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 list generated by the system. With this purpose, they establish a test-value threshold of +3.09 "which means that probability of finding the observed frequency is less than 1/1000" (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, as shown in figure 1. The lexical categories identified by *TermoStat* are: nouns, adjectives, adverbs and verbs. It also detects MWTs having nouns and adjectives as phrase heads.

[5] For more information on the LACELL research group see: http://www.um.es/grupolacell
[6] For more information on the *British National Corpus* visit: http://www.natcorp.ox.ac.uk
[7] Provided by Adam Kilgariff at: http://www.kilgarriff.co.uk/BNClists/lemma.num
[8] The process of lemmatisation consists in retrieving a word's lemma, that is, the root word which other possible realisations of it derive from (e.g. *make* would be the lemma for *made*, *makes*, *making*, etc.). Lemma frequency must be computed by adding up the raw frequency values of all its posible variants.
[9] Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[10] The term *word type* refers to every different wordform in the corpus but not to each of its occurrences known as *tokens*.
[11] Available at: http://olst.ling.umontreal.ca/~drouinp/termostat_web/index.php

Results

List of terms | Cloud | Stat | Structuration | Bigrams

Candidate (grouping variant)	Frequency	Score (Specificity)	Variants	Pattern
section	9694	126.29	section sections	Common_Noun
v	6828	112.55	v	Common_Noun
case	11465	111.79	case cases	Common_Noun
para	5973	108.63	para paras	Common_Noun
article	5686	97.39	article articles	Common_Noun
court	6387	88.65	court courts	Common_Noun
appeal	3993	80.3	appeal appeals	Common_Noun
appellant	3102	78.47	appellant appellants	Common_Noun
not	22062	75.07	not	Adverb
law	5484	73.55	law laws	Common_Noun
judgment	2862	71.67	judgment judgments	Common_Noun
claim	3293	69.8	claim claims	Common_Noun
right	5795	67.98	right rights	Common_Noun
apply	3542	65.5	apply applying	Verb

Fig. 1 *Screenshot of output produced after processing UKSCC with TermoStat*

Based on previous work on lexicon specificity such as Muller's (1979), Lafon's (1980), or Lebart and Salem's (1994), 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, 2003:101).
Drouin uses a corpus comparison approach which provides information on a candidate term'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).
They apply human and automatic validation methods to evaluate the levels of precision and recall of their software. The author resorts to three specialists who identify the true terms 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, they compare the lists of candidate terms 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.

On the other hand, Chung's (2003b: 53) approach to term extraction consists in establishing a threshold to discriminate terms from non-terms affirming that "to be classified as a technical term, a type had to occur at least 50 times more often in the technical text than in the comparison corpus, or only occur in the comparison corpus".
Chung reaches this conclusion after validating their method by comparison with a qualitative one, the *rating scale approach,* with the purpose of assessing the degree of overlap between it and the quantitative technique employed by the author. Thus, two experts are asked to classify the vocabulary in a 5,500 word text from their anatomy corpus, the sublanguage she works with in the design and evaluation of their method.

They classify the words into four different categories depending on their level of specialization.
In contrast, the quantitative method employed by Chung consists in calculating the ratio of occurrence of the word types in the anatomy text given to the experts. The author normalises the frequencies of the text types in both their anatomy corpus and a general one and calculates the ratio by dividing the former by the latter. Then, basing their classification on these results and on the absolute frequency figures obtained, they also produce different groups and compare them to the ones by the specialists. The results of the comparison yield 86% overlap between the author and the experts, especially regarding highly specific words and non-terms.
The author therefore concludes that this ATR method based on statistical data might be reasonably effective, although the last decision to include a word in a given category must be made by the researcher after either consulting the expert or the contexts of occurrence of a given word, since they believe that the most effective approach is the qualitative one. However, it is time-consuming and cannot be applied to large corpora for efficiency reasons.

Kit and Liu's (2008) method measures the degree of termhood of SWTs relying on a corpus comparison technique. It aims at studying the different ways words distribute in a specific subject field, namely, in a specialised 8.8 million-word legal corpus called BLIS (*Bilingual Laws Information System*) against a general domain using BNC as representative of it. Kit and Liu's ATR method focuses exclusively on SWTs, also called *mono-word terms*, basically to avoid "interference from unithood issues" (206), that is, to prevent such questions as establishing the degree of cohesion between the elements in a grammatical pattern from becoming an obstacle for the calculation of a word's level of specificity. These authors acknowledge the greater complexity of classifying a mono-word as a term owing to the fact that the structural information employed to detect the presence of MWTs in a text cannot be applied to SWT automatic mining.
Kit and Liu's method consists in obtaining the rank difference of the vocabulary items in a specialised and a general corpus "given a domain corpus D (with a vocabulary V_D) to represent a subject field and a balanced corpus B (with a vocabulary V_B) as background, the termhood of a candidate word *w* is defined as"(212):

$$\text{ʈ}(w)=\frac{r_D(w)}{|V_D|}-\frac{r_B(w)}{|V_B|}$$

The application of this formula for the calculation of ʈ -value therefore consists in introducing the rank position (r_D) of a given term in the SC (study corpus), the specialised one, and normalise it by dividing it by the total number of items in the list, that is to say, in a vocabulary of 4,500 items, the divisor would be 4,500. After that, the same calculation will be carried out using the normalised datum for the same vocabulary item (r_B) in the RC (reference corpus), the general one. Finally, the normalised value in the RC will be subtracted from the one in the SC obtaining the ʈ-value of the candidate SWT. This result will indicate its level of specialisation thus, the higher it scores, the more specialised it will be considered. Nevertheless, Kit and Liu do not establish a threshold that splits a list into terms and non-terms but rather place words along a termhood continuum "in a way that candidates with a higher termhood value would be pushed to its high end and those with a lower termhood to its low end" (2008: 212).
The method is evaluated using a specialised glossary of legal terms as the gold standard together with a list of true terms extracted from the specialised corpus which were annotated manually by legislators during the drafting of the legal documents in the

corpus. The corpus is tokenised (divided into basic text units) removing all the elements that may cause noise. The text units filtered out of the definite list belong to different categories, namely, punctuation marks, numbering items and numerical expressions and function words. The tokenisation of the corpus results into a list of 8,808,544 tokens which is filtered obtaining a definite one of 13,806 word types.

After comparing their results with those obtained applying Chung's (2003a; 2003b) frequency ratio, they realise that, although the results are similar, it becomes necessary to improve the rank difference calculation to enhance its performance. They propose two alternatives, the second one being slightly more effective. It consists in normalising both the SC and RC ranks using the sum of all the ranks in the respective corpora as the divisor as follows:

$$\mathfrak{v}_2(w)=\frac{r_D(w)}{\sum_{w'\in V_D} r_D(w')} - \frac{r_B(w)}{\sum_{w'\in V_D} r_B(w')}$$

This improved version of the rank difference performs better reaching a precision level of 98.2% on the first 500 candidate terms (out of 12000 evaluated) and 97% on the first 1000, remaining above 90% on the top 20%.

Finally, as opposed to the other four, the TF-IDF measure, used in the fields of information retrieval and text mining, does not employ corpus comparison as a means to determine a word's weight. On the contrary, it measures it by taking into consideration its frequency in a given document and the number of documents it appears in throughout a corpus. A word will display greater weight if it shows high frequency values and appears in fewer documents. As a result, general usage words are ranked lower while more specialised ones tend to appear at higher positions. This measure, or rather more complex versions of it, is very frequently employed by search engines to rank documents after a user query.

IDF was originally proposed by Sparck Jones (1972) meaning "a giant leap in the field of information retrieval. Coupled with TF .... it found its way into almost every term weighting scheme" (Robertson, 2004:503). Sparck Jones believed that the fact that a word appeared in many documents was not a good indicator of its representativeness within that set of documents. Contrarily, it appeared that those words which occurred in fewer texts might potentially have greater relevance and be more representative of the documents under analysis.

TF-IDF, that is, the result of multiplying IDF by a word's frequency in a given document (TF), has evolved throughout time into more sophisticated and complicated measures, as discussed by Robertson (2004). In this study, the classical formula by Sparck Jones will be applied. It is "defined as $-\log_2 df_w/D$, where D is the number of documents in the collection and df_w is the document frequency, the number of documents that contain [the word] *w"* (Church and Gale, 1995: 121).

For the sake of comparison with the lists produced by the other four methods, this measure was slightly modified. Instead of resorting to the frequency of a word within a single document in the corpus, which would leave many of the candidate terms in the other lists out of the rank produced by this measure (they might not be found in the document selected), after calculating a word's IDF value using Sparck Jones' classical formula, it will be multiplied by the normalised frequency value[12] of that word in the whole corpus (our adaptation of TF).

[12] This value is obtained by dividing a word's raw frequency by the total number of tokens in the corpus and then multiplying it by a scaling factor to obtain more manageable figures due to corpus size (for instance, in a 2.6 million-word corpus, the scaling factor employed is 1,000).

5. Method implementation and evaluation

5.1. Pre-processing and implementation

The major difficulties encountered in the evaluation of these five methods were, on the one hand, establishing a similar process to assess their precision levels and on the other hand, the intrinsic differences existing amongst them. To begin with, Drouin's *TermoStat* (2003) and *Keywords* (2008) are fully automatic and do not require pre-processing, that is, filtering the lists *a priori* to eliminate as much noise as possible. However, Chung's, TF-IDF, and specially Kit and Liu's methods need it before producing their lists of candidate terms.

As part of the pre-processing phase, Chung resorts to Heatley and Nation's (1996) software *Range* to obtain a frequency word type list based on both her anatomy corpus, the SC, and the LOB and Wellington corpora used as RCs. Then, they discard those word types which do not occur in the SC and also eliminate the texts that may contain any vocabulary related to the anatomy field from the RCs in order "to maximise the statistical contrast between the two corpora" (Chung 2003a: 233).

Kit and Liu's pre-processing procedure consists in tokenising both BLIS and their background corpus, the BNC, and filter noise using stop word lists and eliminating alphanumerical elements. After that, they lemmatise the corpus so as to apply their calculations on lemmata, as shown above.

Concerning UKSCC, the 193 texts in it were pre-processed with *Wordsmith 5* by Scott (2008a) resulting into a list of 27060 word types. Unlike Chung's pre-processing procedure, the legal texts in LACELL were not eliminated. Neither was a frequency threshold established prior to the application of Chung's, TF-IDF, or Kit and Liu's methods so even hapax and dis legomena were considered with the purpose of maximising the exhaustiveness of the results obtained. UKSCC contains 7339 hapax legomena, that is, vocabulary items occurring only once, which represent 27.12% of the total amount of word types. They include proper names, both English and foreign, such as *Mulliken*, *Kolinsky*, *Jewison* or *Kilmuir;* misspelled words like *spirituall, burmouth, juridicial, tatutory, ntitlement* and also initials and acronyms, i.e. *SIAL, ECHR, BAILII or LJ*.

After obtaining the frequency data of the word types in UKSCC with *Wordsmith*, the corpus was filtered using the function word list and baseword list 15 included in Heatley and Nation's *Range* software. They were imported into an excel spreadsheet employing the search function to eliminate the function words and proper names present in UKSCC. The percentage of function words detected amongst UKSCC word types was low, just 0.99% of the total. As for baseword list 15, it is an ever growing inventory of proper nouns provided by Nation which led to the removal of 2519 of these elements shrinking the list by 9.4%. Judging by the numbers, the use of proper nouns appears to be a relatively outstanding feature of this legal genre representing almost 10% of the whole corpus (leaving aside those which do not form part of Nation's list and cannot be detected automatically). Undoubtedly, removing them automatically does increase the level of precision achieved regardless of the method employed. However, these proper nouns had to be carefully supervised before removing them since some of them corresponded with initials or acronyms belonging to the specialised vocabulary of the genre like *LJ* (*Lord Judge), QB* (*Queen's Bench),* or *EC* (*European Court*), amongst others.

The filtered list was also employed for the calculation of TF-IDF which does not employ corpus comparison. The frequency lists of word types provided by *Wordsmith*

5.0 not only give information about a word's frequency in the corpus (which has been used as TF for this experiment) but also about its distribution throughout it, that is to say, how many documents within the collection include a given word. Therefore, these were the parameters employed in this case.
Another pre-processing step taken exclusively for the implementation of Kit and Liu's method was the lemmatisation of UKSCC. It was lemmatised with Schmid's *Tree Tagger* configuring it so that it would not tag as *unknown* those word types it could not assign to any lemma. It resulted into a list of 4563 lemmata once the function words, proper names and words not found in BNC (following their advice in this respect) were carefully filtered. Kilgariff's lemmatised BNC list was used as the RC, as stated above.
On the other hand, due to the fact that neither *TermoStat* nor *Keywords* require any pre-processing steps, both lists were filtered *a posteriori*. As proof of its efficiency, Drouin's *Termostat* only kept 22 function words (0.94%) and 8 proper names (0.34%) as candidate terms (out of 2,333), while the keywords list of 3618 items retained 61 function words (1,68%) and 222 proper nouns (6.13%).
Regarding the actual implementation of the five methods, it must be highlighted that both *TermoStat* and *Keywords* are fully automatic tools which can perform all tasks without any human intervention. As for Chung's, TF-IDF and Kit and Liu's techniques, excel spreadsheets were used to apply the formulas the authors include in the description of their methods. Once the word type list obtained with *Wordsmith* was imported into a spreadsheet and filtered eliminating function words and proper names, the formulas corresponding to each method were applied to the whole list of word types (the necessary parameters for each calculation were obtained using the *search* function provided by excel). Then, each list was sorted in descending order so that those items displaying the highest values would be ranked at the top of the list. For those methods requiring corpus comparison, LACELL was also processed with *Wordsmith* and imported on a different spreadsheet as well as Kilgariff's BNC lemmatised lists.
The parameters necessary to apply those methods which are not fully automatic go as follows:

- Chung: Relative frequency in the SC and RC.
- Kit and Liu: rank position in the SC and RC (in descending order) obtained after sorting the candidates according to their frequency in both corpora.
- TF-IDF: Normed frequency of candidates in the SC and number of documents they appear in in the whole document collection.

With respect to the parts of speech extracted by each method, the methods designed by Chung or Kit and Liu do not discriminate amongst lexical categories for the identification of terms since they do not resort to POS tagging, neither do *Keywords* or TF-IDF. Hence, any part of speech could potentially be regarded as term depending on the different parameters considered to establish its termhood level. Conversely, Drouin's software does employ POS tagging and can be configured to only extract a given part of speech. Nevertheless, it was configured to include both nouns, adjectives, verbs and adverbs in the process. This validation process is thus carried out taking into consideration all lexical categories.

5.2. Defining a gold standard

The results obtained after applying the five ATR methods on UKSCC were validated automatically against a legal glossary used as gold standard. Instead of asking specialists to gather a terminology database extracted from the study corpus, four

different British and American legal English glossaries[13] in raw text format were merged and filtered resulting into a list of 8715 items containing both single and multi-word 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, as shown above), it appears that only 3031 out of 8715 legal terms (34.78%) are MWTs being distributed as illustrated in figure 2: 1999 bi-grams (22.96%), 728 tri-grams (8.35%), 228 MWTs formed by four units (2.61%) and 76 (0.87%) with more than four constituents.

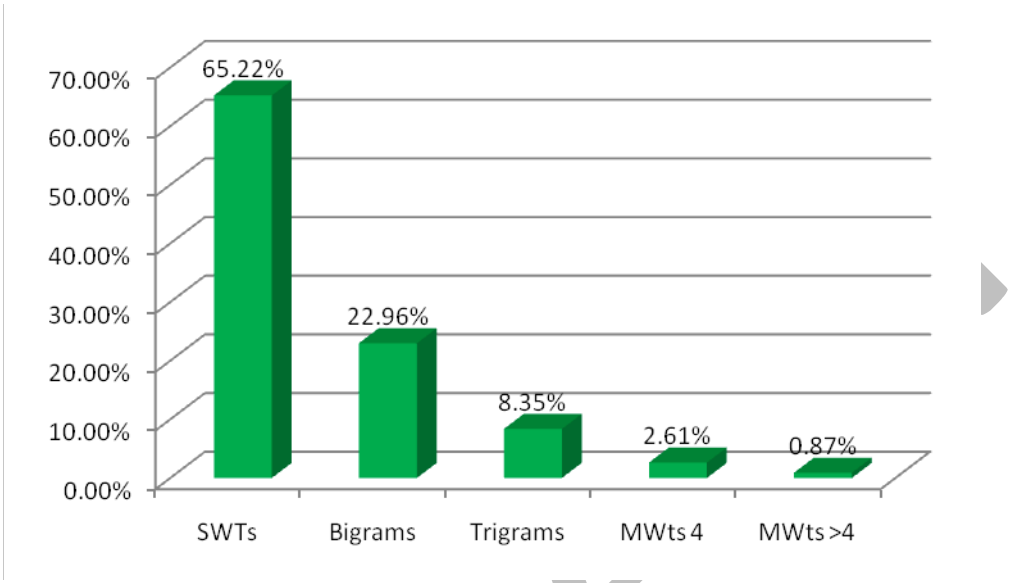

Fig. 2 *Lexical structure of terms in glossary*

Once the true terms in each list were identified by comparison with the gold standard, those candidate terms not qualifying as true terms after applying the methods were analysed manually by the researcher referring to two specialised dictionaries (Alcaraz and Hughes, 2000; Saint Dahl, 1999). This step was taken to contribute to the reduction of silence levels caused by external factors, that is, to guarantee that the glossary, obtained from external sources, would include all the true terms in the corpus. This manual supervision resulted into 10.52% increase of both single and multi-word terms comprised in the glossary list.

5.3. Results

Defining a similar method of comparison amongst the four approaches under evaluation posed certain difficulties due to the different size of the candidate term lists produced by each method. While Chung (2003a, 2003b) and Drouin (2003) establish a threshold to discard non-terms, Kit and Liu (2008), *Keywords* (2008) and TF-IDF provide a much longer inventory of elements which are ranked according to their level of specificity. As a result, since Drouin's list included 2,300 items against 4,654 obtained after applying Chung's ratio, 6,675 keywords, and the 27,060 initial word types appearing in the TF-

[13] 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

IDF and Kit and Liu lists, only the top 2,000 candidate terms in each list were selected so that the comparison could be carried out in similar conditions.

The five methods were assessed in terms of precision and recall. Precision can be measured by establishing the proportion of items that are relevant within a given set. This is why it was calculated progressively, as shown in figure 4 where the five curves plot the precision achieved from candidates 1 to 200, 201 to 400, etc. sorted according to the specifity level established by each method.

As regards recall, which points at the amount of true terms identified with respect to the whole list of terms in the corpus (not in a set), it could be calculated for all methods except for Kit and Liu and TF-IDF since neither of them establish a cut-off point to discriminate terms from non-terms. Figure 3 illustrates both average precision and recall.

Nevertheless, Chung's list posed an additional problem which Kit and Liu address when alluding to the items not in the reference corpus. If an item is not in the RC, Chung automatically classifies it as a term and so do Ahmad et al. (as cit. in ibid.). After examining those elements in BLIS, their study corpus, Kit and Liu (2008: 220) verify that only 20% were true terms and suggest that keeping them "unclassed seems more reasonable when no justifiable solution is available".

Likewise, the number of UKSCC items not in LACELL was also considerably high, 4,367 SWTs were not in the RC and only 280 of them (6.4%) were true terms after comparing them with the gold standard. Thus, it appears that assuming that a word not found in the RC automatically qualifies as a term would not be applicable to our SC either, and following Kit and Liu's advice in this respect might be recommendable. As regards the lists produced by the other methods, they do not include these elements either.

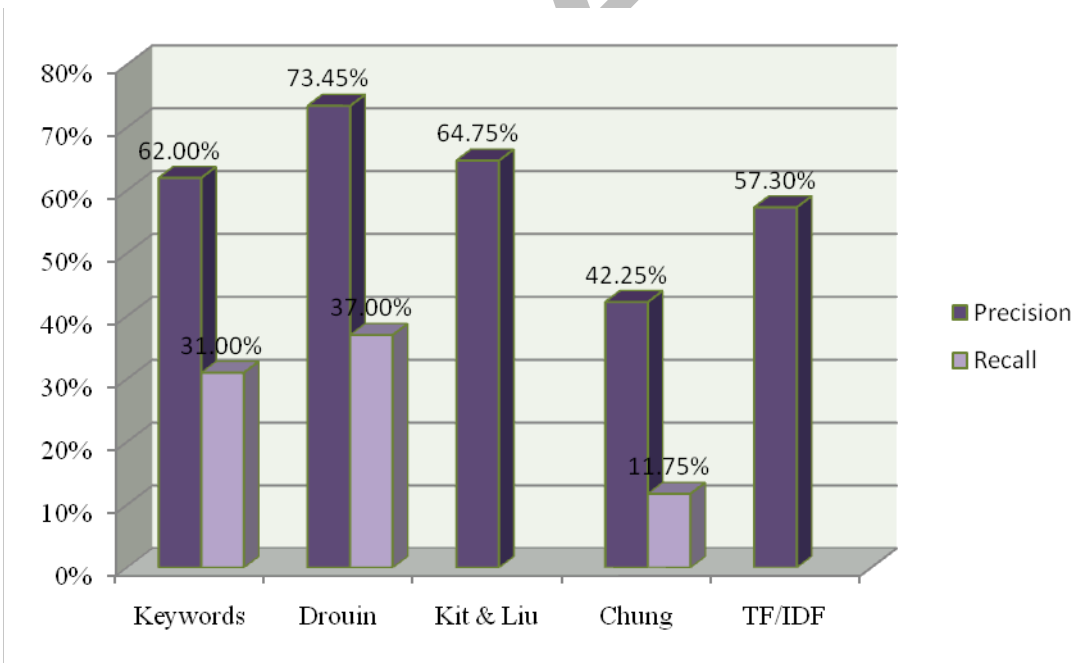

Fig. 3 *Average precision and recall on first 2000 candidate terms*

As shown in figure 3, the overall precision levels attained by the five methods vary with Drouin being the most successful one in identifying terms for this set of 2,000 candidates. It reaches 73.45% being followed by Kit and Liu's which recognises 64.75% of them, TF-IDF manages to extract 57.30%, thus ranking fourth, while

Chung's only identifies 42.25%. As far as *Keywords* is concerned, it ranks third (slightly below Kit and Liu's method) proving to be a considerably effective term extraction tool which detects 62% terms (it reaches 84% precision for the first 200 candidates).

As stated above, calculating recall was not possible for Kit and Liu's method or TF-IDF owing to the fact that the number of candidate terms coincided with the initial list of word types used to implement the four techniques. Kit and Liu believe that there is no such as thing as a cut-off point and establish a termhood continuum where true terms will be pushed to its high end. The TF-IDF measure does not provide such a cut-off point either.

In general terms, recall figures are not high being Drouin's method the one which excels the other two. It reaches 37% recall followed by *Keywords* at 10 points below. Chung's method is the worst performing one achieving only 11.75%.

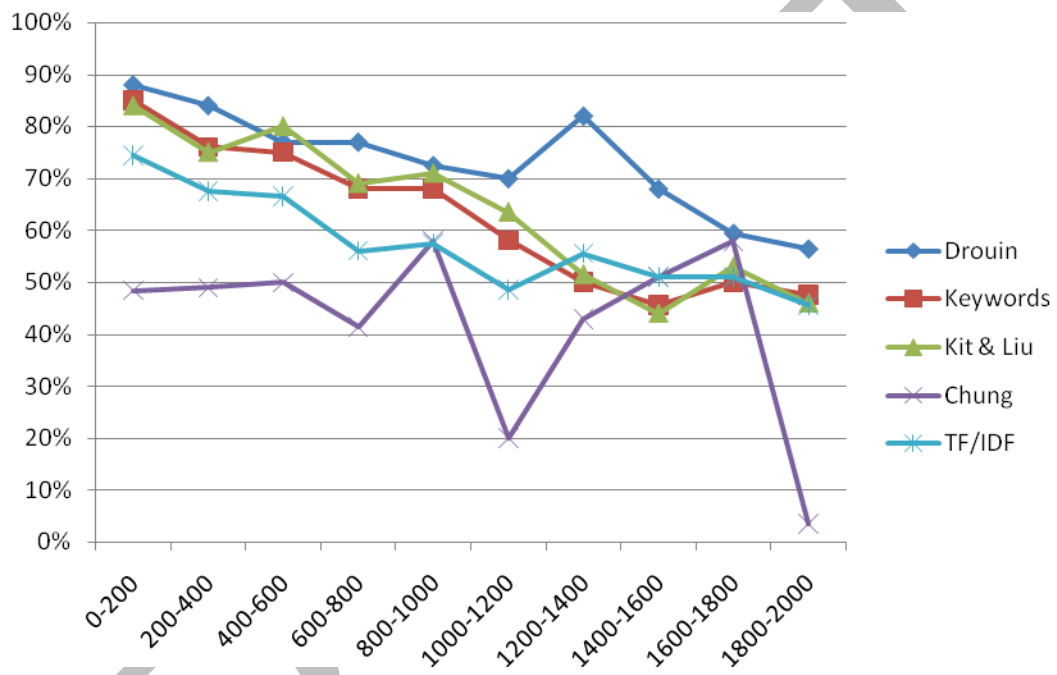

Fig. 4 *Cumulative precision for the first 2,000 candidates*

Figure 4 illustrates cumulative precision across methods where the horizontal axis shows the first 2,000 candidate terms identified in groups of 200 and the vertical one indicates the percentage of precision attained by every method within each group. Drouin's *TermoStat* stands out as the most effective ATR method as it detects 73% terms within the list of 2,000 candidates evaluated. It is closely followed by Kit and Liu's which rises above it only from candidates 400 to 600, where it identifies 80% true terms. The precision levels attained by *Keywords* are reasonably high managing to detect 62% terms (at only 2 points below Kit and Liu). In spite of not resorting to corpus comparison, TF-IDF remains considerably close to *Keywords* and Kit and Liu achieving to detect 57% terms within this range. Nevertheless, Chung's method appears as the least effective one reaching less than 50% precision.

TermoStat, *Keywords,* Kit and Liu and TF-IDF follow a similar trend decreasing their effectiveness smoothly from candidates 1 to 900. Within this range, Drouin achieves

82% precision, Kit and Liu 77%, *Keywords* 76%, and TF-IDF 66% (finding their highest points at 88%, 84%, 84.5%, and 74.5% respectively). On the other hand, Chung remains steady below 50%. From candidates 900 to 1,700 there are greater differences. While *TermoStat* remains ahead reaching a peak of 82% within candidates 1100 to 1300 and then falling down to 60%, Kit and Liu, *Keywords* and TF-IDF continue to descend progressively (more sharply in the case of Kit and Liu) to 53%, 50%, and 54%. Conversely, Chung improves considerably rising to 58%.

Finally, both Kit and Liu and *TermoStat* fall down to 46% and 56.5% while *Keywords* and TF-IDF remain constant at 47.5% and 46% from candidates 1700 to the end of the graph. The case of Chung's method is particularly outstanding as it falls sharply from 58% precision to 3%. It must be emphasised that the 2,000 candidates considered for evaluation do not correspond with what Chung would regard as terms proper. The cut-off pointed suggested by the author would only apply to the first 287. All the same, the average level of precision within this set does not even reach 50%.

On the whole, having compared and assessed the five methods above, there are several generalisations that could be made as regards their effectiveness in extracting terms in a legal English corpus. To begin with, it appears that resorting to corpus comparison yields better results. As a matter of fact, *TermoStat* and *Kit Liu's* methods, the best performing ones, employ this approach to establish a word's termhood level. As regards precision within the list of 2,000 candidates evaluated, both of them stand at 16.15 and 7.5 points respectively above TF-IDF which focuses exclusively on the SC to extract candidate terms.

Another factor that may have influenced their greater rate of success is the fact that, unlike the rest of the methods, both require lemmatisation to be implemented thus indicating that applying calculations on lemmata, not on word types, might be more effective to recognise terms automatically.

Concerning the gold standard employed for evaluation, the fact that it was compiled using external sources does not seem to have affected the results significantly. While Drouin employs a database external to the corpus to assess their method, Kit and Liu resort to a glossary obtained from the texts themselves. However, both methods perform quite efficiently for this experiment being *TermoStat* the most effective one. Even so, there is not enough evidence to relate Kit and Liu's slightly lower rate of success with the fact that the gold standard was not obtained from the legal corpus itself.

Finally, it must be highlighted that the low precision levels achieved by Chung's method might point at its domain dependence. As put forward by Lemay et al. (2005: 233), "lexical units in medical texts bear certain surface-level features (i.e. morphemes or entire words borrowed from Latin and Greek) that, we believe, make them less difficult to identify automatically". Unlike Chung, who resorts to human validation, the use of a gold standard to automatically validate the results in this experiment could have also contributed to the lack of precision of this method.

To conclude, table 1 illustrates the first 25 candidate terms detected by each method ranked in descending order from higher to lower termhood levels according to the different measures proposed by each author.

DROUIN		**KEYWORDS**		**KIT & LIU**		**CHUNG**		**TF-IDF**	
Section	126.29	Court	27965.27	Court	0.3114	Craighead	2198.45	Land	0.998
V (versus)	112.55	Section	24182.76	Judge	0.3110	Appellants	2012.58	Article	0.965
Case	111.79	Para	22007.62	Case	0.3105	CIV	1846.69	Contract	0.926
Para (paragraph)	108.63	Lord	21963.51	Sentence	0.3100	Appellant's	1577.55	Jewish	0.898

Article	97.39	V	19464.25	Contract	0.3095	Paras (paragraphs)	1444.69	Extradition	0.866
Court	88.65	Appeal	18886.16	Appeal	0.3091	Cobbe	1079.23	Possession	0.861
Appeal	80.3	Article	18044.94	Term	0.3086	Estoppel	975.31	Child	0.845
Appellant	78.47	Act	17322.12	Judgment	0.3081	Lessee	639.54	Tenant	0.804
Law	73.55	Case	16541.39	Make	0.3076	PPC	607.57	Company	0.783
Judgment	71.67	Law	10566.68	Issue	0.3072	Respondent's	591.58	Convention	0.775
Claim	69.8	Judgment	8741.90	Order	0.3067	Appellant	582.05	Asylum	0.724
Right	67.98	Convention	7648.50	Offence	0.3062	Realisable	567.59	Data	0.721
Apply	65.5	Rights	7304.34	Appellant	0.3057	Lawfulness	563.60	Directive	0.702
Order	64.39	Whether	7262.35	Costs	0.3053	Tortious	559.60	Equipment	0.701
Decision	63.53	Decision	7056.68	Month	0.3048	Seneschal	535.62	Immigration	0.656
Person	62.83	Appellant	6947.53	Take	0.3043	Para (paragraph)	530.02	Discrimination	0.647
Proceeding	61.7	Proceedings	6927.94	Trial	0.3038	Carnwath	519.63	Suicide	0.645
Relevant	59.02	LJ	6707.16	Say	0.3034	Disapplication	495.65	Rent	0.645
Purpose	58.45	Jurisdiction	5968.92	Evidence	0.3029	Steyn	491.65	Accommodation	0.627
Defendant	57.72	Order	5762.57	Suspended	0.3024	Foreseeability	439.69	Planning	0.614
Provision	57.55	Relevant	5427.42	Defendants	0.3019	Interveners	439.69	Criminal	0.614
Principle	55.77	Ac	5071.04	Fact	0.3015	Abbotsbury	401.71	Commissioners	0.608
Application	55.5	Paras (paragraphs)	5051.25	Conclusion	0.3010	Subsection	384.79	Clause	0.583
Jurisdiction	55.5	Application	4801.27	Give	0.3005	Nuptial	373.73	Property	0.580
Paragraph	54.69	Kingdom	2796.27	Reason	0.3000	Inveresk	371.73	Lease	0.576

Table 1 *First 25 candidate terms ranked by every method*

6. Concluding remarks

This study has presented the results of the evaluation of five single-word term recognition methods implemented on a 2.6 million-word legal corpus. As proved by testing, it appears to be relevant to study such methods since, as far as the gold standard employed for this evaluation is concerned, 65.22% of the items in it are constituted by a single lexical unit.

Establishing a similar procedure to assess these five approaches to term recognition posed certain difficulties and finally, due to the different size of the output lists, just the top 2,000 candidate terms were selected for comparison. As for Chung's list of candidates, it included a large amount of them which were not found to be true terms after comparing them with the gold standard (only 6.4% were identified as such). A great majority of them could not be found in the RC. This is why they were left unclassified following Kit and Liu's suggestion (2008: 220) and only those candidates in the RC were considered for evaluation. Neither were these elements considered for the validation of the other four methods.

As for precision, Drouin's method stands out as the most effective one reaching 88% at its highest point for the first 200 candidate terms and obtaining a mean value of 73.45% for this parameter. Kit and Liu's method remains second at 64.75% finding its peak at 84% within the same range. Their greater rate of success might be related to the fact that both methods resort to corpus comparison, as opposed to TF-IDF (which ranks fourth). They also require the lemmatisation of both the corpora employed for comparison,

hence the importance of applying calculations on lemmata and not on word types for greater efficiency in automatic term recognition.

Curiously enough and in spite of it not being regarded as an ATR method *per se*, *Keywords* achieves 62% precision ranking third and finding its peak at 84.5% for the first 200 candidates evaluated.

Conversely, Chung's method fails to attain the precision percentages established by their own assessments procedure, 86%. The results for UKSCC differ greatly from theirs only achieving 47.5% within the first group of 200 candidates. This might be due to the evaluation process employed and to the fact that, judging by the figures, this method appears to be domain-dependent possibly due to the morphological features of the terms in the original corpus employed by Chung, as pointed out by Lemay et al. (2005).

Finally, recall could only be assessed for three of the five methods under evaluation since two of them, Kit and Liu's and TF-IDF, do not provide a cut-off point to discriminate between terms and non-terms. *TermoStat* reaches 37% recall as opposed to *Keywords* and Chung at 6 and 25 points below, once more pointing at its efficiency as a term extraction tool.

All in all, despite all the automatic steps taken both to implement and evaluate the abovementioned methods, still much remains on the part of specialists to make the last decisions to discriminate terms from non-terms. When words have numerous senses, it is unavoidable to rely on the specialist's criterion to disambiguate them. As Lemay et al. (2005: 245) point out, automatic methods might be of great help for terminologists to confirm their own intuitions and in particular, to "bring to their attention units that might have been considered as trivial and non-domain-specific". Furthermore, they become essential when having to handle large amounts of data which must necessarily be processed and analysed employing automatic means.

References

Ahmad, K., Davies, A., Fulford, H., Rogers, M. 1994. '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* pp. 267-278. Amsterdam: John Benjamins.

Alcaraz, E. 1994. *El inglés jurídico: textos y documentos.* Madrid: Ariel Derecho.

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

Ananiadou, S.

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

----------1994. 'A methodology for automatic term recognition', in *COLING. Proceedings of the 15th International Conference on Computational Linguistics* pp. 1034-1038.

Arson, A., Lang, F. 2010. 'An overview of MetaMap: historical perspective and recent advances. *Journal of American Medical Informatics Association,* 17 (3), pp.229-236.

Barrón-Cedeño, A., Sierra, G. E., Drouin, P. and Ananiadou, S. 2009. '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.

Bernier-Colborne, G. 2012. '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, May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html

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

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

Cabré, M. T., Estopà, R., Vivaldi, J. 2001. 'Automatic term detection: a review of current systems', in D. Bourigault, C. Jacquemin and M.C. LHomme (eds.) *Recent Advances in Computational Terminology*, 53-87. Amsterdam: John Benjamins, Natural Language Processing.

Chung, T. M.

----------2003a. 'A corpus comparison approach for terminology extraction'. *Terminology* 9(2), 221-246.

----------2003b. *Identifying Technical Vocabulary.* Unpublished PhD thesis. Victoria University of Wellington.

Church, K.W., and Hanks, P. 1990. 'Word association norms, mutual information, and lexicography'. *Computational Linguistics* 16(1), 22-29.

Church, K.W., and Gale, W. 1995. 'Inverse Document Frequency (IDF): A measure of Deviations from Poisson'. *Proceedings of the Third Workshop on Very Large Corpora.* Cambridge: Massachusetts Institute of Technology, 121-130.

Dagan , I. and Church, K. 1994 "TERMIGHT: Identifying and Translating Technical Terminology". *4th Conference on Applied Natural Language Processing*. http://www.aclweb.org/anthology-new/A/A94/A94-1006.pdf

Daille, B. 1996. 'Study and implementation of combined techniques for automatic extraction of terminology', in Klavans, J.L., and Resnik, P. (eds.) *The Balancing*

act: Combining symbolic and statistical approaches to language. Cambridge, MA: MIT Press.

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

Drouin, P. 2003. 'Term extraction using non-technical corpora as a point of leverage'. *Terminology* 9(1), 99-117.

Dunning, T. 1993. 'Accurate Methods for the Statistics of Surprise and Coincidence'. *Computational Linguistics* 19(1), 61-74.

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

Flowerdew, L. 2009. 'Applying corpus linguistics to pedagogy: A critical evaluation'. *International Journal of Corpus Linguistics* 14(3), 393-417.

Frantzi, K.T. & Ananiadou, S.

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

---------1999. 'The c/nc value domain independent method formulti-word term extraction'. *Journal of Natural Language Processing* 3(2), 115-127.

Gabrielatos, C. and Marchi, A. 2011. 'Keyness: Matching metrics to definitions', in *Theoretical-methodological challenges in corpus approaches to discourse studies and some ways of addressing them*, 5th November, Portsmouth (unpublished). http://eprints.lancs.ac.uk/51449/4/Gabrielatos_Marchi_Keyness.pdf.

Gilquin, G., Granger, S. 2010. 'How can data-driven learning be used in language teaching?', in OKeefe, A. and McCarthy, M. (eds.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Heatley, A., Nation, P. 1996. *Range* (computer software). Wellington, New Zealand: Victoria University of Wellington. http://www.victoria.ac.nz/lals/about/staff/paul-nation

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

Justeson, J.S. & Katz, S.M. 1995. 'Technical terminology: some linguistic properties and an algorithm for identification in text'. In *Natural Language Engineering* 1 (1), 9-27.

Kit, C and Liu, X. 2008. 'Measuring mono-word termhood by rank difference via corpus comparison'. *Terminology* 14(2), 204-229.

Lemay, C., LHomme, M.C., Drouin, P. 2005. 'Two Methods for Extracting 'Specific' Single-word Terms from Specialised Corpora: Experimentation and Evaluation'. *International Journal of Corpus Linguistics* 10(2), 227-255.

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*, 19th-22nd June, Madrid. http://www.ttc-project.eu/images/stories/TTC_TKE_2012.pdf

Maynard, D. and Ananiadou, S. 2000. 'TRUCKS: A model for automatic multi-word term recognition'. *Journal of Natural Language Processing* 8(1), 101–125.

Marín, M. J., Rea, C. 2011. 'Design and compilation of a legal English corpus based on UK law reports: the process of making decisions', in Carrió Pastor, M.L. and Candel Mora, M.A. (eds.). *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III*

Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politècnica de València. pp. 101-110.

McEnery, T., Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

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

Mondary, T., Nazarenko, A., Zargayouna, H., Berreaux, S. 2012. 'The Quaero Evaluation Initiative on Term Extraction'. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Instambul, May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Nakagawa, H. and Mori, T. 2002. '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.

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

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

Panzienza, M.T., Pennacchiotti, M., Zanzotto, F.M. 2005. 'Terminology extraction: An Analysis of Linguistic and Statistical Approaches'. *Studies in Fuziness and Sooft Computing,* 185, pp. 225-279.

Rea, Camino.

--------- 2008. *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico.* Tesis doctoral. Universidad de Murcia. http://www.tdx.cat/handle/10803/10819.

-------- 2010. **'**Getting on with Corpus Compilation: from Theory to Practice'. *ESP World,* 1 (27), vol. 9. http://www.esp-world.info/articles_27/camino%20rea.pdf

Robertson, S. 2004. 'Understanding Inverse Document Frequency: On theoretical arguments for IDF'. In *Journal of Documentation* 60(5), 503-520.

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

Sánchez, A. P. Cantos, R. Sarmiento and J. Simón 1995. *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Schmid, H. 1995. 'Improvements in Part-of-Speech Tagging with an Application to German', in *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Schmitt, N. 2002. 'Using corpora to teach and assess vocabulary', in Tan, M. (ed.). *Corpus Studies in Language Education*. IELE Press.

Scott, M.

---------- 2008a. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

---------- 2008b. *WordSmith Tools Help*. Liverpool: Lexical Analysis Software.

Sinclair, J. 2005. '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. http://ota.ahds.ac.uk/documents/creating/dlc/index.htm

Sparck Jones, K. 1972. 'A statistical interpretation of term specificity and its application in retrieval'. In *Journal of Documentation* 28 (1), 11-21.

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. 2005. 'Text mining and ontologies in biomedicine: Making sense of raw text'. *Brief Bioinform* 6(3), 239-251.

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

Vivaldi, J. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas.* PhD Thesis. Universidad Politécnica de Cataluña.

Vivaldi, J., Cabrera-Diego, L.A., Sierra, G., Pozzi, M. 2012. '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, May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Wynne, M. 2005. *Developing Linguistic Corpora: a Guide to Good Practice.* AHDS Literature, Languages and Linguistics: University of Oxford. http://ota.ahds.ac.uk/documents/creating/dlc/index.htm.