The contribution of verbal semantic content towards term recognition

Eugenia Eumeridou, Blaise Nkwenti-Azeh, John McNaught

Department of Information and Communication Systems, University of the Aegean, Karlovassi, Samos, Greece (evmoir@icsd.aegean.gr)

Centre for Computational Linguistics, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK (blaise@ccl.umist.ac.uk)

Department of Computation, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK (Jock@co.umist.ac.uk)

Abstract

Automatic term recognition is a natural language processing technology which is gaining increasing prominence in our information-overloaded society. Apart from its use for quick and efficient updating of terminologies and thesauri, it has also been used for machine translation, information retrieval, document indexing and classification as well as content representation. Until very recently, term identification techniques rested solely on the mapping of term linguistic properties onto computational procedures. However, actual terminological practice has shown that context is also important for term identification and interpretation as terms may appear in different forms depending on the situation of use. The aim of this article is to show the importance of contextual information for automatic term recognition by exploiting the relation between verbal semantic content and term occurrence in three subcorpora drawn from the British National Corpus.

KEYWORDS: *Automatic term recognition, BNC, corpora, selectional restrictions, special languages, verbal semantic content, WordNet*

1. Introduction

The rapid growth of science and technology and the globalisation of our societies necessitate the construction of accurate, consistent and comprehensive terminologies as a means to improve communication or access subject specific information. By terminologies, we refer to the concepts that make up the knowledge of a specific domain together with their linguistic realisations, the terms. In the past, such a task used to be particularly laborious, as terms had to be manually extracted. However, the introduction of computers to language analysis has provided terminologists with huge amounts of data to help them construct the conceptual framework of a field and study term usage, as well as techniques to process and filter this data. As a result, terminological practice, is now heavily corpus-based.

This innovation in practice brought a breakthrough in theory as well. Having the opportunity to study a wide range of texts, terminologists came to the realization that terms are not unique manifestations of concepts as was previously held (Wüster 1978), and thus totally independent of linguistic context. On the contrary, they depended on context to convey their specialized meaning, as well as the form they would appear in (Dubuc and Lauriston 1997). Based on this realization, we examine the importance of context to term recognition, as software developed so far has heavily disregarded this aspect.

Current term recognition systems have either used linguistic techniques, mostly exploiting morphosyntactic properties of terms, e.g. term formation patterns to extract candidate terms (Bourigault 1992; Ananiadou 1994; Jacquemin and Royaute 1994;

Nkwenti-Azeh 1994; Dagan and Church 1995; Oueslati et al. 1996) or statistical techniques to measure the degree of unithood or termhood of the candidate multi-word terms (Smadja 1991; Damerau 1993; Haas and He 1993; Enguehard and Pantera 1994; Cohen 1995) or a combination of both resulting in hybrid systems (Daille et al. 1994; Franzi and Ananiadou 1996; Maynard and Ananiadou 1999). In the current state of the art, however, no system performs term recognition in the actual sense of the word. They simply provide researchers with lists of candidate terms which need further validation by a subject specialist. Additionally, the evaluation of such systems has proved difficult. In a review of 12 current systems of term extraction Castellví et al. (2001) note that their evaluation is not carried out in great depth. They conclude that *"Broadly speaking there is neither clear nor measurable explanation of the final results. [….] it is difficult to evaluate and compare them."* This is partly due to lack of a common test bench together with criteria to carry out such an evaluation, and in particular of a marked-up for terms corpus in which all terms will have been previously successfully identified, against which a given technique will be run to measure how well it has done in comparison with this "gold standard" (Kageura et al. 1998). Moreover, a system is usually trained on small and highly specialised corpora with regard to the topic as well as the specialisation degree, which makes it difficult to use or test these systems in different environments. On the whole, all systems propose large lists of candidate terms which have to be manually checked for termhood.

Among the shortcomings of current systems, the review lists their exclusive preoccupation with noun phrases while none of them deals with verbal phrases. Although terms in their majority are nouns, nevertheless special languages have their

own verbs as well, no matter how low the ratio is in comparison with nouns. Additionally, most systems concentrate on noun compounds while little work has been done on single word terms. These latter terms are often general language words which have acquired terminological usage either through extension of their meaning (simile, metaphor), through narrowing of their meaning or through meaning transfer (semantic drift). Such terms, particularly the last type, are the most difficult to identify as semantic information is needed to distinguish between the general usage of a word and its terminological usage. The importance of semantic information and its incorporation in future systems for automatic term recognition is also stressed in the review. Finally, Castellví et al. (2001) mention the importance of taking into account the type of constraints terminological units present with respect to conceptual field and text type.

In this paper, we examine the importance of context for term recognition, focusing on verb ? term relation in three subcorpora belonging to different subject fields and text types. It is important to clarify that, at this point in our research, we have not aimed at the construction of an automatic term extraction or recognition system but rather are interested in establishing whether and to what extent surrounding linguistic context determines the presence of terminological units and thus can serve as an indicator of term presence. As carriers of contextual information, we have chosen verbs and we have so far examined the relation of verbal form, verbal subcategorisation patterns and verbal semantics to term occurrence. Most importantly, the contribution of each dimension of verbal information (form, syntax, semantics) to term recognition has been examined in relation to the particular special language and text type that verbs and terms occur in. In

this paper, we will present one aspect of our research only, namely the contribution of verbal semantic content towards term recognition.

2. Motivation

Our interest in verbs as carriers of contextual information arises from the verb's traditional role as the central organiser and distributor of concepts in a sentence. In many languages, the verb is the only necessary element in the sentence, whereas even the subject can be replaced by a non-referential noun that is a dummy subject *it*, e.g. *It is raining*. Moreover, the predicate-argument structure of the verb determines the syntactic and semantic structure of the sentence, since its arguments carry the grammatical relations and semantic roles in the sentence. Furthermore, the selectional restrictions imposed by the verb's subcategorisation frame determine the semantic properties of the nouns that occupy argument positions. The above verb properties have led linguists such as Fillmore (1968) and Chafe (1970) to consider the verb as the central element in the sentence and to argue for a verb-based model of sentence meaning.

The central position of the verb in the sentence has however been challenged by Gentner and France (1988). Having carried out a number of experiments in which subjects were more likely to paraphrase the meaning of a verb rather than of its argument noun in cases of semantic strain, they arrived at the conclusion that it is verbs which change their meanings according to the noun that precedes or follows rather than the other way round. The tendency of the verb to yield in semantic change is referred to as the **mutability of verbs** (Gentner and France 1988). Among the reasons cited to account for this tendency were the high degree of verbal polysemy[1], especially for the most frequent verbs,

together with the fact that the representations of verbal concepts are less internally cohesive than the representations of nominal concepts.

However, it should be noted that even if nouns are considered to be more concrete semantically and less likely to yield to semantic change in cases of semantic strain, this does not invalidate the fact that verbs exercise semantic control over their arguments either by means of their semantic content or via the thematic roles they assign to them. Additionally, we should bear in mind that the above experiments concerned the combinatorial semantics of nouns and verbs in general language usage. In special language texts, however, vocabulary is far more restricted, and verbs appear in fewer senses (Basili et al. 1996) while some of them are terms themselves. Hence, at times, their conceptual representations can be equally if not more concrete semantically than the representations of their argument nouns.

3. Verbal semantic content and term recognition

Our testing hypothesis has been that verbal semantic content could contribute to term recognition. Verbs impose selectional restrictions on their argument nouns. As a result, they subcategorise for nouns of compatible semantic content. Taking the argument further, verbal semantic classes combine with compatible nominal semantic classes. Considering the fact that term rate varies across nominal classes, certain verbal classes will prove better probes for termhood than others.

However, the combinatorial semantics of nouns and verbs as potential indicators of termhood can be exploited in the opposite way as well, that is, by breach of their

regularity. In other words, we can examine whether violation of the selectional restrictions a verb imposes on its argument nouns could signal term presence.

Finally, having tested the validity of our hypotheses in each subcorpus, we need to examine whether and to what extent their effectiveness is constrained by the subject domain or text type they apply in. An explication of our findings is given at the end of this paper.

4. Methodology

In this section, we briefly describe the tools and resources used and the process followed to extract our results. The tools and resources used for the analysis of verb ? term relation include 1) the corpus from which sample subcorpora were selected, 2) the software used to extract the necessary environments for the study of verb and term behaviour, 3) the set of subcategorisation patterns and their tags, employed for the manual parsing of each subcorpus sentences, 4) the set of WordNet semantic classes for nouns and verbs used in the semantic annotation of the three subcorpora, 5) The dictionaries we used to decide which nouns and verbs in our subcorpora are terms.

To investigate the contribution of verbs towards automatic term recognition, we studied a wide range of verbs in different environments based on a selection of texts found in the British National Corpus (BNC) (Leech 1993). The BNC is a general corpus, yet contains a wide number of specialized texts classified for domain, time of publication, medium (book, periodical etc.) and level of technical difficulty. Additionally, all texts are grammatically tagged using the CLAWS tagging system (Leech et al. 1994). They also

share the same encoding system, which makes use of Standard Generalised Markup Language (SGML) (ISO 8879 1986) and conforms to Text Encoding Initiative (TEI) (Dunlop 1995). These features were the main reasons we chose the BNC over a simple collection of technical documents. The three subcorpora we extracted from the BNC for our analysis are comparable in size and are all marked for a high level of technicality, as we are interested in rich terminologically environments. However, they fall into different subject fields as our aim was to investigate how our hypotheses are constrained by the special language factor. As the BNC classification of medium is rather too broad to adequately define the specialized character of the three subcorpora in terms of text type, we additionally used Sager's classification of text types (Sager et al. 1980). Thus, our first subcorpus, the FRT subcorpus (34,136 words), is an extract from a textbook on *futures regulations.* In terms of text type, it is a regulation and in terms of subject matter it belongs to the commercial special language. The second subcorpus, the CMT subcorpus (20,293 words), is a sample from a *Microsoft manual.* It is a handbook and belongs to the computing special language. Finally, the third subcorpus, the EMT subcorpus (26,467 words), comprises a series of *lectures on electromagnetic theory.* In terms of text type, it is a lecture and belongs to the electrical engineering special language.

To carry out our corpus analysis of the three special language subcorpora, we used Xtract (Smadja 1991), a collocation retrieval tool, which provided us with long enough contexts to study different occurrences of the same verb. The program was run in three stages:

1. During the first stage, a file was output with all the frequencies of occurrence of the verbs. Only verbs with frequency higher than three were extracted, as verbs of a lower frequency were not considered informative enough of a subcorpus' semantic content, thus most likely they would not be frequented by terms.

2. Sentences were extracted, containing instances of these verbs.

3. For each verb, n-grams were extracted of all the words occurring within five positions (the maximum position range allowed for in Xtract) both before and after the verb. However, this last stage did not prove very useful in our analysis, as the arguments of the verb were frequently found at a greater distance than five preceding or following positions.

Once the sentences containing instances of each subcorpus verb were extracted, they were manually parsed. For the syntactic parsing, we used the Oxford Advanced Learner's Dictionary of Current English (Hornby 1974) due to its detailed and comprehensive list of verbal syntactic patterns. Tables for each verbal syntactic pattern were then constructed containing all the instances of verbs following this pattern together with their argument frames.

In a second phase, all selected verbs together with their argument nouns were manually semantically annotated, using WordNet (Miller et al. 1990) semantic labels. At this point, tables were constructed for each verbal class in the corpus containing all instances of selected verbs falling into this semantic class, together with their arguments classified in terms of nominal semantic class and argument position.

Next, all instances of nouns falling into the same nominal semantic class and argument position were counted for each verbal semantic class, in order to derive which are the dominant nominal semantic classes in each argument position.

Finally, instances of terms were counted for each nominal class to establish those nominal classes that were the richest in terminological units. To decide on the termhood of nouns and verbs in our research, we consulted expert opinion as well as a variety of dictionaries, e.g. *A Dictionary of Finance* (Butler and Isaacs (eds.) 1993), *Dictionary of Commercial, Financial and Legal Terms* (Herbst 1966), *Dictionary of Electrical Engineering* (Jackson and Feinberg 1981), *A Dictionary of Law* (Curzon 1982), *Elsevier's Dictionary of Personal and Office Computing* (Vollnhals 1984) and *The 3-D Visual Dictionary of Computing* (Graham 1995). At this point, it should be noted that our approach to term definition is the pragmatic one (Pearson 1998). That is, we do not include as terms only items belonging to the specialised vocabulary of the given special language but also items which belong to different special languages or which at first sight appear to belong to the general language but have acquired a precise meaning in a specialised context and thus have acquired terminological status, e.g. *bank, customer, rules* (FRT subcorpus), *copy, file, line, character* (CMT subcorpus), *factor, force, current, surface* (EMT subcorpus).

The results of our analysis showing the overall distribution of verbal and nominal semantic classes in each corpus are presented in tables 2, 4 and 7 in the following sections. Once such tables were compiled, we were able to see the overall semantic patterns emerging in each subcorpus and the significance of each verbal semantic class in each subcorpus as a probe for terminologically rich environments.

In the following sections, we examine how evidence drawn from the three subcorpora supports our hypotheses.

5. Verbal semantic content and term recognition in the FRT

In this section, we explore the contribution of verbal semantic content to term recognition in the FRT subcorpus. To start with, we carried out a frequency analysis to establish which are the prevalent verbal semantic classes in the FRT subcorpus. The results of this analysis are presented in Table 1, in which verbal classes are listed in decreasing order of frequency of occurrence together with the number of occurrences for each class. Additionally, columns 4 and 5 present their relative frequency in comparison to the whole of the subcorpus together with their synset size in WordNet.

Verbal classes	**FRT rank**	**Freq.**	**Total%**	**WordNet Rank**
Stative verbs	1	407	18%	9
Possession verbs	2	375	17%	11
Communication verbs	3	326	15%	3
Social verbs	4	325	15%	4
Consumption verbs	5	278	12.5%	5
Cognition verbs	6	210	9%	6
Change verbs	7	155	7%	2
Creation verbs	8	106	5%	7
Competition verbs	9	33	1%	8
Contact verbs				1
Perception verbs				10

Table 1: Semantic verbal class distribution in WordNet and in the FRT

According to table 1, the prevalent verbal classes in the corpus are stative, possession, communication and social verbs. Our results are in perfect accord with the subcorpus which in terms of text type is a regulation and in terms of subject matter belongs to the commercial special language. More specifically, stative verbs are used to describe the legal state concerning futures regulations, communication and social verbs are used to convey the legal activities in the subcorpus, whereas possession verbs are used to express

the commercial activities. It is also interesting to note that certain verbal semantic classes which have a high rate of occurrence in general language, i.e. which rank highest in the WordNet scheme, do not occur or are among the lowest ones in the FRT subcorpus, e.g. *contact verbs, change verbs*, whereas verbal semantic classes with a limited presence in general language verb classifications have a dominant presence in our subcorpus, e.g. *possession verbs.* The difference suggests that, in special language texts, the distribution of verbal classes is different than in general language and can serve as an indicator of the special language a text belongs to.

Once the prevalent verbal semantic classes have been established, the next step is to determine the main nominal classes they subcategorise for as well as the potential contribution of verbal ? nominal combination for term prediction.

At this point, new tables were constructed, this time for each verbal class in the subcorpus containing all instances of selected verbs falling into this semantic class together with their arguments classified in terms of nominal class and argument position. Next, all instances of nouns falling into the same nominal class and argument position were counted for each verbal semantic class, in order to derive the dominant nominal classes in each argument position. Finally, instances of terms were counted for each nominal class to establish the nominal classes which are richest in terminological units.

The results of this process are presented in the table 2 which shows the prevalent verbal semantic class ? nominal semantic class combinations in the FRT subcorpus, as well as the rate for each combination. More specifically, in the first column we list the prevalent nominal semantic class preceding each verbal semantic class in the subcorpus; in the

second column, we present the term frequency and percentage rate for each such combination; in the third column, we have the verbal classes encountered in the FRT subcorpus listed in decreasing order of frequency; in the fourth column, we have the predominant nominal semantic class following each given verbal semantic class. Finally, in the fifth column, we have again the term frequency and percentage rate estimated for each such pair.

Nominal class	**Term %**	**Verbal class**	**Nominal class**	**Term %**
State nouns	17 (44%)	Stative verbs (intr)		
Communication nouns	27 (93%)	Stative verbs (tr)	Communication nouns	19 (69 %)
Person nouns	36 (100%)	Possession verbs	Communication nouns/ Possession nouns	72 (78%) 43 (96%)
Group nouns	40 (100%)	Communication verbs	Communication nouns	100 (78%)
Communication nouns	54 (94%)	Social verbs	Act nouns	71 (81%)
Communication nouns	54 (98%)	Consumption verbs	Communication nouns	53 (84%)
Person nouns	64 (100%)	Cognition verbs	Communication nouns	32 (92%)
Communication nouns	38 (100%)	Change verbs	Possession nouns	45 (100%)
Person nouns	27 (96%)	Creation verbs	Communication nouns	49 (92%)

Table 2: The prevalent verbal semantic class – nominal semantic class combinations and term occurrence in the FRT subcorpus

Table 2 shows that verbal classes vary in which nominal classes will precede or follow. Additionally, it shows that term rate varies depending on which nominal class precedes or follows. On the whole, the best combinations of verbal ? nominal classes in order of decreasing frequency are person nouns ? possession verbs, group nouns ? communication verbs, person nouns ? cognition verbs, communication nouns ? change verbs, change verbs ? possession nouns, communication nouns ? consumption verbs, person nouns ? creation verbs, possession verbs ? possession nouns, communication

nouns ? social verbs, communication nouns ? stative verbs (transitive), cognition verbs ? communication nouns and creation verbs ? communication nouns.

The above information is crucial for term recognition purposes as it shows which are the semantically richest environments for terms in a corpus. Therefore a term recognition system could be guided to first look into these verb ? noun combinations, assigning to them higher term weights with respect to the rest of the corpus.

However, verbs are important for term recognition not only in terms of the semantic regularity of the patterns they display with their arguments but also when this semantic regularity of the verb ? noun relationship is violated. The above argument is supported by the following examples (underlined words in the following examples are terms):

1. Chinese wall authorises the withholding of information.
2. Apply derivatives instruments.
3. Give liquidity.
4. Trade transparency.
5. Take margin.

In the first example, we have an instance of a social verb, the verb *authorise*. *Authorise*, in the FRT subcorpus, strictly subcategorises for person, group and communication nouns in subject position, e.g. *the customer authorises, the FSA authorises, the trade custom authorises, the 1986 Act authorises,* etc. However, *wall* being an artefact noun falls into none of the above nominal categories. This is a case of semantic violation which points to terminological usage. *Wall* in the first example is a legal term referring to a *rule* and

rule is clearly a communication noun. In the second example, we have an instance of a consumption verb, the verb *apply*. *Apply* equally subcategorises for communication nouns in object position in the FRT subcorpus, e.g. *apply the rules*, *the COB rules*, *the rule against churning*, *regulatory rules*, *blanket consents*, etc. However, *instruments* is an artefact noun. This is also a case in which a noun has changed its meaning (an artefact noun has become a communication noun), showing once more that verbs largely determine the semantics of their argument nouns. In the last three examples, all verbs are possession verbs. According to the corpus analysis, possession verbs are followed by possession nouns. However, none of the above argument nouns is a possession noun. *Liquidity* is a state noun, *transparency* is an attribute noun and *margin* is a shape noun. The above nouns are all used with a possession noun sense in our subcorpus. They are all general language words which have acquired a different meaning in our subcorpus and have become terms. The change of their meaning has been signalled by the violation of the selectional restrictions the preceding verb imposes. It is worth mentioning at this point that our findings contradict Gentner and France's position that verbal meaning depends on nominal meaning.

6. Verbal semantics and term recognition in the CMT

The same type of analysis is followed in the CMT subcorpus as well. The CMT subcorpus is an extract from a Microsoft manual. In terms of special language, it falls into the computing special language and in terms of text type, it is a handbook, presenting the right way to do something, outlining and detailing operations and processes. A

semantic analysis of the corpus in terms of its verbal semantic classes yields the results presented in table 3.

Verbal classes	CMT rank	Freq.	Total%	WordNet rank
Communication verbs	1	731	26%	3
Change verbs	2	475	17%	2
Contact verbs	3	442	16%	1
Perception verbs	4	241	8%	10
Cognition verbs	5	222	8%	6
Creation verbs	6	175	6%	7
Possession verbs	7	145	5%	11
Stative verbs	9	126	4%	9
Social verbs	10	114	4%	4
Motion verbs	8	109	4%	
Consumption verbs	11	48	2%	5

Table 3: Verbal semantic class distribution in WordNet and in the CMT

Table 3 shows that the prevalent verbal semantic classes in the CMT subcorpus are communication verbs, change verbs, contact verbs and perception verbs. Communication verbs are a prominent verbal class in the subcorpus, for it is a handbook. They are largely used to convey directions. Change verbs are the second most frequent verbal class, since the corpus largely describes processes. Contact verbs are the third most frequent verbal class, since contact is necessary to carry out these processes and perception verbs are the fourth ranked verbal class, since the results of these processes are visually perceived. Considering the fact that all the above classes, apart from communication verbs, rank low or are non-existent in the FRT subcorpus, the observed differences strongly suggest that different verbal semantic classes are prominent in different special language texts.

Next, a distributional analysis of nominal classes in terms of argument positions around corpus verbal semantic classes together with the term rates for each class was carried out along the lines described in section 4. The results of the analysis are reported in table 4.

Nominal class	Term %	Verbal class	Nominal class	Term %
		Communication verbs	Communication nouns	14 (100%)
		Change verbs	Communication nouns	30 (86%)
		Contact verbs	Artefact nouns	63 (92%)
		Perception verbs	Communication nouns	11 (64%)
		Cognition verbs	Communication nouns	21 (81%)
		Creation verbs	Communication nouns	5 (100%)
Artefact nouns	2 (25%)	Possession verbs	Communication nouns	11 (73%)
		Motion verbs	Artefact nouns	5 (83%)
Communication nouns	4 (100%)	Stative verbs (intr)		
		Consumption verbs	Communication nouns	13 (100%)

Table 4: Semantic verbal class ? nominal class combinations and term occurrence in the CMT subcorpus

According to table 4, the best verbal class ? nominal class combinations for term recognition are communication verbs ? communication nouns, creation verbs ? communication nouns, consumption verbs ? communication nouns and contact verbs ? artefact nouns.

It is worth noticing that we have a small number of nominal classes preceding verbal semantic classes in the CMT subcorpus with a low term rate. Additionally, the majority of nouns following CMT verbs are communication nouns, regardless of the verbal semantic category. As a result, we note that verbal semantic category plays an insignificant role in determining which nominal class will precede or follow. This is probably due to the highly restricted vocabulary of the CMT subcorpus, in which choices are too limited for patterns to emerge.

However, verbal semantics have proved a significant factor for term recognition in the CMT subcorpus in a different sense. The CMT subcorpus, being a directive text which largely describes processes, has a great number of verbs which are terms themselves.

Such verbs are change verbs ? *inset*, *clear*, *merge*, *start*, *insert*, *highlight* and *delete*, perception verbs ? *display*, contact verbs ? *operate*, *indent*, *tap*, *switch on*, stative verbs ? *start* and *hold*, communication verbs ? *type*, *enter*, *underline* and *record*, the motion verb ? *move*, creation verbs ? *undo* and *format*, possession verbs ? *retrieve*, *save* and *get*, and the cognition verbs ? *read*, *check* and *select*.

The significance of verb terms for term recognition, apart from being terms themselves, is that they are always preceded or followed by terms. This is due to the fact that, having a very specific meaning themselves, they exercise a greater number of selectional restrictions on their argument nouns than general language verbs. Therefore, they can only be accompanied by a noun argument with an equally specific meaning, that is a term. As a result, such verbs constitute the perfect environments for term recognition. Table 5 provides some examples (terms are given in boldface):

Verbs	Subject	Object
Merge		**Documents**
Start	**page numbers, inset paragraphs**	another **command**
		a new **document**
Insert		**document**, a **character**, a **line** , a **hard hyphen**
Highlight		**text**, the whole **document**
Load		**Files**
Delete	the **backspace key**	**text**, a **character**, **file**, **block of text**, **highlighted text**, the **phrase**
Display		the **ruler line**
Operate	**commands**, the **computer**	**Toggles**
Tap		**tab key, Enter, E, Esc,** the **function key,** the **backspace key,** the **Enter key, Caps Lock**
switch on		**toggles, page numbering**
Type	**Caps Lock**	
Enter	**Numbers**	
Underline	the **heading**	
Undo		**printing, edit, commands**
Format		the **page**
Retrieve	**DBMS**	
save	**Word**	**documents, files**
Get		**help information**
Read		**programs,** the **character**
Select		**Copy,** the **phrase,** the **text,** the **non-printing symbols, left indent, right indent**

Table 5: Verb terms and term density in the CMT

An additional function of verbs in the CMT subcorpus is as a part of a term in a compound. Verbs can occur in nucleus position, modifier position or both positions in a two word or a multi-word compound. Examples of such cases are the following compound terms:

1. Verb in nucleus position: *Line* ***Draw****, Column* ***Select****.*

2. Verb in modifier position: ***replace*** *facility*, ***read-write*** *head*, ***Shift*** *key*, ***Shift****-Lock key*, ***Function*** *key*, ***Scroll*** *Lock*, ***Break*** *mark*, ***write-protect*** *notch*, ***PRINT*** *PREVIEW*, ***Print*** *Screen*, ***print*** *server*, ***Extend*** *selection key*, ***Scroll*** *lock keys*.

3. Verb in both nucleus and modifier position: ***Transfer - Merge***.

Thus, once we encounter a verb term followed or preceded directly by one or more nouns, we can extract the sequence and guarantee its termhood.

7. Verbal semantic content and term recognition in the EMT

In this section, we finally examine the validity of our hypotheses in the EMT subcorpus. The EMT subcorpus is a series of lectures on electromagnetic theory. Its special language lexicon belongs to the electrical engineering special language. This field has a long established terminology, a fact that has helped eliminate the number of grey areas between terms and words. In terms of text type, it is classified as a lecture (Sager et al. 1980). However, as our corpus constitutes a series of lectures, it mostly functions as a textbook. As such, it is by definition a special text and taking into account the fact that textbooks "*ideally exemplify the relativity of the notions of general and special knowledge*" and that "*all new items of information are new for the reader and are special*" (Sager et al. 1980), it makes this corpus a good candidate for the study of terms.

To start our analysis, we carried out a frequency analysis for all verbs in the corpus with range of occurrence higher than three to establish the main verbal semantic classes in the subcorpus. The results of the frequency analysis are presented in table 6.

Verbal classes	**EMT rank**	**Freq.**	**Total%**	**WordNet rank**
Stative verbs	1	432	23%	9
Possession verbs	2	344	18%	11
Cognition verbs	3	336	18%	6
Change verbs	4	245	13%	2
Communication verbs	5	242	13%	3
Motion verbs	6	110	5%	--
Consumption verbs	7	65	3%	5
Creation verbs	8	52	3%	7
Perception verbs	9	35	2%	10
Contact verbs	10	31	1.5%	1
Social verbs	11	11	0.5%	4

Table 6: Distribution of verbal semantic classes in the EMT subcorpus and WordNet

According to this table, the prominent verbal classes in decreasing order are stative verbs, possession verbs, cognition verbs, change verbs, communication verbs and motion verbs. The extensive use of stative, change and motion verbs can be explained in terms of the special language the subcorpus belongs to, that is, the corpus describes states, motions and changes involving physical phenomena. Moreover, the significant number of cognition and communication verbs is largely due to the fact that the corpus is a lecture. Furthermore, the extensive use of possession verbs is due to the fact that among possession verbs we have a number of very general usage verbs such as *give, take* and *get* which have a great number of occurrences, yet they have little terminological value.

We have previously suggested that verbs determine which noun classes will precede or follow by means of selectional restrictions they impose on their arguments, thus subcategorizing for nouns of compatible content. Corpus analysis in the EMT subcorpus further supports our hypotheses, as cognition nouns and communication nouns are the most prevalent nominal classes in our subcorpus. Furthermore, the various changes, motions and states described in the subcorpus involve substances, phenomena and attributes of those substances and events, thus rendering the following nominal classes — phenomena nouns, event nouns, attribute nouns, object nouns and substance nouns — the main nominal classes in the subcorpus.

To clearly show which verbal class ? nominal class combinations are the best environments for term recognition, we carried out frequency analyses concerning nominal class distribution around each verbal class and term rates in each class. The best combinations both in terms of frequency of occurrence and term rate are presented in table 7.

Nominal class	**Term %**	**Verbal class**	**Nominal class**	**Term %**
Phenomenon nouns	14 (100%)	Stative verbs (intr)		
Communication nouns	24 (96%)	Possession verbs	Communication nouns	42 (86%)
		Cognition verbs	Phenomenon nouns	20 (95%)
Phenomenon nouns	19 (93%)	Change verbs (intr)		
		Change verbs	Phenomenon nouns	13 (100%)
Person nouns	0 (0%)	Communication verbs	Communication nouns	15 (79%)
Phenomenon nouns	16 (100%)	Motion verbs (intr)		
Person nouns	0 (0%)	Consumption verbs	Cognition nouns	7 (36%)
Phenomenon nouns	4 (100%)	Creation verbs	Phenomenon nouns	6 (100%)
Communication verbs	3 (100%)	Perception verbs (intr)		
		Contact verbs	Artefact nouns	8 (100%)
Person nouns	0 (0%)	Social verbs	Artefact nouns	3 (92%)

Table 7: Semantic verbal class – nominal class combinations and term occurrence in he EMT

According to this table, the best environments for terms are provided by the following combinations in order of decreasing frequency of occurrence: phenomenon nouns ? stative verbs, phenomenon nouns ? motion verbs, change verbs ? phenomenon nouns, communication nouns ? possession verbs, cognition verbs ? phenomenon nouns, phenomenon nouns ? change verbs (intr). Remaining pairs with high term rates are not listed as they have too low frequencies.

8. Conclusions

The above findings strongly suggest that distinct verbal classes subcategorise for distinct nominal classes. As the rate of term occurrence varies between nominal classes, the importance of verbal semantic classes for term recognition varies as well. Results drawn from all three subcorpora suggest that certain verbal class ? nominal class combinations are exceptionally good environments for terms whereas others are to be totally avoided. However, the importance of verb ? noun combinatorial semantics varies in different corpora. It seems to work better in less technical corpora, e.g the FRT subcorpus, which allow for a wide range of nominal classes, thus allowing for an equally wide range of patterns to emerge in which verbs play a discriminating role. However, its effectiveness declines as we move into more technical texts. The CMT subcorpus for instance is a highly technical text both due to subject matter ? computing is considered to be a more technical special language than commerce ? and due to text type, as manuals and handbooks have a highly controlled language. As a result, they are poor environments to support our hypotheses as nominal classes are few, highly repetitive and occur irrespectively of which verbal class precedes or follows.

Along the same line that verbs, through the selectional restrictions they impose on their argument nouns, should subcategorise for nouns of compatible semantic content, we have found that verbs with a highly specific content, mostly terms, also subcategorise for nouns of an equally specific content that are also terms. Such terms in our corpus, the CMT subcorpus, are general language words which have become terms through a restriction of their general language meaning.

The second aspect we chose to study across the three corpora has been the contribution of the violation of the regularity encountered in verb ? noun combinatorial semantics to the prediction of term occurrence. Additionally, we mentioned the close link between this verbal semantic dimension and the specific type of term formation pattern, namely the semantic drift. Terms resulting from semantic drift are general language words which have acquired a different meaning in the special language text they are encountered in. As a result, they are difficult to detect as there is no way to distinguish between their general language and terminological usage, unless context is taken into account. In the FRT subcorpus, we encountered words from general language that had acquired terminological meaning. In this case, breach of the selectional restrictions of the preceding verb pointed to the terminological usage of the word.

At this point, it should be noted that the last two cases mentioned concern single word terms which have been transferred from general language usage and thus contextual information is necessary to disambiguate between their general and specialized usage.

The third subcorpus, the EMT subcorpus, has no instances of single word terms, as its main term formation pattern is the combination of adjective followed by a word. In this

subcorpus, we can only exploit the combinatorial semantics of its nouns and verbs, yet to a lower extent than in the FRT subcorpus.

To conclude, our results, though promising, are of an indicative nature only. The most significant aspect of our research is that we have shown verbs add a new dimension in our efforts to extract terminological units from texts. The findings presented in the previous sections form the basis of further research in this direction. In particular, we can

(a) analyse other, larger corpora in the same domain to arrive at domain-level universals, i.e. study of verbal contribution to term recognition in corpora belonging to one special language, e.g. commercial, but to different text types to see whether we can make generalisations about the behaviour of certain verbs in a given domain irrespective of text-type;

(b) analyse other larger corpora of the same text type to arrive at text-type universals, i.e. study of verbal contribution to term recognition in corpora belonging to the same text type but to different special languages;

(c) incorporate the results in an automatic term recognition system, by examining the most productive verbal semantic patterns for term recognition in each special language corpus.

---

[1] [i]Statistics of verb ? noun distribution in lexicographic works may seem to justify this claim. Miller et al. (1990) note, for example, that the Collins English Dictionary lists 43,436 different nouns and 14,190 different verbs which are highly polysemous. In

particular, we notice that the more common a verb is, the more polysemous it is, e.g. *have, be, run, make, set, go, take* and others. For example, in the sentences *I have a Mercedes* and *I have a headache* the meaning of the verb depends totally on its object.

References

Ananiadou, S. 1994. "A Methodology for Automatic Term Recognition". *Proceedings of COLING-94*, Kyoto, Japan, 1034-1038.

Basili, R., Pazienza M.T., Velardi, P. 1996. "Interpreting General-purpose and Corpus-based Verb Classification". *Computational Linguistics.* Volume 22 (4), 559-568.

Bourigault, D. 1992. "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases". *Proceedings of COLING-92*, Nantes, France, 977-981.

Butler, B. and Isaacs A. (eds.) 1993. *A Dictionary of Finance*. Oxford: Oxford University Press.

Castellví, M. T. C., Bagot, R. E. and Palatresi J. V. 2001. "Automatic Term Detection: A Review of Current Systems". In *Recent Advances in Computational Terminology.* Bourigault D., Jacquemin C. and L'Homme M-C. (eds). Amsterdam/Philadelphia: John Benjamins Publishing Company, 53-89.

Chafe, W. 1970. *Meaning and the Structure of Language.* Chicago: University of Chicago Press.

Cohen, D. J. 1995. "Highlights: Language- and Domain-independent Automatic Indexing for Abstracting". *Journal of the American Society for Information Science* 46(3), 162-174.

Curzon, L.B. 1982. *A Dictionary of Law*. Plymouth: McDonald and Evans Ltd.

Dagan, I. and Church, K. 1995. "Termight: Identifying and Translating Technical Terminology". *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL ' 95,* Dublin, Ireland, 34-39.

Daille, B. Gaussier E. and Lance Jean-Marc. 1994. "Towards Automatic Extraction of Monolingual and Bilingual Terminology". *Proceedings of the 15th International Conference on Computational Linguistics, COLING ' 94,* Kyoto, Japan, 515-521.

Damerau, Fred J. 1993. "Generating and Evaluating Domain-oriented Multi-word Terms from Texts". *Information Processing and Management* 29 (4), 433-47.

Dubuc, R. and Lauriston, A. 1997. "Terms and Contexts". In *Handbook of Terminology Management.* Volume 1. Wright, S.E. and Budin, G. (eds). Amsterdam/Philadelphia: John Benjamins Publishing Company, 80-87.

Dunlop, D. 1995. "Practical Considerations in the Use of TEI Headers in Large Corpora". In Ide and Veronis *Text Encoding Initiative: Background and Context* . Dordrecht: Kluwer, pp 85-98.

Enguehard, C. and Pantera, L. 1994. "Automatic Natural Acquisition of a Terminology". *Journal of Quantitative Linguistics* 2(1), 27-32.

Fillmore, C. 1968. "The Case for Case". In *Universals in Linguistic theory*. Bach, E. and Harms, R.T (eds.). New York: North Holland, 1-88.

Franzi, K. and Ananiadou S. 1996. "Extracting Nested Collocations". *Proceedings of the 16th International Conference on Computational Linguistics, COLING ' 96,* 41-46.

Gentner, D. and France, I. 1988. "The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs". In *Lexical Ambiguity Resolution.* Small S., Cottrel G., and Tanenhaus M. (eds.). Los Altos, Calif.: Morgan Kaufmann, 343-382.

Graham, G. 1995. *The 3-D Visual Dictionary of Computing*. Foster City CA: IDG Books from MaranGraphics.

Haas, S.W. and He, S. 1993. "Toward the Automatic Identification of Sublanguage Vocabulary". *Information Processing and Management* 29(6), 721-731.

Herbst, R. 1966. *Dictionary of Commercial, Financial and Legal Terms.* Switzerland: Translegal Ltd.

Hornby, A, S. 1974. *Oxford Advanced Learner's Dictionary of Current English*. London: Oxford University Press.

ISO 8879. 1986. *Information Processing: Text and Office systems in Standard Generalised Markup language.* Geneva: International Organisation for Standardization.

Jackson, K.G. and Feinberg, R. 1981. *Dictionary of Electrical Engineering.* London: Butterworth and Co.

Jacquemin, C. and Royaute, J. 1994. "Retrieving Terms and their Variants in a Lexicalised Unification-based Framework". *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94).* Berlin: Springer Verlag, 132-141.

Justeson, John S. and Katz, Slava M. 1995. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Corpus". *Natural Language Engineering* 1 (1), 9-27.

Kageura, K., Yioshioka, M. and Nozue, T. 1998. "Towards a Common Testbed for Corpus-based Computational Terminology". *Proceedings of Computerm'98*, Montreal, Canada, 81-85.

Lauriston, A. 1996. *Automatic Term Recognition: Performance of Linguistic and Statistical Techniques.* PhD thesis, University of Manchester Institute of Science and Technology.

Leech, G. 1993. "100 Million Words of English". *English Today* 9, 9-15.

Leech, G., Garside, R., and Bryant, M. 1994. "CLAWS4: The Tagging of the British National Corpus". In *Proceedings of COLING 94*, Kyoto, Japan, 622-628.

Maynard, D. and Ananiadou, S. 1999. "Identifying Contextual Information for Multi-word Term Extraction". *Proceedings TKE'99: Terminology and Knowledge Engineering*, 212-221. Vienna: TermNet.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Katherine J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography* (special issue), 3(4), 235-312.

Nkwenti-Azeh, B. 1994. "Positional and Combinatorial Characteristics of Terms: Consequences for Corpus-based Terminography". *Terminology* 1(1), 61-97.

Oueslati, R., Frath, P. and Rousselot F. 1996. "Term Identification and Knowledge Extraction". *Proceedings NLP + IA 96*. Moncton, N.B, Canada, 191-196.

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

Sager, J.C., Dungworth, D. and McDonald, P.F. 1980. *English Special Languages*. Wiesbaden: Brandstetter Verlag.

Smadja, F.A. 1991. "Retrieving Collocations from Text: Xtract". *Computational Linguistics* 19(1), 144-177.

Vollnhals, O. 1984. *Elsevier's Dictionary of Personal and Office Computing*. Netherlands: Elsevier Science Publishers B.V.

Wüster, E. 1978. *Einfuhrung in die Allgemeine Terminologielehre und Terminologische Lexicographie*. 2 volumes. Wien: Springer.