



ELSEVIER

System 32 (2004) 251–263

SYSTEM

www.elsevier.com/locate/system

Identifying technical vocabulary

Teresa Mihwa Chung, Paul Nation*

*School of Linguistics and Applied Language Studies, Victoria University, PO Box 600,
Wellington, New Zealand*

Received 21 May 2003; received in revised form 12 November 2003; accepted 12 November 2003

Abstract

This study compared four different approaches to identifying technical words in an anatomy text. The first approach used a four step rating scale, and was used as the comparison for evaluating the other three approaches. It had a high degree of reliability. The least successful approach was that using clues provided by the writer such as labels in diagrams, typographical marking, and definitions. Using a technical dictionary was more successful, but had an accuracy rate around 80%. The fourth approach compared frequency of occurrence in the specialized text with frequency in a large more general corpus. This worked well, but failed to identify words like *neck*, *chest*, *skin* which were also in common usage. It also could not separate collocates of technical words (*superior*, *posterior*, *transverse*) from technical words. If collocates are included the accuracy rate is close to 90%. Being able to reliably identify technical vocabulary provides an essential starting point for looking at how learners and teachers should deal with technical vocabulary.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Terminology; Technical; Vocabulary; Special purposes; ESP; Computer; Term; Extraction; Dictionary

1. Introduction

Technical vocabulary is a major concern for learners who have special purposes in language learning. However, surprisingly little is known about such vocabulary, largely because there are no well established approaches for deciding which words are technical terms and which are not, and there are virtually no studies that compare the effectiveness of approaches. The major difficulty is that technicalness is a functional aspect of a word and thus the particular use of a word must be taken into account when deciding whether it is a technical term or not. The aim of this study is to

* Corresponding author.

E-mail address: Paul.Nation@vuw.ac.nz (P. Nation).

compare various approaches to identifying technical vocabulary. If we can reliably identify what words are technical in a specific discipline, we can then begin to look at the role these words play in technical text and the demands they place on a learner of English for special purposes. We can calculate how large a technical vocabulary might be, how often and with what density these words occur in text, and how teachers and learners should deal with them.

Technical vocabulary is subject related, occurs in a specialist domain, and is part of a system of subject knowledge. Each of these features can provide a basis for the deliberate identification of terms.

The subject related aspect of technical vocabulary goes to the heart of what technical vocabulary is. The meanings of technical terms are closely associated with a particular subject area. The best way to determine this for any word is to use a rating scale that classifies words according to how closely related they are to a particular subject area (Baker, 1988; Farrell, 1990; Sutarsyah et al., 1994).

Technical vocabulary is part of a system of subject knowledge. It could thus be identified by referring to specialists who have a good knowledge of the subject area. This can be done in two ways, by using a technical dictionary compiled by a subject specialist or group of specialists (Nation, 2001: 201; Oh et al., 2000: 304), and by making use of clues that the most relevant specialist, the actual writer of the text, used to mark the words considered to be important for the message of the text, since when new terms are introduced in a text, the writers deliberately provide contextual clues to help readers manage new terminology (Bramki and Williams, 1984; Flowerdew, 1992; Williams, 1981).

Technical vocabulary occurs in a specialist domain, so one of the approaches for identifying terms would be to compare the frequency of occurrence of words in a particular subject area with their frequency of occurrence or non-occurrence in another area or range of areas. Technical terms should either only occur in a specialist area or occur with much greater frequency in that area than other areas (Bečka, 1972; Yang, 1986; Baker, 1988; Farrell, 1990; Sutarsyah et al, 1994). This kind of comparison can be done using a computer and any of the various corpora available.

The text used in this study was taken from *Clinically Oriented Anatomy* (Moore and Dalley, 1999, 4th edition, pages 60–71). This text is widely used in first year anatomy courses around the world, is current, and has stood the test of time.

Technical vocabulary is referred to by a variety of labels such as ‘terminological words’ (also ‘terms’ and ‘terminology’, Bečka, 1972), ‘specialised lexis’ (Baker, 1988), ‘technical terms’ (Yang, 1986), ‘specialist vocabulary’ (Kennedy and Bolitho, 1984), and ‘technical words’ (Farrell, 1990). In the discipline of terminology, ‘terminological units’ (Cabré, 1999: 81; Desmet and Boutayeb, 1994), ‘terms’, ‘technical terms’, or ‘terminology’ are used in the sense of ‘technical vocabulary’.

2. Using a rating scale

A rating scale approach to identifying terms involves deciding whether the individual meanings of words fall into the sphere of specialised meaning or not.

Deciding on or interpreting the individual meanings of words depends on the ability of the researchers to draw on their own domain knowledge and to make inferences from domain information within the context in question (Asher and Lascarides, 1996; Bečka, 1972; Štambuk, 1998). At the decision stage, researchers ultimately have to rely on their intuition and knowledge of the field.

One of the possible ways of avoiding the variability of intuitive judgement is to use a carefully designed rating scale. The rating would depend on how closely related the meaning of a word is to the particular subject area, and the scale would range from the least related meaning to the most. There is no previous research that has used a rating scale for this purpose. However, Nagy and Anderson (1984), when estimating the number of words in school reading materials, designed a six-point scale to define the different forms of a word that could be defined as the same word at various degrees of transparency of semantic relatedness. For example, the semantic relatedness of *overlap* and *overlapping* were at the highest degree of transparency whereas *fix* and *prefix* were at the lowest degree.

To ensure the rating scale is used reliably, one possible solution is to measure interrater reliability. Interrater reliability is used to estimate whether there is a reasonable degree of agreement by different raters as to where a lexical item falls on the scale. To make sure that the interrater reliability check works efficiently, the training of raters should be done using the same kinds of materials that are used for the research.

In the first part of this study, words were classified as being technical or non-technical words by rating them on a four point scale designed to measure the strength of the relationship of a word to a particular specialised field. The scale used to do this is in Fig. 1. Items classified at steps 3 and 4 were considered to be technical words. Items at steps 1 and 2 were not. Using such a scale requires good knowledge of the subject area. The first author trained as a nurse and taught public health and so was able to bring specialist subject area knowledge to bear on the classification of the words in the anatomy text. A high level of inter-rater reliability agreement (95%) was achieved with the scale.

The 5500 token sample was made up of 876 lemmas, 227 of which were rated as being technical terms, and 649 as being non-terms (A lemma consists of a headword and its inflected forms which are all the same part of speech as the headword) (see Table 1). The 131 words classified as being at step 3 included *bar*, *column*, *canal*, *disc*, *facet*. Note that some of these words are likely to occur outside the field of anatomy with the same meaning, and some like *bar*, *trunk*, *invest* can occur outside the field of anatomy with a quite different meaning. When using the rating scale, each word was looked at in context to see what meaning it conveyed.

Because this is the most thorough and direct way of identifying terms, it was used as the basis of comparison for the other approaches that follow.

3. Using a technical dictionary

One way of deciding whether a word is a technical term or not is to let a specialist decide, namely to use a technical dictionary, such as a dictionary of anatomy. There

<p>Step 1</p> <p>Words such as function words that have a meaning that has no particular relationship with the field of anatomy, that is, words independent of the subject matter. Examples are: <i>the, is, between, it, by, 12, adjacent, amounts, common, commonly, directly, constantly, early, and especially.</i></p>
<p>Step 2</p> <p>Words that have a meaning that is minimally related to the field of anatomy in that they describe the positions, movements, or features of the body. Examples are: <i>superior, part, forms, pairs, structures, surrounds, supports, associated, lodges, protects.</i></p>
<p>Step 3</p> <p>Words that have a meaning that is closely related to the field of anatomy. They refer to parts, structures or functions of the body, such as the regions of the body and systems of the body. Such words are also used in general language. The words may have some restrictions of usage depending on the subject field.</p> <p>Examples are: <i>chest, trunk, neck, abdomen, ribs, breast, cage, cavity, shoulder, girdle, skin, muscles, wall, heart, lungs, organs, liver, bony, abdominal, breathing.</i> Words in this category may be technical terms in a specific field like anatomy and yet may occur with the same meaning in other fields and not be technical terms in those fields.</p>
<p>Step 4</p> <p>Words that have a meaning specific to the field of anatomy and are not likely to be known in general language. They refer to structures and functions of the body. These words have clear restrictions of usage depending on the subject field. Examples are: <i>thorax, sternum, costal, vertebrae, pectoral, fascia, trachea, mammary, periosteum, hematopoietic, pectoralis, viscera, intervertebral, demifacets, pedicle.</i></p>

Fig. 1. A rating scale for finding technical words (as applied to the anatomy text).

are numerous technical dictionaries and most established specialist fields have at least one and in many cases several.

Oh et al. (2000) used bilingual technical dictionaries in 57 subject areas to develop Japanese term extraction software that involves a series of mathematical formulas. They used information from dictionaries such as the number of dictionaries where a possible term appears, and the ‘distance score’ between the subject area of a possible term and the subject area of a dictionary where the possible term occurs. (For example if a possible term was in the field of medicine and it appeared in a chemistry dictionary, the ‘distance’ between medicine and chemistry was calculated by a specially designed formula). A simpler approach was used in this study. The criterion

for deciding if a word is a technical word or not is to see if it occurs in a technical dictionary for that specialist area. If it does, it is a technical word.

This criterion is relatively straightforward to apply. The first decision is about what technical dictionary to choose if there is a choice. The largest dictionary is not necessarily the best as the largest dictionary may include many words that are not technical terms. There do not seem to be specialist dictionaries only for anatomy, so *Dorland's Illustrated Medical Dictionary* (2000) was chosen. This dictionary is large enough, has been in existence for 100 years, and is up-to-date. Single word items have a main entry and multiword items occur as subentries.

Decisions also need to be made about whether to be a technical word, the word has to appear as a main entry or can appear in a sub-entry, whether the word has to appear in the same form or can include closely related family members and spelling variants, and whether a word has to occur alone in the dictionary or can occur as part of a multi-word group. In this part of the study, the lemma was used as the unit of relationships, so any inflected forms were considered to be part of the same word family. Inclusion in a multi-word unit was not counted, and the lemma could appear as a main entry or a sub-entry.

Out of the 876 lemmas in 5500 token sample, 420 were not in the dictionary or in the rating scale list of terms.

The four words only in the rating scale list and not in the dictionary were *bony*, *clinically*, *nonarticular* and *regenerate*. Related forms of all these words, *bone*, *clinical*, *articular* and *regeneration* were in the dictionary, but because the lemma was chosen as the unit of counting, these were not included.

Of the 229 words occurring only in the dictionary, three were Latin words related to technical words, but not used as technical terms in English. 96 words of the 229 were related to structures of the body but did not have enough subject specificity of meaning to be counted as terms. They included words like *anterior*, *inferior*, *flexibility*, *elasticity*, *mobility*, *triangular*. The remaining 130 of the 229 were words like *breadth*, *diameter*, *adults*, *children*, *defective*, *coughing*, *pain*, *needle*, *knife*. Several of these words may be considered to be technical words in other areas of medicine, but not in anatomy.

If we add together the words not in the dictionary or rating scale list (420) (non-terms accurately identified), the words in both the dictionary and rating scale lists (223), and the four words only in the rating scale list (terms correctly identified), we get a total of 647 out of 876 or an accuracy rate of 73.9%. It is possible that if there was a specialist dictionary of anatomy, this accuracy rate could have been higher. 223 of the 227 terms identified by the rating scale were in the dictionary, and the missing four could reasonably be considered to be in the dictionary because their related forms were there. However only 420 of the 649 non-terms were correctly identified because the dictionary included a lot of words that were not terms in anatomy.

This part of the study was repeated using a different dictionary, the *Oxford Concise Colour Medical Dictionary* (1998, 5th edition). 151 words occurred in both the dictionary and the rating scale list, 45 occurred only in the *Oxford* dictionary, 76 occurred only in the rating scale list, and 604 did not occur in the dictionary or in

the rating scale list (see Table 2). Although fewer of the technical words were identified (151 compared to 223), the overall accuracy rate was higher (86.2%), because the Oxford dictionary included fewer words which were not technical terms.

The major disadvantage of using a dictionary is that we do not know how the compiler(s) of the dictionary selected the words for inclusion. At worst, it is one person's intuitive judgement. At best, it is a group of specialists' systematic judgement. This means that there could be little consistency of decision making between dictionaries.

4. Using clues provided in the text

Writers sometimes signal in the text that a word is a technical term. The most deliberate way that this is done is to explicitly provide a definition for the word while marking through bolding or italicization that the word is special.

The thorax (chest) is the superior part of the trunk between the neck and the abdomen.

Sometimes this defining is done through the use of a synonym in brackets.

bone marrow (hematopoietic tissue) that forms blood cells

Another way in which words are signaled as terms is through labeling in diagrams. Not surprisingly, this is particularly common in the anatomy text, accounting for 45% of the words that have clues provided in the text (see Table 3). The three major types of clues counted in this part of the study were (1) definitions, (2) typographical clues like bolding, italics, and brackets, and (3) labels in diagrams or illustrations.

Table 1
Number of terms and non-terms in the 876 lemma sample of the anatomy text

Total	Step 3	Step 4	
Terms	227	131	96
Non-terms	649		
Total	876		

Table 2
A comparison of the rating scale approach and the dictionary approach

In both the dictionary and rating scale list	223
Only in the dictionary	229
Only in the rating scale list	4
Not in the dictionary or rating scale list	420
Total	876

Table 3
Frequency of types of clues provided by the writer

Pictures	Brackets	Bolding	Italics	Definitions	Total	
Number	223	134	68	48	15	488
Percentage	45.7%	27.5%	13.9%	9.8%	3.1%	100%

Table 4
A comparison of the rating scale approach and the clues-based approach

In both the clues list and the rating scale list	135
Only in the clues list	56
Only in the rating scale list	92
Not in the clues list or rating scale list	593
Total	876

There are several difficulties in finding clues. Firstly, definitions can take a variety of forms (Bramki and Williams, 1984; Flowerdew, 1992) and semi-formal definitions are not always easy to recognise. Secondly, typographical signals can have a variety of functions besides indicating that a word is a technical term. Thirdly, not all labels on diagrams are technical terms and function words like *the*, *of* and *a* can be parts of labels. As Table 3 shows, only 3.1% of the items identified by clues were signaled by definitions.

Out of the 876 lemmas in the 5500 token sample, 593 were not in the rating scale list and were not signaled by clues.

The 135 words in both the clues list and the rating scale list were items correctly identified as terms (see Table 4). The 593 words not in the clues list or the rating scale list were correctly identified as non-terms. This gives a total accuracy rate of 83.1% (135 plus 593 divided by 876) for the clues based approach. However only 135 (59.9%) of the 227 terms identified by the rating scale approach were also identified by the clue-based approach. This is not a satisfactory identification rate. Although clues provided in the text are a useful source of information about terms, they are not an effective means of coming up with a comprehensive list of terms. One major reason for this is that the topic of a particular section of text strongly affects what words the writer focuses on. For example, in the anatomy text, terms related to the thorax are indicated by clues, but not terms which are not so directly related to the thorax even though they occur in the description.

Interpretation of clues in the text requires a lot of judgement similar to that of using the rating scale but without the inclusiveness of the rating scale approach.

5. Using a computer-based approach

Since the early 1990s, there have been many studies of automatic term extraction (Ahmad et al., 1994; Fukushige and Noguchi, 2000; Daille, 1995; Fulford, 2001; Kavanagh, 1995; Heid, 1998/1999; Paziienza, 1998/1999; Vivaldi and Rodríguez,

2001) due to a growing demand for information exchange. Computer scientists researching on extracting terms are constantly developing new computer software in order to obtain more accurate results. The process is called automatic term extraction (Heid, 1998/1999; Pazienza, 1998/1999), automatic term recognition (Kageura and Umino, 1996) or computer-assisted term acquisition (Gamper and Stock, 1998/1999).

Typically term extraction software has used two different approaches: statistical, and linguistic. Statistical approaches basically compare the number of occurrences of a word in a technical corpus with the number of occurrences in a comparison corpus. Statistical approaches differ from each other in two ways: (1) the size and nature of the comparison corpus and (2) the formula used to compare the occurrences.

In the most extreme of the statistical approaches, a “stop list” of non-technical words is used as the comparison. A stop list is a list of words, such as function words, that the computer program is instructed not to consider when performing a count. More often a corpus of general texts is used. Usually the comparison corpus is larger than the technical corpus. Kavanagh’s (1995) *Text Analyser* and Fulford’s (2001) *Textprobe* are based on the ‘stop list’. After the stop words are eliminated, the remaining words are considered to be possible terms. A cut-off point based on frequency can be used to eliminate further words. The list of stop words and the cut-off point can be adjusted by the user.

Kavanagh’s stop word list included the 100 or so most frequent words from the Brown corpus plus some other words such as the names of months. Fulford’s (2001: 263) stop word list included more items than Kavanagh’s: function words, connectors such as *additionally*, *for example* and *otherwise*, “connective verbs” such as *effect*, *perform* and *result in*, and hyphens.

Formulas that compare occurrences are based on the fact that the frequency and range of word forms are different in different types of text. Simple formulas use the difference between the raw frequency and range of word forms in specialised texts and in general language texts (Yang, 1986). More complex formulas exploit the ratio of the percentage of frequency of a word in a technical corpus to the percentage of frequency of the same word in a general corpus (Ahmad et al., 1994). Recently introduced software uses even more complex formulas which combine other information with the frequency and range information such as the frequency of regular expressions over affixes and morphemes (Heid, 1998/1999), mutual information between head and modifier (Pazienza, 1998/1999; Vivaldi and Rodríguez, 2001), the frequency of medical senses of a word (Vivaldi and Rodríguez, 2001), the number of dictionaries where a possible term occurs (Oh, Lee, Lee and Choi, 2000). The common denominator of these formulas is the use of frequency and range of word forms.

Linguistic approaches take advantage of linguistic cues such as word form analysis, part of speech, grammatical structure of possible terms such as noun-noun or adjective-noun structures, lemmatization for inflected forms and/or tokenisation to identify word boundaries using a tagged and/or parsed corpus (Yang, 1986; Ahmad et al, 1994; Heid, 1998/1999; Pazienza, 1998/1999; Vivaldi and Rodríguez, 2001; Oh, Lee, Lee and Choi, 2000). This approach is used as the basis for a statistical approach, that is, it identifies possible terms, and then a statistical approach is applied in order to identify true terms.

Of the two approaches: statistical and linguistic, the statistical approach based on the use of the frequency and range of word forms is the one most often used in term extraction software. This is because terms are generally of higher frequency in specialised texts than in general texts—the most common and typical characteristic of terms from a statistical viewpoint. In order to take advantage of a statistical comparison, the use of formulas is more valid than the use of a stop list. The reason for this is that though the stop word list can be adjusted by the user, it seems that even after eliminating the stop words, the most frequent words from a specialised corpus are not all true terms but include many general words used across a wide range of subjects (Nation, 2001: 203; Sutarsyah et al., 1994).

Technical terms are likely to occur only in a specialised field or to occur with a much higher frequency in a specialised field than in a different field or in a variety of other texts. Such a frequency comparison can be readily carried out by a specially designed computer program with the help of a spreadsheet. To be effective the following conditions need to be met. A large technical text needs to be used in order to provide enough opportunities for the technical terms to occur and be repeated. This text should be free from mis-spellings, and certain kinds of hyphenation would need to be removed (hyphenation of compound words and line-breaking hyphenation).

The non-technical comparison corpus should not contain any texts from the specialised area being investigated and should be large enough for all the high frequency words of the language to appear. Pilot testing showed that the comparison corpus would need to be at least around 2,000,000 words long for this to happen.

The frequency data from the technical text and the comparison corpus figures can be adjusted to match the size of the technical text, and so that a ratio can be calculated comparing the frequency in the technical text with the adjusted frequency in the comparison corpus. Testing showed that a ratio of 1:50 provided the most effective cut-off point. That is, to be classified as a technical term, a word had to occur at least 50 times more often in the technical text than in the comparison corpus, or occur only in the technical corpus (a ratio of infinity). The comparison corpus used in the study was the LOB corpus (Johansson, 1978) combined with the Wellington corpus of written English (Bauer, 1993) with the medical sections (J01–J17 and J22–J31) removed. The large number of words that occur only in the comparison corpus are of no interest in this kind of study and can be removed from the analysis.

The data in Table 5 uses only the 876 words in the sample text used in the investigation of the previous approaches, but uses their frequency in the whole 452,192 token anatomy text to put up against the frequencies in the 1,892,000 token comparison corpus. The program used was RANGE, which is freely available from XX.

The 167 words in both the computer-based list and the rating scale list were correctly identified as terms, and the 595 words in neither list were correctly identified as non-terms making a combined accuracy rate of 87%. However, 60 items in the 227 lemma rating scale list were not identified as terms in the computer-based approach. Here are some of these items starting from those nearest to the ratio of 50 cut-off point—*bone*, *vessel*, *disc*, *facet*, *cardiac*, *trunk*. These are all words that can readily occur in non-technical texts. In a few cases *articulate*, *trunk*, they can occur

Table 5

A comparison of the rating scale approach and the computer-based approach

In both the computer-based list and the rating scale list	167
Only in the computer-based list	54
Only in the rating scale list	60
Not in the computer-based list or rating scale list	595
Total	876

with quite unrelated meanings. In anatomy, *articulate* refers to the connection and movement of body parts, outside the field it can also refer to speaking. Similarly, *trunk* refers to a part of the body in anatomy, but can refer to a suitcase or a major rail route outside that field. These homonyms however make up only a small proportion of the non-overlapping items identified only by the rating scale. The majority are items like *bone*, *bypass*, *jaw*, *vessel*, *skin*, *treatment* which are commonly used both within and outside of the field of anatomy with essentially the same meaning. Thus neither tagging a corpus for part of speech nor tagging homonyms would solve the problem of distinguishing these words as terms.

Fifty-four items were identified by the computer-based approach which were not classified as terms by the rating scale. Here are some of these items starting with those with very high frequencies in the anatomy corpus—*posterior*, *anterior*, *lateral*, *inferior*, *transverse*, *connective*. These words are common in anatomy texts but were not classified as technical terms because they do not directly refer to parts of the body. Rather they are common collocates of technical terms—*anterior ligament*, *transverse process* (here *process* means “a protruding part”). These words would be at home in a technical dictionary but on their own they are not technical terms. If the goal of a computer-based approach is to identify terms *and their collocates*, then it is more successful. The proportion of terms and collocates correctly identified would be 77.8% and the overall accuracy rate 91.9%.

6. Comparison of the four approaches to identifying terms

Table 6 compares the four approaches. However, not all measures are equally important. The percentage of terms and non-terms correctly identified is an overall measure of accuracy. The disadvantage of this measure is that most of the words (649 out of 876) are non-terms and so the accuracy of identifying non-terms has the major effect on this measure. This can be seen in the clues approach where the identification of non-terms is high (91.4%) and the identification of terms is low (59.5%), but there is a reasonable overall accuracy rate of 83.1%.

To be effective, primarily the number of terms correctly identified needs to be high. This is the most important measure. The dictionary approach using Dorland’s dictionary has an excellent rate of almost 100% (98.2%). However it is not effective for identifying non-terms (64.7%) and this has a major effect on the overall rate. To take account of this imbalance, the average of the percentage of terms and non-terms

Table 6
Effectiveness of the four ways of identifying technical terms

Approach	Percentage of terms and non-terms correctly identified	Percentage and number of terms correctly identified	Percentage and number of non-terms correctly identified	Average percentage of terms and non-terms correctly identified
Rating scale	100% (876 lemmas)	100% (227 lemmas)	100% (649 lemmas)	100%
Clues	83.1%	59.5% (135 lemmas)	91.4% (593 lemmas)	75.5%
Dictionary	73.9% (Dorland's)	98.2% (223 lemmas)	64.7% (420 lemmas)	81.5%
	86.2% (Oxford)	66.5% (151 lemmas)	93.1% (604 lemmas)	79.8%
Computer	87%	73.6% (167 lemmas)	91.7% (595 lemmas)	82.7%
	91.9% (with collocates)	77.8% (with collocates)	98.2% (with collocates)	88%

correctly identified is listed in the right hand column of Table 6. This is the average of columns 3 and 4, and is the best overall measure.

The rating scale approach was assumed to be the most accurate so it is given a 100% success rate for all measures. The inter-rater reliability check showed that the scale could be applied consistently and there is no reason to doubt the validity of using the rating scale. There was an almost perfect overlap with terms identified by Dorland's dictionary. It has strong construct validity in that when it is being used each rating involves asking the question "How strongly related is the meaning of this word to the specialist field of anatomy?" which is the essence of being a technical term. It is however laborious to apply.

The clues based approach is also laborious and difficult to apply. It identifies just over half of the terms, although not surprisingly it does well with non-terms because the writer has no reason to highlight or define words not closely connected with the subject matter. However, it cannot be considered an effective way of identifying terms. Too many terms are not explicitly signalled. It has an overall averaged rate of 75.5%.

The dictionary based approach was probably not given its best trial because there is no specialist dictionary of anatomy- this subject area is included in general medical dictionaries. In a subject area where there is a relevant specialist dictionary, this approach may be more successful than in this study. The data from the two dictionaries studied shows how dictionaries can differ from each other and thus how choice of the best dictionary is critical. Dorland's included all the terms but also included a lot of other words. Oxford did not include all the terms, but also did not include many non-terms giving an overall averaged rate of 79.8%.

The computer based approach can be easily applied. It requires a small amount of skill in using a counting program like RANGE and a spreadsheet program like Excel. It has quite a good overall rate and reasonably consistent success in identifying terms and non-terms. To quickly arrive at a rough estimate of the number of technical

terms, this approach is satisfactory, particularly if common collocations of terms are included with terms. To come up with a definitive list of terms, it is not inclusive enough, largely because it also identifies collocates and it has difficulty identifying terms that are also commonly used outside the specialised field. It has an overall averaged rate of 82.7%, rising to 88% if collocates are included as correct identifications.

The goal of this study was to compare approaches to identifying terms so that the most reliable, valid, and practical approach could be determined. The rating scale approach was assumed to be the most valid and comparison with the other approaches has confirmed the wisdom of this assumption. It is however a time-consuming approach as virtually every word has to be checked against the scale. It has been applied to a 93,445 token applied linguistics text containing 5137 different words (Chung, forthcoming 2004; Chung and Nation, 2003) and proved to be manageable. In terms of practicality, the computer-based approach works very well and if common collocates are included as well as terms, it is quite successful.

The reason for carrying out these studies was to arrive at a good method for identifying technical terms so that it is possible to gain some idea of how large technical vocabularies are, how important technical words are in texts, and what technical words are like (Chung and Nation, 2003). Studies of anatomy and applied linguistics texts have shown that technical vocabularies can be very large (4270 in the anatomy text, 835 in the applied linguistics text), technical vocabulary accounts for a very large proportion of the running words in texts (31.2% in anatomy, 20.6% in applied linguistics), and technical vocabularies can differ greatly in the kinds of words they contain (64.4% of the different technical terms in the anatomy text are terms peculiar to anatomy—*costovertebral*, *synovial*, *xiphoid*. 88.4% of the different technical terms in the applied linguistics text studied are words that commonly occur in other uses of the language—*negotiation*, *interaction*, *incidental*) and many of these amongst the commonest words of English.

There are other applications of the methodology for identifying terms, such as dictionary construction, the preparation of glossaries, indexes, and databases, and the preparation of teaching materials. It is hoped that the findings reported in this article will contribute to such research.

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., Pochhacker, F., Kaindl, K. (Eds.), *Translation Studies: An Interdiscipline*. John Benjamins, Amsterdam/Philadelphia, pp. 267–278.
- Asher, N., Lascarides, A., 1996. Lexical disambiguation in a discourse context. In: Pustejovsky, J., Boguraev, B. (Eds.), *Lexical Semantics: The Problem of Polysemy*. Clarendon Press, Oxford, pp. 69–108.
- Baker, M., 1988. Sub-technical vocabulary and the ESP teacher: an analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language* 4, 91–105.
- Bauer, L., 1993. *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English*. Victoria University of Wellington, Wellington.
- Bečka, J., 1972. The lexical composition of specialized texts and its quantitative aspect. *Prague Studies in Mathematical Linguistics* 4, 47–64.
- Bramki, D., Williams, R., 1984. Lexical familiarization in economics text, and its pedagogic implications in reading comprehension. *Reading in a Foreign Language* 2, 169–181.

- Cabré, M., 1999. *Terminology: Theory, Methods, and Applications*. John Benjamins, Amsterdam/Philadelphia.
- Chung, T.M., 2003. Identifying technical vocabulary. Unpublished PhD thesis, Victoria University of Wellington.
- Chung, T.M. 2003 forthcoming. A corpus comparison approach for terminology extraction. *Terminology*.
- Chung, T.M., Nation, I.S.P., 2003. Technical vocabulary in specialised texts. *Reading in a Foreign Language* 15 (2), 103–116. <http://nflrc.hawaii.edu/rfl>.
- Daille, B., 1995. Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. Unit for Computer Research on the English Language Technical Papers 5. Lancaster University.
- Desmet, I., Boutayeb, S., 1994. Terms and words: Propositions for terminology. *Terminology* 1, 303–325.
- Dorland's Illustrated Medical Dictionary. 29th ed. 2000. W. B. Saunders Company, Philadelphia.
- Farrell, P. 1990. Vocabulary in ESP: a lexical analysis of the English of electronics and a study of semi-technical vocabulary CLCS Occasional Paper No. 25 Trinity College.
- Flowerdew, J., 1992. Definitions in science lectures. *Applied Linguistics* 13, 202–221.
- Fukushige, Y., Noguchi, N., 2000. Statistical and linguistic approaches to automatic term recognition: NTCIR experiments at Matsushita. *Terminology* 6, 257–286.
- Fulford, H., 2001. Exploring terms and their linguistic environment in text: A domain-independent approach to automated term extraction. *Terminology* 7, 259–279.
- Gamper, H., Stock, O., 1998/1999. Corpus-based terminology. *Terminology* 5, 147–159.
- Heid, U., 1998/1999. A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology* 5, 161–181.
- Johansson, S., 1978. *Manual of Information to Accompany the Lancaster–Oslo/Bergen Corpus of British English, for Use with Digital Computers*. University of Oslo, Oslo.
- Kageura, K., Umino, B., 1996. Methods of automatic term recognition: A review. *Terminology* 3, 259–289.
- Kavanagh, J., 1995. *The Text Analyser: A Tool for Extracting Knowledge from Text*. Master's of Computer Science Thesis. University of Ottawa, Canada.
- Kennedy, C., Bolitho, R., 1984. *ESP (English for Specific Purposes)*. Pergamon Press, Oxford.
- Moore, K., Dalley, A., 1999. *Clinically Oriented Anatomy*. (4th edition). Lippincott, Williams and Wilkins, Philadelphia.
- Nagy, W., Anderson, R., 1984. How many words are there in printed school English? *Reading Research Quarterly* 19, 304–330.
- Nation, I.S.P., 2001. *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge.
- Oh, J., Lee, J., Lee, K., Choi, K., 2000. Japanese term extraction using dictionary hierarchy and a machine translation system. *Terminology* 6, 287–311.
- Oxford Concise Colour Medical Dictionary. 5th ed., 1998. Oxford University Press, Oxford.
- Pazienza, M., 1998/1999. A domain-specific terminology-extraction system. *Terminology* 5, 183–201.
- Sutarsyah, C., Nation, P., Kennedy, G., 1994. How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal* 25, 34–50.
- Štambuk, A., 1998. Tenets of cognitive linguistics as a framework for terminology research. *Papers and Studies in Contrastive Linguistics* 34, 233–243.
- Vivaldi, J., Rodríguez, H., 2001. Improving term extraction by combining different techniques. *Terminology* 7, 31–47.
- Williams, R., 1981. Lexical familiarization in content area textbooks. In: Chapman, L. (Ed.), *The Reader and the Text*. Heinemann Educational Books Ltd, London, pp. 49–59.
- Yang, H., 1986. A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing* 1, 93–103.