Measuring the Degree of Specialisation of Sub-Technical Legal Terms through Corpus Comparison: a Domain-Independent Method

Abstract
One of the most remarkable features of the legal English lexicon is the use of sub-technical vocabulary, that is, words frequently shared by the general and specialised fields which either retain a legal meaning in general English or acquire a specialised one in the legal context. As testing has shown, almost 50% of the terms extracted from BLaRC, an 8.85m word legal corpus, were found amongst the most frequent 2,000 word families of West’s (1953) GSL, Coxhead’s (2000) AWL or the BNC (2007), hence the relevance of this type of vocabulary in this English variety. Owing to their peculiar statistical behaviour in both contexts, it is particularly problematic to identify them and measure their termhood based on such parameters as their frequency or distribution in the general and specialised environments. This research proposes a novel termhood measuring method intended to objectively quantify this lexical phenomenon through the application of Williams’ (2001) lexical network model, which incorporates contextual information to compute the level of specialisation of sub-technical terms.

Keywords: Legal English; sub-technical terms; lexical networks; ESP; corpus linguistics
Measuring the Degree of Specialisation of Sub-Technical Legal Terms through Corpus Comparison: a Domain-Independent Method

1. Introduction

David Mellinkoff, one of the first scholars devoted to a systematic description of legal English (also known as legalese), affirms that “the language of the law has a strong tendency to be: wordy; unclear; pompous [and] dull” (Mellinkoff 1963:63). Following his initial steps towards the characterisation of legalese, other authors such as Alcaraz Varó (1994; 2000), Tiersma (1999), Borja (2000) or Orts (2006), also highlight its convoluted character. Regarding its lexicon, the presence of Old French and Latin phrases or the abundance of synonyms, archaisms, redundant elements and shared vocabulary contributes to such obscurity. From this list of features, it is probably the highly frequent use of “common words with uncommon meanings” (Mellinkoff 1963, 63) which differentiates legal English vocabulary from the lexicon of other LSP (Language for Specific Purposes) varieties.

This statement could be supported by the results obtained after applying Heatley and Nation’s (2002) software Range to the comparison of the list of single-word legal terms extracted from the British Law Report Corpus (BLaRC) — an 8.85 million-word legal English corpus of written judicial decisions — with the most frequent 2,000 word families (a lemma and all its possible realisations) of English found in West’s General Service Word List (GSL) (1953), Coxhead’s Academic Word List (AWL) (2000) and the British National Corpus (BNC) (2007). 40.47% overlap was found between our term

1 The software Range was designed by Heatley and Nation for different purposes. It can calculate text range, that is, the percentage of running words in a text covered by a given word list. By default, it uses the lists designed by West (1953), including the most frequent 2,000 words of English; Coxhead’s (2000) addition of 570 most frequent academic word families and the most frequent 3,000 words of the British National Corpus. These lists were used to compare them with our term list as regarded overlap percentage (using an excel spreadsheet). It can also process corpora and produce frequency word lists.

2 See Author and Rea (2012) on the design and structure of BLaRC.
inventory and the first two general vocabulary lists, whereas the level of coincidence was slightly higher, 45.41%, when comparing it with the BNC. These figures thus attest that approximately half of the legal terminology identified in BLaRC$^3$ is shared with the general field since almost 50% of it matched the general vocabulary lists used as reference.

Furthermore, the distinction between highly specialised and general English vocabulary appears to be clear-cut. While the former could be found in general vocabulary inventories such as the ones mentioned above, the latter can be recognised by resorting to a plethora of ATR, capable of identifying the most representative terminology of different varieties of the language automatically.

Nonetheless, there are words standing somewhere in between highly general and specialised vocabulary whose specificity level is hard to define (especially using quantitative criteria) mainly due to their distribution and frequency in both contexts. They are the so-called sub-technical terms, whose presence in BLaRC is noteworthy, as demonstrated above.

This research thus suggests a novel approach designed to measure the degree of specialisation of these terms based on the implementation of Williams’ (2001) lexical network model. In spite of the fact that this research was carried out using a legal and a general English corpus, the algorithm proposed in it could also be implemented on other corpora belonging to any LSP (Language for Specific Purposes) variety, since it is not domain dependent. It is rather based on the difference existing between the average

---

$^3$ The general English corpus used for comparison was LACELL, a 21 million-word general English corpus including both written texts from diverse sources and also oral language samples (only 10%). 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 (BLaRC only includes British texts), reducing the original size to 14.6 million words.
collocate frequency of a given sub-technical term and its collocates in a general corpus and the same datum obtained from a specialised one.

In an attempt to try and delimit as accurately as possible the concept sub-technical, section 2 discusses several authors’ views and criteria for its characterisation, providing a definition of the term sub-technical which stands at the core of the algorithm proposed in section 3. Section 4 also comprises a literature review on ATR methods and the justification, description and results obtained by Sub-Tech, a termhood measuring method designed for the quantitative analysis of sub-technical vocabulary based on its context of usage by comparing the data obtained after applying Williams’ (2001) lexical network model. Finally, section 4 presents the validation process of ten different ATR methods applied on a pilot corpus of 2.6 million words, followed by the conclusion, in section 5.

2. Sub-Technical Vocabulary: a Review of the Concept

As stated in the introduction, the use of vocabulary shared by the general and specialised fields in legal English should be regarded as a relevant feature of legal English. However, there is little agreement amongst specialists on how to define these terms. Generally speaking, authors tend to favour the use of the term "sub-technical" basically defined as vocabulary common to both the general and the specialised fields or amongst scientific disciplines. Only few of them (Cowan 1974; Flowerdew 2001) employ the label "semi-technical" either to refer to the same concept or as a synonym. In addition, most scholars underline the importance of these terms in LSP instruction due to the fact that they might become an obstacle in the learners’ acquisition of the vocabulary in any scientific field.
Within the group of specialists who have addressed the issue from a didactic perspective we find the work by Cowan (1974)—who coined the term "sub-technical"—, Baker (1988) and Flowerdew (2001). They agree on the confusing and obscure character of sub-technical terms, putting special emphasis on the relevance they should be given in the LSP curriculum precisely because of their semantic ambiguity. Flowerdew defined them as "words in general usage [...] which have a special meaning within the technical area" (2001: 82).

Sub-technical terms have also been explored from a quantitative perspective. Farrell (1990) came to the conclusion that distribution plays an essential role in the identification of sub-technical terms. According to him, these terms are well distributed across a corpus and display high frequency counts. Trimble also focused on frequency but, similarly to Flowerdew, added a qualitative viewpoint to their definition noticing that "those words [...] in technical contexts, take on extended meanings" (1985: 129).

Finally, only Chung and Nation (2003) and Wang and Nation (2004) managed to delimit the semantic features of technical and sub-technical vocabulary although they did not label shared vocabulary as "sub-technical".

Therefore, considering all the perspectives involved in the study and analysis of sub-technical terms, a label which is preferred by most authors specifically dealing with this issue, and given the main objective of this research, that of designing a quantitative method which can provide a measure to establish the degree of specialisation of sub-technical vocabulary using statistical data, for a term to considered sub-technical, it should comply with the following requirements:
• It should display high frequency counts both in the specialised and the general corpora used as source, being found within the first third of the frequency word list obtained from each corpus.

• It should also be evenly distributed in both corpora, being present in at least a third of the texts comprised in each collection.

• Having obtained the list of legal terms from BLaRC, the legal corpus, applying TermoStat (Drouin, 2003) as the ATR method to extract them, it should have been assigned a high specificity score in the term rank, placing it within the first third of the term list.

• It could either retain the same meaning in both corpora or rather activate a specialised one when in contact with the legal environment. For example, there are terms like “legislation” which could be found in concordances such as “... all the complications of the legislation on the powers of local authorities to recover payments ...” (extracted from BLaRC) or “… the government’s recent care in the community legislation has forced many people with mental health problems onto the streets ...” (LACELL), referring to a set of rules or laws passed by some official institution or governing body in both contexts. Conversely, there is a second group of sub-technical terms which activate a specialised meaning when in contact with the legal context. The term “conviction” belongs in this group since it can denote “a belief, a strong opinion” in general English, for instance, “… in the conviction, explicit or implicit, that Joyce is an outstanding modern writer …” (LACELL), or rather the legal concept “judgment by a jury or a judge that a person is guilty of a crime”, e.g. “… nothing was to be found elsewhere in the Act which justified a conviction for theft in its absence …” (BLaRC).

---

The first three characteristics mentioned above were checked automatically using software tools such as *Wordsmith* (Scott, 2008) and Drouin’s *TermoStat* (2003), available online. Nevertheless, the last feature required the manual supervision of the concordances associated to each term in each context. In order to carry out this manual supervision, two specialists in legal terminology were consulted. Since this supervision implied reading a high number of concordance lines to ensure the subtechnical character of the terms before actually implementing William’s lexical network method, a selection was carried out first by the author. As already stated by Drouin (2003), this working method implied a high degree of subjectivity, however, whenever there was a lack of consensus between the specialists about a given term, it was automatically removed from the list. Therefore, all the sub-technical terms in the set examined below were identified as such by both specialists.

As a final remark, it should also be stated that none of the scholars devoted to the study of sub-technical vocabulary consider multiword terms within their definition of the concept, this is the reason why the selection of words and terms present in section 3 only includes single-word units.

3. **Sub-tech**: an Algorithm Designed to Measure the Degree of Specialisation of Sub-technical Terms

3.1. **A Literature Review on ATR Methods**

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 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

In spite of the large number of ATR methods existing to date, very few concentrate solely 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 (see Author, 2014 on this issue). Nevertheless, 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 multi-word terms (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 used as gold standard for comparison in the ATR validation procedure applied for this study (described in section 4), which was compiled by merging and filtering four different specialised glossaries of British and American legal English, 65.22% of 10,088 terms in the list were SWTs.
As a matter of fact and as stated in the previous section, most of the scholars devoted to the study of sub-technical vocabulary do not consider multiword units within the category, therefore, the ATR methods discussed in greater depth below are focused on the extraction of SWTs.

3.2. Justification of the Algorithm

The major objective of this research was to try and characterise more accurately such a relevant feature of the legal lexicon as sub-technicality applying corpus linguistics techniques. To that end, we attempted to create a method capable of quantifying the termhood of sub-technical terms. The statistical behaviour of these terms makes them particularly hard to spot, especially when resorting to corpus comparison techniques, as is the case of some of the methods presented and assessed below, given their high frequency and distribution figures in both general and specialised speech. In fact, these methods do not incorporate contextual information, which we deemed necessary to study these terms as additional information to complement other parameters like frequency or distribution.

Thus, would it be possible to quantify the distance or proximity of a sub-technical term to the general or specialised fields based on the context of usage of those terms in both environments? We believed it could and decided to design a termhood measuring method that would process the data obtained from the collocate networks (Williams, 2001) formed by each of these terms in general and specialised English, after comparing ten different ATR methods (Chung, 2003; Church and Gale, 1995; Drouin, 2003; Frantzi and Ananiadou, 1999; Kit and Liu 2008; Nazar and Cabrè, 2012; Park et al., 2002; Sclano y Velardi, 2007; Scott, 2008; Sparck-Jones, 1972). Sub-Tech, the algorithm defined below, attempts to quantify the level of specialisation of a sub-
10 technical term by subtracting the average frequency of its collocate network in a general corpus from the same value in the specific one, likewise incorporating contextual information to its implementation.

Regarding Williams’ (2001) lexical network model, it relies on the concept of collocation as defined by the Birmingham school, based on statistical information for the identification of a word’s collocate. The author defines collocates as “the habitual and statistically significant relationship between word forms within a predefined window and for a defined discourse community, expressed through an electronic corpus of texts” (2001:5), not considering such aspects as grammaticality or lexical transparency in the initial phase to obtain a word’s lexical network.

Williams’ model takes into consideration not only a word’s capacity to generate collocates by itself but also the associations of its collocates and co-collocates (the collocates of those collocates), regardless of the LSP (Language for Specific Purposes) variety, thus increasing the contextual information provided in each case. Consequently, it was used as the means to obtain the necessary data to calculate a word’s sub-technicality level by comparing the networks generated by it both in BLaRC, the specialised corpus, and LACELL, the general one.

The ATR methods cited above were assessed so as to measure their degree of efficiency in legal term extraction and, in turn, to obtain a reliable list of terms to use as reference. Drouin’s (2003) TermoStat excelled the other nine by achieving 88% precision for the top 200 candidate terms (CTs) extracted and 73% on average. Six of these methods (Chung, 2003; Church and Gale, 1995; Drouin, 2003; Kit and Liu 2008; Scott, 2008 and Sparck-Jones, 1972) focused on SWTs, which is why they were studied in greater detail.

---

5 See section 4 for further details on the validation process.
From this group of six methods, only two of them, RIDF (Church and Gale, 1995) and TF-IDF (Sparck-Jones, 1972) do not resort to corpus comparison but rather to one single corpus and the distribution of its lexicon throughout the documents comprised in it. Actually, Sparck-Jones (1972) 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 seemed that those words which occurred in fewer texts might potentially have greater relevance and were more representative of the documents under analysis. TF-IDF (Term Frequency-Inverse Document Frequency) could measure a word’s weight by taking into consideration its frequency in a given document and the number of documents it appeared in throughout a corpus. A word would display greater weight if it showed high frequency values and appeared in fewer documents. As a result, general usage words were ranked lower while more specialised ones tended to appear at higher positions. Church and Gale (1995) reformulated Spark-Jones’ measure proposing a new method, RIDF (Residual Inverse Document Frequency), with the aim of accounting for certain deviations from a chance-based model. It could be calculated by subtracting the predicted IDF from the observed measure. Even so, neither TF-IDF nor RIDF reached high precision percentages, only managing to identify 57.35% and 48% true legal terms on average respectively (see table 2).

The rest of SWT recognition methods assessed for this study require the comparison between a general corpus and a specialised one to identify and rank the terms in the latter according to their termhood. The parameters employed by each of them are, basically, frequency and distribution in both corpora, although the methods vary as regards the sophistication of their algorithms. Whereas Chung (2003) calculates the frequency ratio of a given type by dividing its absolute frequency in the specialised corpus by the same value in the general corpus, and Kit and Liu (2008) obtain the level...
of specialisation of a term by subtracting its rank position the general corpus\(^6\) from the same value in the specialised one, the algorithms proposed by Drouin and Scott are far more complex.

The tool *Keywords*, included in the software package *Wordsmith* (Scott, 2008), extracts terms by applying different measures. In this case, it was configured to implement Dunning’s (1993) log-likelihood algorithm which takes into consideration the size and relevance of a word’s frequency of use in two texts or sets of texts and then measures the discrepancy existing between this value and the expected frequency of the same word. As for Drouin (2003), he proposes a corpus comparison approach which provides information on a candidate term’s standard normal distribution in both contexts giving “access to two criteria to quantify the specificity of the items in the set: (1) the test-value, which is a standardized view of the frequency of the lexical units, and (2) the probability of observing an item with a frequency equal to or higher than the one observed in the AC [the general corpus]” (2003: 101).

Another key factor is lemmatisation\(^7\), which both Kit and Liu (2008) and Drouin (2003) resort to, whereas Chung (2003) and Scott (2008) do not. Given the average precision percentages attained by the first two methods, 73% and 64% respectively, as opposed to Chung’s method, which only reaches 42.25% precision, it could be stated that computing lemma frequency might also be a very helpful procedure in term identification, although Scott’s (2008) *Keywords* does not require such process and, even so, obtains 62% average frequency, at only 2 points below Kit and Liu’s method. Only Drouin’s *TermoStat* POS (part of speech) tags\(^8\) the corpus.

---

\(^6\) The terms are ranked according to their frequency in both corpora, requiring normalisation.

\(^7\) It entails identifying the rootwords or lemmas of every type in the corpus and compute lemma frequency, not type frequency, which would compute, for instance, the word *defendant* and its plural as two different types, in spite of deriving from the same rootword or lemma.

\(^8\) It consists in indicating the grammatical category of the lemmas.
In sum, comparing general and specialised corpora was attested to be an essential procedure in the identification of legal terms as well as focusing on such parameters as their frequency and distribution in both contexts, however, none of the methods above take into consideration the capacity of a given term to attract other words/terms to its vicinity depending on its relevance in each context. Precisely here lies the novelty of *Sub-Tech*, the algorithm described below, which applies Williams’ (2001) lexical network model as one of the steps to be taken in its implementation, owing to the capacity of these networks to provide plenty of contextual information about a term and its collocates and co-collocates, as stated above.

### 3.3. Description of the Algorithm

After obtaining the lexical networks of twenty technical, sub-technical and general words and arranging the data on an excel spreadsheet for their later processing, a direct observation of the information provided was carried out leading us to the formulation of *Sub-Tech*. The data available for each of the terms included in the sample showed that sub-technical terms displayed a clear tendency to generate not only a larger amount of collocates and co-collocates in the specialised corpus —294.80 items integrated the specialised networks of these terms on average as opposed to 39.82 elements in the general networks—, but also higher frequency counts as collocates. Based on these observations, our hypothesis was formulated, that is, the comparison of these data could provide objective information on the greater or lesser proximity of sub-technical terms to the general or specialised fields based on the density and consistency of the lexical networks generated in each case.
In order to test our hypothesis, general words as well as highly specialised terms were also sampled to try to delimit the results after implementing the algorithm Sub-Tech. The collocate span established was 5 to the left and the right of the network node (the word under analysis) and its collocates subsequently. Nevertheless, a >30 collocate frequency threshold was applied to prevent the networks from becoming unmanageable.

Even so, the average number of elements in each network was 2,609 constituents for BLaRC (294.80 after normalisation) and 596 for LACELL (39.83 after normalisation). The networks expanded at two levels, that is, they comprised the main node’s collocates and the collocates of those collocates (the co-collocates) so as to limit their size in a way that the information could be properly handled.

Both BLaRC and LACELL were processed using Keywords (Scott 2008) applying mutual information (Church and Hanks, 1990) for the identification of the terms’ collocates. Function words were filtered out. Following Williams’ procedure, ungrammaticality and lexical transparency were not considered for the initial selection of the collocates in each network. On the contrary, once function words had been purged automatically, all the patterns identified by mutual information were included as part of the collocate inventory. Then, the resulting lists of words forming each network were transferred to a spreadsheet and the data obtained was then processed applying step 4 of Sub-Tech.

Table 1 illustrates the average number of elements in the specialised and general networks of all the words selected as well as the average number of occurrences of all

---

9 For a term to be regarded as highly specialised, according to the same criteria employed in the definition of sub-technical terms, it should have high frequency counts in the specialised context (being amongst the top 50 terms in the term list) and not occurring in the general corpus; be poorly distributed in the specialised corpus (being restricted to a small number of texts given its highly specific meaning) and also display a high specificity score assigned by TermoStat (as a result of such data).

10 The figures were normalised for the sake of comparison owing to the different size of the corpora used in the study. The frequency data provided by Wordsmith was divided by the number of millions of words in each corpus, 8.85 for BLaRC and 14.6 for LACELL.
the collocations comprised in each network. The figures were normalised for the values to be comparable because of the different size of both corpora.

**Table 1. Number and frequency of the collocates and co-collocates in the lexical networks analysed**

<table>
<thead>
<tr>
<th>Word</th>
<th>Specialised collocates/ co-collocates (normalised)</th>
<th>General collocates/ co-collocates (normalised)</th>
<th>Normalised frequency (BLaRC)</th>
<th>Normalised frequency (LACELL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTY</td>
<td>708.36</td>
<td>274.13</td>
<td>9.22</td>
<td>4.73</td>
</tr>
<tr>
<td>TRIAL</td>
<td>666.66</td>
<td>2.33</td>
<td>9.22</td>
<td>3.84</td>
</tr>
<tr>
<td>OFFENCE</td>
<td>522.93</td>
<td>28</td>
<td>8.91</td>
<td>5.03</td>
</tr>
<tr>
<td>SENTENCE</td>
<td>491.25</td>
<td>1.53</td>
<td>9.5</td>
<td>2.98</td>
</tr>
<tr>
<td>LIABILITY</td>
<td>421.69</td>
<td>0</td>
<td>8.2</td>
<td>0</td>
</tr>
<tr>
<td>PURSUANT</td>
<td>404.4</td>
<td>0</td>
<td>10.34</td>
<td>0</td>
</tr>
<tr>
<td>DISMISS</td>
<td>338.64</td>
<td>3.2</td>
<td>10.06</td>
<td>3.81</td>
</tr>
<tr>
<td>CONVICTION</td>
<td>281.35</td>
<td>1.33</td>
<td>10.41</td>
<td>3.23</td>
</tr>
<tr>
<td>LEGISLATION</td>
<td>246.44</td>
<td>39.7</td>
<td>9.23</td>
<td>4.2</td>
</tr>
<tr>
<td>RELIEF</td>
<td>184.18</td>
<td>6.08</td>
<td>9.88</td>
<td>4.45</td>
</tr>
<tr>
<td>COMPLAINT</td>
<td>180.22</td>
<td>18.18</td>
<td>8.79</td>
<td>4.7</td>
</tr>
<tr>
<td>CHARGE</td>
<td>167.68</td>
<td>64.77</td>
<td>9.08</td>
<td>4.89</td>
</tr>
<tr>
<td>SOLICITOR</td>
<td>159.77</td>
<td>0.33</td>
<td>8.23</td>
<td>2.39</td>
</tr>
<tr>
<td>ESTOPPEL</td>
<td>114.57</td>
<td>0</td>
<td>8.65</td>
<td>0</td>
</tr>
<tr>
<td>GUILTY</td>
<td>66.55</td>
<td>11.96</td>
<td>6.87</td>
<td>4.25</td>
</tr>
<tr>
<td>WARRANT</td>
<td>30.39</td>
<td>1.6</td>
<td>7.91</td>
<td>3.01</td>
</tr>
<tr>
<td>BATTERY</td>
<td>27.57</td>
<td>0.73</td>
<td>7.89</td>
<td>2.27</td>
</tr>
<tr>
<td>EAT</td>
<td>0</td>
<td>2.2</td>
<td>0</td>
<td>3.27</td>
</tr>
<tr>
<td>BLUE</td>
<td>0</td>
<td>13.43</td>
<td>0</td>
<td>3.52</td>
</tr>
<tr>
<td>MORNING</td>
<td>0</td>
<td>268.36</td>
<td>0</td>
<td>4.94</td>
</tr>
</tbody>
</table>

As it can be observed, the words *pursuant, estoppel* and *liability* could be deemed highly technical since they do not generate any collocates in the general corpus applying the >30 frequency threshold established in this study (apart from the high specialisation score assigned to them by ATR methods). Conversely, *eat, blue* and *morning*, general words employed in everyday language, attract no collocates in the specialised corpus under the same conditions. The rest of them are sub-technical, either because they are shared by the general and specialised areas of language without changing their meaning (*guilty; legislation; solicitor*) or because they acquire a technical meaning in the legal
context (trial; sentence; conviction). All of them have a greater number of collocates and co-collocates in the legal corpus, as stated above. In addition, the average frequency of these collocations is higher for all of them in the specialised corpus (5.1 points higher in BLaRC on average) except for eat, blue and morning, which behave conversely due to their highly general character.

The application of the algorithm proposed below allowed us to calculate a score for the ranking of the set of terms and words selected depending on the value assigned to each of them, as already stated. This value was not delimited either with a maximum or a minimum, therefore, the scores obtained could not be studied in isolation but rather as part of a continuum of specialisation whereby words would tend towards one or the other end depending on the density and consistency of the lexical networks they could generate in each field. Those words displaying a higher score would be considered more specialised than those associated to a lower or even negative value. Figure 1 displays the results.

An algorithm was therefore proposed including all the steps followed towards the ranking of this type of vocabulary. The algorithm, owing to its major objective, was called Sub-Tech. It goes as follows:

**Step 1:** Identification and extraction of the specialised terms in BLaRC applying Drouin’s (2003) TermoStat ATR method.

**Step 2:** Selection of the set of words which would be used for the implementation of Williams’ lexical network model. The selection would be carried out applying the criteria established in the definition of sub-technical term provided in section 2.

**Step 3:** Application of Williams’ (2001) lexical network model to the list of words selected both in the specialised and general corpora with the aim of comparing results.
Step 4: Implementation of the formula below in order to place sub-technical terms along a continuum of specialisation:

\[ ST(w_i) = \frac{\mu^G}{|C^G|} - \frac{\mu^T}{|C^T|} \]

Where \( ST(w_i) \) represents the sub-technicality score of a given word, \( w_i \), calculated by subtracting the average frequency of the collocates and co-collocates in the general usage network \( \mu^G \) from the same parameter in the specialised one \( \mu^T \). Both values had to be normalised by dividing them by the total number of tokens in each corpus, that is, \( |C^T| \) and \( |C^G| \) represent the number of elements in the specialised and general corpora respectively. This normalisation was necessary so as to obtain a coherent value due to the different size of both corpora. The average frequency values in each corpus could solely be compared if they were normalised, since LACELL (14.6m words) was almost twice as big as BLaRC (8.85m). In order for the figures obtained to be manageable, \( |C^T| \) and \( |C^G| \) were expressed in millions of words.

The implementation of this formula produced a score which allowed us to rank the terms in the sample and place them along a continuum of specialisation, which is exemplified in figure 1.
3.4. Results and Discussion

Figure 1. Sub-technicality score

Figure 1 illustrates how pursuant, estoppel and liability, the most specialised words in the group (their specificity scores are the highest in the group and they cannot be found in the general corpus, not generating any collocates in it) present the highest values (10.34, 9, and 8.2 respectively), ranking at the very top of the continuum of
specialisation, whereas *eat, blue* and *morning* (whose sub-technicality scores are -3.22 -3.52 and -4.94 respectively), appear at the bottom owing to their highly general character. They were employed to validate the method showing that, while the rest of the words are distributed between these two extremes, they would gather at opposite ends of the continuum as a result of their lack of collocates in either the specialised or the general networks once the algorithm *Sub-Tech* was fully implemented.

The highest ranking set amongst the sub-technical group are *conviction* (7.18 score), *sentence* (6.61) and *dismiss* (6.25). What they all have in common, given the importance of this parameter in the calculation of their sub-technicality score, is the reduced number of elements in the general networks they generate, ranging from 1.33 collocates and co-collocates to 3.2, as shown in table 1. This datum clearly signals their lesser weight in the general context because of their smaller capacity to attract other words to them and their collocates as well as their lower frequency and distribution values in the general corpus. Conversely, the number of constituents of their specialised networks is much greater varying from 281.35 collocates (*conviction*) to 491.25 (*sentence*). This fact coupled with the higher average frequency of those collocates in the specialised corpus (at almost 8 points above the same datum in the general corpus), as well as their own higher frequency and distribution in BLaRC, points at their greater relevance in the specialised field.

On the other hand, such terms as *charge, complaint or offence*, which stand at the bottom of the sub-technical ranking (with 4.09, 3.88 and 2.62 score respectively), behave differently. Although they generate a similar amount of collocates in the specialised field (290 on average) to the top-ranking terms above, their relevance and capacity to attract other words in the general context is much greater. The number of constituents in the general networks associated to these terms ranges from 18.18
To 64.77 (charge), 18 times as much as the most specialised ones within the set. The difference between the mean frequency of the constituents in their general and specialised networks is also smaller, around 4 points on average, half as much as the top-ranking terms above, hence their smaller distance from the general words at the lower end of the continuum.

Nevertheless, for the method to work properly, the statistical data of the words selected had to be significant. If it was not, in spite of its robustness, mutual information, as well as other statistical measures, would produce misleading results. Therefore, if step 4 of the algorithm was applied to a word like lessor, which occurs only 50 times in BLaRC and none in LACELL, it would not be located within the highly technical term range (where it belongs) but rather within the general one. Even so, this method has not been designed to measure highly technical words (associated with very low frequency and distribution counts in the general field), but rather those which tend to be at one end of the continuum or the opposite one due to the number and frequency of the constituents of its lexical network in both corpora.

In relation to the meaning of the sampled sub-technical terms, applying Chung and Nation’s criteria for the classification of specialised vocabulary, it appears that, on the whole, those sub-technical terms whose legal meaning is “minimally related” (Chung and Nation, 2003: 105) to their most frequent general sense such as battery (understood as “a power source” in the general field and “the act of beating up somebody” in the legal context), conviction (“a strong belief” against “the act of declaring someone guilty of a given crime”) or warrant (“a product guarantee” v. “a written order from a judge”), tend to be more distant from the general set of words (0.6 points difference on average) than those terms whose specialised meaning is more related to their general meaning, for instance, offence (“an offensive act” in the general context” and “a crime” in the
legal one), or sentence (“a grammatical unit” and also “a penalty dictated by a law court”).

Nonetheless, this difference is too small to lead to a generalisation on the possible correlation between this measuring method and the meaning of sub-technical terms. A wider sample of words would be needed in order to reach more definite conclusions in this respect as well as a qualitative study of the characteristics of the collocates in the lexical networks, yet the size of these networks might pose difficulties for this method of analysis. This is why a partial automatisation of the algorithm (for example, the initial steps taken to obtain a term’s lexical network comprising more than one level) might be desirable.

Word sense disambiguation (WSD) methods might certainly add to a further development of this measuring method in the future and to deepen into the meanings of sub-technical words, which are so typical of this LSP variety. The work of authors like Geffet and Dagan (2005), who study distributional similarity and lexical entailment, could contribute to a more detailed study of the collocate networks generated for this research. So would the work by Weeds et al. (2004), which proposes a measure to determine the degree of compositionality of collocations, that is, how lexically transparent they may result depending on the possibility of inferring the meaning from their constituents. Furthermore, the research by Joslyn et al. (2008) may also be of great use in the construction of specific corpus-based semantic taxonomies which could help to organise the vast amount of data comprised in the lexical networks of the sub-technical terms included in this study.

To conclude, in spite of the above-mentioned limitations, one of the major strengths of this ranking method, together with the fact that it allows us to objectively measure the degree of specialisation of sub-technical terms, is that it is not domain-dependent.
However, the statistical data associated with each word must be significant enough for such measures as mutual information to work properly, therefore, low frequency words could not be studied applying this technique since mutual information does not extract any collocates when this parameter is too low. On the other hand, Williams’ method is capable of producing such a vast amount of data that the networks usually become unmanageable thus requiring the researcher to establish a frequency threshold and to limit the network levels analysed for practical reasons.

4. ATR Method Validation

As a final remark related to the design and implementation of Sub-Tech, it would be noteworthy to describe the validation process of ten different ATR methods, which were assessed as regards precision with the aim of finding the most efficient one in legal term identification. They were tested on a 2.6 million-word pilot corpus extracted from BLaRC to facilitate the process. After doing so, they were applied on BLaRC. A legal glossary of 10,088 entries (including both single and multi-word term units) was employed as gold standard to calculate precision, the glossary resulted from merging and carefully filtering four different online legal glossaries. The glossaries contained legal terms which belonged to different legal areas and jurisdictions and were downloaded/copied and saved in raw text format and then imported to an excel spreadsheet to be compared with the lists of candidate terms extracted from both BLaRC.

11 Available online at:
http://www.copfs.gov.uk/glossary-of-legal-terms
http://sixthformlaw.info/03_dictionary/index.htm
http://www.nolo.com/dictionary
Once the candidate term (CT) output lists were generated\(^\text{12}\), they were compared with the gold standard using a spreadsheet to calculate the percentage of CTs found in it and thus confirmed as true terms (TTs), in other words, to establish the levels of precision achieved by each ATR method. The total number of CTs confirmed as TTs was obtained by applying an excel formula which allows for the automatic search of elements in different sheets within the same excel book, therefore, if a CT (listed on one of the spreadsheets) was found among the list of terms in the glossary (listed on a different spreadsheet), a value was assigned to the former, if not, the symbol “N/A” was provided by the system, which meant that the CT could not be found in the glossary and could not be confirmed as a TT. Hence, the list of TTs was arranged in order and noise (CTs wrongly identified as TTs) was eliminated. The last step consisted in calculating the percentage of TTs extracted by each method with respect to the whole list of elements included in each CT list.

Precision was measured both cumulatively (in groups of 200 CTs to observe how it progressed as the number of these items augmented) as well as on average.

Table 2 illustrates the average precision percentages attained by the ten methods evaluated as well as the ones achieved by each of them for the top 200 CTs. With the aim of selecting the most efficient methods, they were grouped according to the type of terminological units they could extract. Drouin’s (2003) TermoStat was validated twice due to its configuration possibilities allowing for SWT (single-word term) and MWT (multi-word term) recognition separately. The comparison amongst methods was therefore carried out separately so as to produce a list of both single and multi-word legal terms. If all the methods had been compared as a single group, those capable of

\(^{12}\) In order for the evaluation process to be homogeneous and owing to the different size of the CT lists obtained, precision was measured for the top 2,000 and 1,400 CTs in single and multi-word term identification respectively.
identifying only single-word units would have excelled the other set, as shown in table 2, thus affecting the selection of the most efficient techniques to be implemented on our corpus.

Recall was not taken into consideration in any case owing to the fact that a few of these ATR methods do not establish a limit (related to the specificity value assigned to each CT) to discriminate terms from non-terms, only a few of them do. As a matter of fact, absolute recall can be calculated by dividing the number of true terms identified by each ATR method by the total number of terms in the whole corpus, however, as Frantzi et al. (2000: 122) already acknowledge when explaining the validation procedure of the C-Value/NC Value method, “if we wanted to calculate the absolute value for recall, a domain expert would have had to find all the multi-word terms from the corpus (or a sufficiently large sample of it)”, which would be unattainable due to the size of the corpus. Using the frequency list as reference would not be an option either. Actually, the glossary would only work as the gold standard against which the list of terms would be compared to be validated but not as the reference list to calculate recall, since it would comprise terms not extracted from the pilot corpus.

Consequently, such comparison was not possible for those methods not providing a cut-off point as reference. As we did not have a complete list of all the terms in the corpus identified by specialists (using the lists produced by one of these methods could lead to skewedness in the results), recall was not computed in those cases. As for those methods which did establish a limit for the distinction between terms and non-terms, TermStat also excelled the other two, reaching 37% recall as opposed to Keywords (Scott, 2008), which achieved 31% and Chung’s (2003) method, reaching only 11.75%.
Table 2. ATR method validation results

<table>
<thead>
<tr>
<th>SWT RECOGNITION METHODS</th>
<th>ATR Methods</th>
<th>Average Precision</th>
<th>Precision for top 200 CTs</th>
<th>Number of TTs extracted from pilot corpus (out of 2,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TermoStat (Drouin 2003)</td>
<td>73%</td>
<td>88%</td>
<td>1,463</td>
</tr>
<tr>
<td></td>
<td>Kit and Liu’s rank difference (2008)</td>
<td>64%</td>
<td>84%</td>
<td>1,282</td>
</tr>
<tr>
<td></td>
<td>Keywords (Scott 2008)</td>
<td>62%</td>
<td>85%</td>
<td>1,241</td>
</tr>
<tr>
<td></td>
<td>TF-IDF (Sparck-Jones 1972)</td>
<td>57.35%</td>
<td>74.50%</td>
<td>1,147</td>
</tr>
<tr>
<td></td>
<td>RIDF (Church and Gale 1995)</td>
<td>48%</td>
<td>28%</td>
<td>960</td>
</tr>
<tr>
<td></td>
<td>Chung’s ratio value (2003)</td>
<td>42.25%</td>
<td>48.50%</td>
<td>845</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SWT AND MWT RECOGNITION METHODS</th>
<th>ATR Methods</th>
<th>Average Precision</th>
<th>Precision for top 200 CTs</th>
<th>Number of TTs extracted from pilot corpus (out of 1,400)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Terminus 2.0 (Nazar and Cabré 2012)</td>
<td>71.50%</td>
<td>84.50%</td>
<td>1,001</td>
</tr>
<tr>
<td></td>
<td>Textract (Park et al. 2002)</td>
<td>52.71%</td>
<td>39%</td>
<td>737</td>
</tr>
<tr>
<td></td>
<td>C-value (Frantzi and Ananiadou 1999)</td>
<td>52.43%</td>
<td>69%</td>
<td>734</td>
</tr>
<tr>
<td></td>
<td>Termextactor (Selano and Velardi 2007)</td>
<td>48.29%</td>
<td>64.50%</td>
<td>676</td>
</tr>
<tr>
<td></td>
<td>TermoStat (Drouin 2003)</td>
<td>35.86%</td>
<td>44%</td>
<td>502</td>
</tr>
</tbody>
</table>

The figure shown in column 4 of table 2 corresponds with the final number of true terms (TTs) confirmed after validating the list of candidate terms extracted by the single and multi-word term recognition methods in the table out of 2,000 and 1,400 CTs respectively. Establishing a cut-off point was necessary for the evaluation of each of these methods because the size of the output lists generated by them ranged from 27,060 units measured by TF-IDF (there is no threshold to discriminate terms from non-terms according to this method and thus all the types in the frequency list are included in the output list), to 2,300 candidate terms extracted by TermoStat (Drouin, 2003), which does apply a specificity limit of +3.09 to differentiate terms from non-terms.
As illustrated above, the results differ greatly showing that Drouin’s (2003) TermoStat and Nazar and Cabré’s (2012) Terminus were the most efficient methods in identifying the terms in our legal corpus. They managed to recognise 73% and 71.50% single and multi-word terms respectively on average reaching peaks of precision of 88% and 84.5% for the top 200 CTs. They were therefore singled out to be implemented on BLaRC resulting into a list of 2,848 specialised legal terms. In order to reduce noise and silence levels to the minimum, the list was also supervised manually following the procedure described in section 2.

6. Conclusion

This article has presented a novel approach to try and quantify the degree of specialisation of sub-technical terms in legal English which, according to scholars and the data provided, stand out as a significant trait of the legal lexicon. In general, sub-technical terms are shared by the general and specialised fields, being evenly distributed and presenting high frequency counts in both contexts, apart from often displaying new specialised meanings when in contact with the technical environment.

Given its capacity to provide information about a word’s context by taking into consideration not only the word’s collocates but also its co-collocates, Williams’ (2001) lexical network model was used as a way to access such information about the set of sub-technical terms sampled for this study. We departed from the hypothesis that the more specialised a sub-technical term was, the greater its capacity to attract other words in a specialised context, generating more populated networks around it and its collocates, whose mean frequency was much higher. On the contrary, the more general a sub-technical term tended to be, the smaller its influence on the specialised field and therefore the lesser its capacity to associate to other words within it.
Based on this hypothesis, the algorithm Sub-Tech was formulated so as to establish a comparison between the data obtained from both corpora applying Williams’ method. Once the lexical networks of each of the words analysed were obtained from BLaRC and LACELL, a sub-technicality score was calculated by subtracting the average frequency of the collocate network in the general corpus from the same datum in the specialised one after normalisation. This value allowed us to place sub-technical terms along a continuum of specialisation where they would tend to one or the other end depending on the density and consistency of the networks obtained in each case. However, more accurateness in the results could be achieved if the networks were expanded to lower levels, adding more contextual information to be processed with Sub-Tech, yet such expansion could also cause the networks to become unmanageable because of the large number of elements comprised in them.

To sum up, one of the major novelties introduced by the termhood measuring method proposed herein is the fact that contextual information is added to the equation through the implementation of Williams’ (2001) lexical network model. Moreover, Sub-Tech is a domain-independent method that could be applied to any LSP corpus. In fact, it allows us to objectively characterise such ambiguous lexical elements as sub-technical terms, something which, judging by the literature on ATR methods and to the best of our knowledge, has not been accomplished to date.

Nonetheless, there is still a long way to go as regards the automatisation of the process, which is still time-consuming when it comes to the formation of the lexical networks and their processing. It would also be interesting to explore the contribution of WSD approaches as a further extension of this proposal, which could indeed add up to a better understanding of such a relevant trait of the legal lexicon.
References


Very Large Corpora, ed. by Yarowsky, D., and K. Church, 121-130. Cambridge: Massachusetts Institute of Technology.


