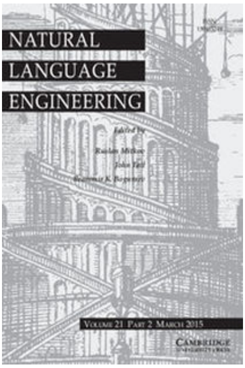

Technical terminology: some linguistic properties and an algorithm for identification in text

John S. Justeson and Slava M. Katz

Natural Language Engineering / Volume 1 / Issue 01 / March 1995, pp 9 - 27
DOI: 10.1017/S1351324900000048, Published online: 12 September 2008

Link to this article: http://journals.cambridge.org/abstract_S1351324900000048

How to cite this article:
John S. Justeson and Slava M. Katz (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1, pp 9-27 doi:10.1017/S1351324900000048

Technical terminology: some linguistic properties and an algorithm for identification in text

JOHN S. JUSTESON
Department of Anthropology
SUNY at Albany
Albany, NY 12222, USA

SLAVA M. KATZ†
IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598, USA

(*Received 12 June 1994; Revised 9 November 1994*)

Abstract

This paper identifies some linguistic properties of technical terminology, and uses them to formulate an algorithm for identifying technical terms in running text. The grammatical properties discussed are preferred phrase structures: technical terms consist mostly of noun phrases containing adjectives, nouns, and occasionally prepositions; rarely do terms contain verbs, adverbs, or conjunctions. The discourse properties are patterns of repetition that distinguish noun phrases that are technical terms, especially those multi-word phrases that constitute a substantial majority of all technical vocabulary, from other types of noun phrase.

The paper presents a terminology identification algorithm that is motivated by these linguistic properties. An implementation of the algorithm is described; it recovers a high proportion of the technical terms in a text, and a high proportion of the recovered strings are valid technical terms. The algorithm proves to be effective regardless of the domain of the text to which it is applied.

This paper outlines some linguistic properties of technical terms that lead to the formulation of a robust, domain-independent algorithm for identifying them automatically in continuous texts. In particular, it addresses multi-word noun phrase terms. Judging from data in dictionaries of technical vocabulary, the majority of technical terms do consist of more than one word; among these, the overwhelming majority are noun phrases, which constitute the vast majority of multi-word terminological units in probably all domains.

† Current address: Weston Language Research, 138 Weston Road, Weston, CT 06883, USA.

The algorithm we present is quite simple conceptually, yet it performs very well: it recovers a high proportion of the valid terms in a text, and the proportion of nonterminological word sequences recovered is low. It has been tested on a variety of text types and domains. The list of candidate terms produced by the algorithm is useful for a variety of tasks in natural language processing, such as text indexing and construction of glossaries for translation.

While 'technical terminology' is the fundamental notion of this paper, this notion has no satisfactory formal definition. It can be intuitively characterized: it generally occurs only in specialized types of discourse, is often specific to subsets of domains, and when it occurs in general types of discourse or in a variety of domains it often has broader or more diverse meanings. In this paper, we treat it as an undefined term for a basic, intuitively recognizable construct.

The first part of the paper discusses some of the properties of technical terms, which provide the linguistic underpinnings of the algorithm: the patterns of use of terminological units in text (section 1), and the grammatical structures of these units (section 2). Section 3 describes an efficient implementation of these ideas. The performance of the algorithm is illustrated in section 4, in a detailed analysis of terms recovered from three recent papers in different domains. Section 5 relates this research to other recent work.

1 Repetition of technical terms

Terminological noun phrases (NPs) differ from other NPs because they are LEXICAL – they are distinctive entities requiring inclusion in the lexicon because their meanings are not unambiguously derivable from the meanings of the words that compose them. An example is *central processing unit*, whose referent is much more specific than the words themselves might suggest. Lexical NPs are subject to a much more restricted range and extent of modifier variation, on repeated references to the entities they designate, than are nonlexical NPs. This applies to variation in the omission of modifiers, in the insertion of modifiers, and in selection among alternative modifiers.[1] This section outlines the differences and their sources.

After an entity is introduced into a discourse via a nonlexical NP, it can be referenced simply by a noun or NP head of that phrase: such omission of modifying words and phrases is semantically neutral, if the meaning of a phrase is compositionally derivable from that of its head and those of its modifiers. In addition, an entity introduced by a nonlexical NP can be and often is reintroduced via a variety of other NPs. In fact, several factors promote variation and inhibit exact

[1] We limit the class of MODIFIERS (as e.g. in Huddleston 1984:233–5) by excluding the general class of DETERMINERS, premodifiers that are applicable to virtually any NP, regardless of its meaning; unless otherwise stated, 'NP' is used in this paper to refer to the core of an NP, excluding its determiners. This is because determiners tend to inform discourse pragmatics rather than lexical semantics, or to serve as quantifiers (see section 2); these functions are generally applicable to all NPs, so the tendency of determiners to be repeated or not is independent of the lexical vs. nonlexical status of the NP they modify.

repetition of these NPs on repeated references. When an entity is introduced with one set of modifiers in a nonlexical NP, these modifiers typically function as means for specifying the entity or type of entity referred to, an aspect of the entity that is in focus, or an orientation to the entity. When this is the function of the NP's modifiers, inclusion of the same modifiers on a subsequent reference to the same entity is, usually, pragmatically anomalous. Accordingly, the typical follow-up reference to the entity is by a definite NP that was either a head of the original NP, or an approximate synonym for either the NP or its head.

Repetition including the modifiers of a nonlexical NP can be appropriate pragmatically, when repetition of the specifying function is motivated; this can occur when the specified attribute is being emphasized, or when the referent of the NP is being distinguished from that of another NP with the same head. The more modifiers are involved, the less likely such possibilities are. Even when repetition of the full NP might be pragmatically appropriate, precise repetition can create a tedious or monotonous effect, the more so the longer the NP and the more recently the repeating phrase was used; some sort of stylistic variation is usual. Exact repetition of nonlexical NPs is expected to occur primarily either when they are widely separated in relatively large texts or else as an accidental effect.

In contrast, omission of modifiers from a lexical NP normally involves reference to a different entity. Lexical NPs – even those with compositional semantics – are much less susceptible to the omission of modifiers.[2] When a lexical NP has been used to refer to an entity, and that entity is subsequently reintroduced after an intervening shift of topic, the reintroduction of reference to it is very likely to involve the use of the full lexical NP, especially when the lexical NP is terminological.

Lexical NPs are also far less susceptible than nonlexical NPs to other types of variation in the use of modifiers. Modifying words and phrases can be inserted *within* a nonlexical NP but not, without a change of referent, within a lexical NP. Similarly, the precise words comprising a nonlexical NP can be varied without a change of referent, but usually not in a lexical NP. Variations either in the choice of some words or in the presence vs. absence of some words in terminological NPs reflect distinct terms, often differentia of a noun or NP head.

In technical text, which is the sole concern of the remainder of this paper, lexical NPs are almost exclusively terminological. Accordingly, the above considerations suggest that variation in the form of an NP in repeated references to the entity it designates is a major textual difference in the uses of terminological vs. nonterminological NPs that can be exploited in building a terminology identification algorithm.

[2] Consider, for example, the terminological unit *word sense*, as used in a paper analyzed in section 4 (Pustejovsky and Boguraev 1993). This is by far the most frequent technical term extracted from the paper. The construct occurs 49 times, in 42 sentences. In 33 instances it occurs in the form *word sense*; in 16 it is used simply as *sense*. The contexts of the reduced form are quite limited. Usually, it occurs when a nearby sentence, or even successive sentences, containing a reference to this construct use the form *word sense(s)*; when the expression occurs more than once in a sentence is it more likely than not that the reduced form will be used, and very often this is along with the full form. It also occurs in the reduced form when it appears in other technical terms (*sense selection*), and when senses of a particular word are being discussed.

This difference applies primarily to *multi-word* terms. All 1-word NPs (nouns) are by definition lexical. The differences between lexical and nonlexical NPs discussed above involve variations in modifier usage. The primary variation involving 1-word NPs (nouns) is noun substitution (e.g. via synonyms, hypernyms, and hyponyms), to which both terminological and nonterminological nouns are subject, and the tendency of nonterminological NPs to avoid exact repetition is least pronounced in the shortest NPs. Accordingly, 1-word terminological NPs are less resistant to and nonterminological NPs less prone to variability in expression than are multi-word NPs of the corresponding types. Accordingly, the repetition of 1-word NPs – i.e. of nouns – does not provide as powerful a contrast between terminological and nonterminological NPs as does the repetition of multi-word NPs. It is primarily for this reason that multi-word NPs are the focus of our terminology identification algorithm, presented in section 3. As it happens, multi-word NPs constitute the majority of all terminological units in technical vocabularies, so this focus helps us to capture the majority of technical terms.

There is also a restricted difference in the susceptibility to repetition of the entities referred to by terminological vs. nonterminological NPs. This difference is specific to the use of novel terminology, i.e. terms that are newly introduced and not yet widely established, or terms that are current only in more advanced or specialized literature than that with which the intended audience can be presumed to be familiar. Whether or not a novel technical term is used for the construct to which it refers, a discursive statement of the construct must be made at or near the first reference to it (or, with suitable indication, in a glossary) in cooperative discourse. If the context of this explanatory statement is the only one in which the construct is referenced, then the use of the term itself does little to advance the exposition. We expect that the use of novel terminology is most often justified by the convenience of its use in further instances, and probably in fact in more than one paragraph.

Established terminological NPs, such as *semantic load* or *binary tree*, may but need not be repeated in a text. But when an entity designated by such an NP is a topic of significant discussion within a text, that entity is almost certainly repeated; as previously discussed, terminological NPs tend to be repeated intact on repeated references to the entities they designate. Accordingly, established, topically significant terminological NPs do tend to be repeated in a text. Nontopical terminological NPs may or may not be repeated; nonrepeated terminological NPs are mostly nontopical.

Some nonterminological NPs behave much like terminological NPs. Such NPs are likely to be repeated, word for word, only as a way of aiding recognition that the reference is to the same construct that was designated earlier. Arguably, however, such uses are effectively coinages of intentionally temporary terms for nonstandard constructs.

2 Structure of technical terms

The previous section describes a pattern of constraints on the *uses* of terminological NPs. It is generally recognized that terminological NPs differ also in *structure*, at least statistically, from nonlexical NPs. This recognition is embodied in the observation that technical jargon makes heavy use of noun compounds. Based both on general considerations and on empirical study of terminology in technical vocabularies, we propose a specific set of structural constraints on terminological NPs that hold in so high a proportion of cases as to be useful for automatic terminology identification.

The structures of technical terms can be illustrated by sampling from available sources for different domains. We selected dictionaries of technical terminology in fiber optics (Weik 1989), medicine (*Blakiston's Gould* 1984), physics and mathematics (Lapedes 1978), and psychology (English and English 1958). From each dictionary, we extracted random samples of 200 technical terms. Noun phrases constitute 185 of the 200 medical and psychological terms, 197 of the mathematical terms, and 198 of the fiber optics terms, i.e. from 92.5% to 99.0% of the terms in each domain. Of the 35 non-NPs among these 800 terms, 32 are adjectives and 3 are verbs. We then extracted additional terms at random until 200 noun phrase terms had been extracted from each dictionary. Out of these 800 NP terms, 564 have more than one word and thus might have words other than nouns. Not one of these 564 terms has either a determiner or an adverb; only 2 have a conjunction (*and*); and just 17 have a preposition (in 15, this preposition is *of*). Thus, 97% of multi-word terminological NPs in these sources consist of nouns and adjectives only, and more than 99% consist only of nouns, adjectives, and the preposition *of*.

This prevalence of noun phrases containing only nouns and adjectives follows from generalizations concerning the typical structures of technical semantic domains. Such domains are organized largely as taxonomies. Ethnolinguistic investigations have established that the terms for taxonomic categories are quite regular in structure (Berlin, Breedlove, and Raven 1973). Those at a level that can be considered a 'basic' or 'generic' level for discourse in the field tend to consist of a single word, or of a single word and a modifier. Furthermore, single words in general vocabulary are rarely appropriate for technical usage in a more specialized meaning because they are thereby inherently ambiguous; when native English forms are used to create new terms, it most often takes at least two words to adequately specify a meaning, and when this is done they usually have just one meaning and are relatively transparent semantically. Often, well established one-word terms are Greek or Latin forms made up of more than one root, e.g. *aerodynamics*; these would often be multi-word terms had they been based on English forms (*air flow*). Daughter nodes of a taxonomy are normally labelled by a term of the same complexity, or by one including one additional modifier; the typical form is the label for the mother node plus a modifier. Furthermore, modifiers applied to the label for one taxon in designating a more specific level are also often applied to other taxons in designating their differentia, leading from hierarchical toward paradigmatic (cross-classificational) structure.

As a result of these trends, 2-word terms are the modal length in systems that have been subject to thorough investigation. We find the same in our dictionary samples. Overall, the average length of NP terms in these samples is 1.91; individual dictionaries provide values ranging from a low of 1.78 for medical terms to a high of 2.08 for fiber optics terms. In the typical distribution of term length, the number of 2-word terms is substantially larger than the number of 1-word terms, with the

Table 1. *Frequencies of NP terms of different lengths in samples from four domains. (Only 3 out of 800 terms have more than 4 words; none has more than 6 words.)*

	Term length (in number of words)			
Dictionary	1	2	3	4 or more
fiber optics	43	109	36	12
medicine	88	80	22	10
physics & mathematics	41	125	29	5
psychology	64	120	12	4

numbers decreasing thereafter as term length increases. This pattern characterizes our samples of terms from the fiber optics, physics/mathematics, and psychology dictionaries (Table 1). Medical terms are anomalous in this respect; while term frequency declines with increasing length for multi-word terms, 1-word terms are more rather than less frequent than 2-word terms. This difference is an effect of an extensive use of compounded Latin and/or Greek roots to form medical terms. The structure of these terms parallels that of English noun compounds; these 'words' amount to *multi*-word terms in Greco-Latinate translation,[3] even corresponding to the same sorts of taxonomic organizations observed above. Conservatively, 39 of the 88 1-word terms are disguised compound forms of this sort; when they are removed, or reanalyzed as being compound forms in fact, the overall qualitative distribution of term lengths agrees with those of the other three dictionaries. The near absence in NP technical terms of prepositions, conjunctions, determiners, and adverbial modifiers of adjectives reflects in part the preference for adjectives and nouns just discussed. In addition, however, the excluded classes may be actively avoided in technical terms, for reasons suggested in the remainder of this section: determiners for discourse reasons; certain conjunctions for semantic reasons; and adverbs, prepositions, and other conjunctions for coding efficiency.

Determiners in the narrowest sense – articles, possessives, and demonstrative pronouns – are used primarily for such discourse functions as indicating the use of an NP in defining an entity, referring to a previously introduced entity, etc. The other determiners consist mostly of a variety of quantifiers and of measure words and phrases. Determiners tend to inform discourse semantics rather than lexical semantics, and to perform very broadly applicable functions such as quantifying, classifying, and measuring. As a result, they readily modify but seldom form a part of a terminological NP. Because the basis for this tendency is semantic, it applies equally to fragments of NPs that are semantically comparable but that do not function syntactically as determiners – e.g. *a lot of* or *a number of.* The syntactic arguments for excluding such fragments from the class of determiners

[3] For example, *melanuria* (*melan-* + *uria*) is literally 'black urine'; *synarthrophysis* (*syn-* + *arthro-* + *physis*) equates with 'together-growing joints'.

follow essentially from their NP+*of* structure (Akmajian and Lehrer 1976), which is not a syntactic constituent of any kind.

The absence in terminological NPs of conjunctions in logically disjunctive uses (a typical use for *or* and a common use for *and*) can be explained on semantic grounds. Conjunctions are useful in compositional, descriptive references, using nonlexical NPs, to entities with multiple functions or parts. However, most entities that require technical terms, and whose general specification involves multiple attributes, are defined in terms of those attributes by logical *con*junction; this has been shown or assumed in numerous studies in componential semantic analysis (see for example the collection by Shepard and Romney 1972). *Dis*junctive uses of grammatical conjunctions, then, are rarely applicable semantically in technical terms.

The avoidance in terminological NPs of logically *con*junctive usage of grammatical conjunctions, as well as of adverbs and prepositions, is understandable largely in terms of coding efficiency. Terminology is coined in order to facilitate communication among people with expertise in an area, providing a compact means for referring to often quite complex constructs; for frequently used constructs, technical terms should, ideally, be relatively short. Study of term coinage among expert users of an artificial command language (Ellis and Hitchcock 1986) indicates that terms do evolve, as expertise increases, toward shorter and shorter designations.

Factors working *against* greater compactness of terminological NPs help to determine the types of words they come to contain. Descriptively explicit NPs have greater semantic transparency than corresponding NPs with some words eliminated; the latter in general are either less transparent or less specific. This effect can be expected to retard the tendency to shorten NPs with repeated use in task-oriented situations; the net effect should be that words with the highest semantic content are the most resistant to loss, and grammatical function words the most susceptible.

Such observations may account for the fact that both prepositions and conjunctions like *and* are infrequent, while noun compounds are among the favored structures, in terminological NPs.[4] Prepositions within NPs provide relational information; eliminating these prepositions, thereby simply compounding the nouns, leaves the relations among the nouns unspecified. Eliminating this information under pressure of terminological compactness, however, obscures the meaning of the phrase less than eliminating one of the nouns, so eliminating prepositions is a typical response to the pressure for compactness. Note that such reduction, by making the relation more ambiguous, will derive a lexical NP whose specific meaning must be given in a lexicon from a nonlexical NP that may have been more interpretable compositionally (e.g. *character string* vs. *string of characters*). Empirically, only about 3% of terminological NPs contain prepositions (when they do, the preposition usually encountered is *of*).

Adjective and noun modifiers can be equally meaningful elements of lexical NPs,

[4] In a section of a computer manual at our disposal, *string of characters* occurs twice, *character string(s)* 37 times. The two instances of *string of characters* occur in definitions of terms in the text, one for *character string*, the other for *character string constant*. This illustrates the fact that the fuller NP expression, in which the NP consists of NP PP, is more compositionally transparent.

but adverbials – as modifiers of modifiers – play a tertiary semantic role; they form a new adjectival modifier of a noun or phrase within an NP. So, although NP terms containing adverbs do occur (e.g. *almost periodic function*), they are quite rare. Their semantic role may be more prominent in *adjective* phrase technical terms, as in *statistically significant*; adjective terms constitute overall 4% of our dictionary samples, and only 2 consist of more than one word.

3 A terminology identification algorithm

Section 1 suggests that exact repetition should discriminate well between terminological and nonterminological NPs. Genuinely large numbers of instances in particular are almost certain to be terminological: excessive repetition is truly anomalous for purely descriptive NPs. Conversely, repetition of *non*terminological NPs at *any* rate is unusual, except in widely spaced occurrences in larger documents; raw frequency should provide a powerful cue to terminological status, without regard to the probability of co-occurrence of the constituent words under assumptions of randomness.

Accordingly, one effective criterion for terminology identification is simple repetition: an NP having a frequency of two or more can be entertained as a likely terminological unit, i.e. as a candidate for inclusion in a list of technical terms from a document. The candidate list that results from the application of such a criterion should consist mainly of terminological units. In fact, this list should include almost all technical terms in the text that are novel and all that are topically prominent.

Structurally, section 2 indicates that terminological NPs are short, rarely more than 4 words long, and that words other than adjectives and nouns are unusual in them. Among other parts of speech, only prepositions occur in as many as 3% of terms; almost always, this is a single preposition between two noun phrases.

3.1 Constraints

The proposed algorithm requires satisfaction of two constraints applied to word strings in text. Strings satisfying the constraints are the intended output of the algorithm. Various parameters that can be used to influence the behavior of the algorithm are introduced in section 3.2.

Frequency: Candidate strings must have frequency 2 or more in the text.

Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A \mid N)^{+} \mid ((A \mid N)^{*}(NP)^{?})(A \mid N)^{*})N$, where

A is an ADJECTIVE, but not a determiner.[5]

[5] Determiners include articles, demonstratives, possessive pronouns, and quantifiers. Some common determiners (after Huddleston 1984:233), occupying three fixed positions relative to one another, are as follows. *Pre-determiners*: all, both; half, one-third, three-quarters, ...; double, twice, three times; such, what(exclamative). *Determiners proper*: the; this, these, that, those; my, our, your; we, us, you; which, what(relative), what(interrogative); a, another, some, any, no, either, neither; each, enough, much, more, less; a few(positive), a little(positive). *Post-determiners*: every; many, several, few(negative), little(negative); one, two, three...; (a) dozen.

N is a LEXICAL NOUN (i.e. not a pronoun).
P is a PREPOSITION.

In words, a candidate term is a multi-word noun phrase; and it either is a string of nouns and/or adjectives, ending in a noun, or it consists of two such strings, separated by a single preposition. Concerning the exclusion of determiners from adjectives admitted in candidate strings, see note above.

There are $(l+2)\cdot 2^{l-3}$ admissible term patterns of length l. Candidate terms of length 2 (with two admissible patterns) and length 3 (with five admissible patterns) are by far the most commonly encountered, and all of the permitted grammatical sequences are attested in strings of this length. The following examples of each permitted pattern are taken from articles analyzed in section 4, drawn from three different domains:

AN:	linear function; lexical ambiguity; mobile phase
NN:	regression coefficients; word sense; surface area
AAN:	Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase
ANN:	cumulative distribution function; lexical ambiguity resolution; accessible surface area
NAN:	mean squared error; domain independent set; silica based packing
NNN:	class probability function; text analysis system; gradient elution chromatography
NPN:	degrees of freedom; [*no example*]; energy of adsorption

3.2 Implementation

Different implementations of the algorithm, or the use of several parameters, provide control over the trade-off between COVERAGE, the percentage of valid technical terms in a text that are extracted by the procedure, and QUALITY, the percentage of extracted candidate strings that are valid technical terms. The algorithm is intended to provide both high coverage of a text's technical terminology and high quality of the candidate terms extracted. Our implementation is based on a preference for coverage over quality, except when a substantial gain in quality can be attained by a minimal sacrifice of coverage. However, the algorithm itself embodies an inherent trade-off, with a definite loss of coverage for the sake of quality, via the frequency constraint. This is an intentional characteristic of the algorithm; using the grammatical constraints alone would result in the recovery of a large number of the unremarkable noun phrases in a text. This design decision leads to a loss of those valid terms that occurred only once. We expect that such terms will typically be the least relevant or important, since they were not so topically significant as to bear repetition. For applications in which this is not the case, the frequency constraint can always be placed under user control, but at the expense of the quality of terms recovered.

The selection of structural constraints also affects the coverage/quality trade-off. In particular, if prepositions are allowed, relatively few of the candidates including them turn out to be valid terms. Quality typically declines, moderately, but coverage is improved: almost all valid terms are recovered. In absolute terms, however, the proportion of valid terms that are not covered when prepositions are not allowed is so low that we usually favor the exclusion of prepositions for the sake of higher quality.

How the grammatical constraints get implemented will strongly affect the coverage/quality trade-off. Since the grammatical class of a given word can be ambiguous, and since automatic parsers and part-of-speech taggers are not entirely reliable, any automatic implementation of the grammatical constraints can only approximate them. Two approaches to this approximation are sentence parsing (or part-of-speech tagging), and part-of-speech filtering.

Under the most straightforward parsing/tagging implementation of the algorithm, a string is accepted as fitting the grammatical constraint if it constitutes a noun phrase of appropriate structure in at least one parse of each of at least two instances of the candidate. Modern parsers and taggers make this approach computationally feasible, but because errors are inevitable in the best of them, this approach will usually fail to recognize some noun phrases that do in fact fit the constraints. The result is that a parsing/tagging approach is very likely to have incomplete coverage, though the quality of the candidate list may be very high.

In our implementation, we adopted a part-of-speech filtering approach that is substantially more efficient computationally than parsing or tagging, and that guarantees greater coverage. For each word in a string being tested, a list of its known parts of speech is retrieved from a lexical database, using morphological analysis as needed; our implementation used OD (Mueller 1990), a lexical database and morphological analyzer containing approximately 100,000 entries. Then, the word is identified as a noun, adjective, or preposition, in that order of preference, if any of these is retrieved as a part of speech for the word; otherwise, the string is rejected. The string is also rejected if it does not end in a noun, if more than one word is identified as a preposition, or if any preposition is not immediately preceded by a noun. Otherwise, the string is accepted as a possible technical term, provided it passes the frequency constraint. To properly implement the frequency constraint, we lemmatize the head nouns of the candidate string.

The coverage provided by this filtering approach is guaranteed to be at least as good as what can be attained by parsing or tagging: it will accept any string that is accepted by a parser that uses the same lexical database for part-of-speech information. Quality should be lower, because the actual part-of-speech sequence may be different from that assigned by this procedure; for example, *fixed* will be identified as an adjective (as is appropriate in, e.g. *fixed disk drives*), even when it is in fact a verb (as presumably in *fixed malfunctioning drives*). Computation time for this procedure is a linear function of sentence length; over an entire text, it is much faster than parsing or tagging.

In one respect it is desirable to limit the increase in coverage of a *pure* part-of-speech filtering approach. Since a part-of-speech filter does not permit contextual constraints to eliminate strings with alternative parses, words used only rarely in an admissible part of speech will provide at most a small proportion of valid candidate terms. For example, *or* is usually a conjunction, but in heraldry is used as a noun meaning 'gold', and in more general text *can* is usually a modal verb but also occurs as a noun (not apt to appear in technical terms in most fields). For the best performance, low risk of a small loss of coverage permits a much higher percentage of valid technical terms among candidates when strings containing such forms are eliminated.

These exclusions are of course tailorable to specific purposes. Some might not be excluded in particular applications. For example, *or* should not be excluded when extracting terms concerning heraldry, and *can* might be removed from the set of exclusions in applications involving packaging and waste management texts. Similarly, additional words or phrases might be excluded. We exclude mostly common verbs interpretable as nouns (e.g. *do, go, see*), and some adjectives that are fairly empty semantically (e.g. *following*) – generally, only the commonest words in English. Exclusions could be much more aggressive, with genuine nouns or adjectives excluded as experience shows that the candidates they yield are seldom valid terms. For example, the noun *show* occurs in terms such as *fashion show* and *trade show*, so that in general *show* should not be excluded. However, if in a domain such as computer manuals the word is used only as a verb, excluding *show* could noticeably improve the quality of the candidate list.

Finally, control over the frequency constraint also enables more flexibility in the quality/coverage trade-off. Raising the minimum frequency of a candidate string improves the quality of the list of candidate terms. Generally, groups of candidate terms of higher frequency have higher quality than groups of candidate terms of lower frequency; and the most frequent candidate strings recovered from technical text are almost always valid technical terms. Particularly for longer documents, minimum frequencies greater than 2 normally will substantially improve quality. However, there is a definite loss of coverage associated with raising the minimum frequency threshold.

One of us has implemented this terminology identification algorithm as a program called TERMS, versions of which have been in use since 1989. The program allows user control over some of the implementation alternatives, e.g. whether or not to allow prepositions in candidate terms. The structural constraint is approximated by part-of-speech filtering. Our algorithm has since been implemented by others, using the alternative approaches discussed above to approximate this constraint. McCord (personal communication, 1990) implemented our algorithm using his parser (McCord 1990); Dagan and Church (1994) implemented an abbreviated version of it using a part-of-speech tagger (Church 1988).[6]

4 Results

TERMS has been extensively tested on a wide variety of materials. It has been successfully used to extract domain-specific multi-word terminology from large text

[6] Dagan and Church mislead their readers by representing part-of-speech tagging as the 'technology' that identifies technical terminology. It is rather a set of noun phrases of a particular type, repeated within a text, that is the crucial clue to terminological status. Part-of-speech tagging is merely one way to implement an approximate assignment of parts of speech to words; it does not identify terms.

collections in a variety of domains – metallurgy, space engineering, and nuclear energy. It is now actively used in several IBM translation centers to assist in identifying technical terminology.

We illustrate the effectiveness of the algorithm using the results obtained by running TERMS. Both coverage and quality can be estimated by analyzing the technical terminology in a text, and the candidate terms recovered from the same text using TERMS. Here we report the results on quality from our analysis of current technical papers in three areas: in order of size, these are statistical pattern classification (Nádas 1995), lexical semantics (Pustejovsky and Boguraev 1993), and liquid chromatography (Cox 1995). Coverage is addressed through a complete analysis of just one of these (Nádas 1995). All three texts were available to us in computer-readable form and were preprocessed to remove nontextual data. Formulas and equations were removed from all three, and blocks of linguistic example sentences were removed from Pustejovsky and Boguraev (1993). The program was run with our preferred settings of the program parameters – the minimum candidate frequency being 2, and with prepositions not being allowed in candidates. The stoplist was not tailored to any or all of the papers; the standard TERMS stoplist was applied to all of them.

Deciding whether a given candidate is a *plausible* technical term is usually very easy, even without the context of the string's occurrence. However, deciding whether a given noun phrase *is* a technical term is necessarily somewhat subjective: many items not found in a terminological dictionary have the 'feel' of terms within the context of a particular paper, and the line between a nonlexical, topical NP and a lexicalized NP can be difficult to draw. This decision is often facilitated by internal clues to the authors' intentions, e.g. by accompanying definitions, but there is no way to completely eliminate a subjective element from these assessments. In our view, then, the best possible judge must be the authors themselves. For the papers analyzed here, these decisions were made by either the paper's author (Nádas, for Nádas 1995; Boguraev, for Pustejovsky and Boguraev 1993) or its editor (Elena Katz, for Cox 1995). In spite of some subjectivity in these judgments, our experience with many users in different domains is that all or almost all of the more frequently occurring candidates are always judged to be valid terms.

4.1 Coverage

Section 2 established the coverage attainable by the algorithm using only the grammatical constraints: more than 97% of the multi-word technical terms in samples from various terminology dictionaries fit the structural constraints imposed by the algorithm (99% if prepositions are allowed in candidates). TERMS does accomplish these high levels of coverage when run with the minimum frequency set to 1, but quality is relatively low for unrepeated term candidates. Coverage decreases and quality improves when the frequency constraint is used as intended.

Coverage was evaluated only on Nádas (1995), the shortest of the three papers, as the tasks involved are too onerous. Nádas identified all multi-word technical terms in his text, yielding 97 distinct terms having a total of 207 occurrences in the text.

Structurally, all were noun phrases except *jointly distributed, standard normal,* and *uniformly approximated*; and of the 94 noun-phrase terms, only *degrees of freedom* and *force of mortality* have a preposition. By token frequency, most terms are recovered by the program, and constitute 146 (71%) of their 207 occurrences; they derive from 36 (37%) of the 97 distinct term types.

In addition, Nádas assessed which terms were a methodological or topical focus of the paper. He identified 39 terms, having 149 instances, as being topical; 36 of these terms, and 146 of the instances, were recovered by the program. Only 3 of the 61 noun phrase terms that were not recovered by the algorithm,[7] because they occurred only once in the text, are topical. Thus, in Nádas (1995), topical terms are almost all repeated, and in fact the valid recovered candidates (which are all repeated) are all topical terms. Candidate terms recovered by the algorithm therefore provide a good indication of the content of a text. The appendix gives the most frequent recovered candidates from the analyzed papers, and well illustrates their topicality.

Nádas's evaluations also suggest that term frequency increases with the degree of 'topicality' of the terms. He not only separated the topical terms from the nontopical, but rated the topicality of topical terms in 3 grades, from topical through more topical to highly topical. The average frequencies of terms in these groups rises from 3.00 for 9 topical terms, to 3.29 for 14 more topical terms, to 4.75 for 16 highly topical terms. The difference between the topical and more topical terms is quite small, but the difference between more topical and highly topical terms is substantial. These differences are not statistically significant; however, we believe the apparent correlation is real, given the generally high topicality of the most frequent terms from each paper (see Appendix). In any event, the difference between the average frequencies of these groups and the 1.00 average for nontopical terms is striking, and is statistically significant at any level; this is the crucial difference, suggested by the discussion of section 1, that lies behind the design of the algorithm.

These results indicate that the valid terms the algorithm fails to recover are, as postulated in section 1, mainly background terms: they reflect assumed knowledge that was used to advance the topic, but were not themselves the focus of discussion. For example, *Heaviside function* and *approximation theory* are drawn upon in the mathematical development of Nádas's stochastic neural net formalism; *speech recognition* and *Viterbi alignment* occur in a brief, one-paragraph section describing an application of his algorithm.

Finally, it should be noted that terminological dictionaries cannot provide a useful and objective measure of the algorithm's coverage. They are not useful for this task, because they contain too few of the terms actually used in any given paper; for example, only 13 of the 97 terms in Nádas (1995) are found in the dictionary of mathematics and physics analyzed in section 2, containing more than 20,000 terms (Nádas identified all 13 as valid terms). Neither is the coverage of these terms an objective estimate. The algorithm's coverage, estimated from the proportion of these 13 dictionary terms recovered by TERMS,

[7] These terms are *EM training, maximum likelihood,* and *ML estimation* (= *maximum likelihood estimation*).

Table 2. *Quality of candidate terms in three recent technical articles. (Entries are the numbers of candidates recovered by* TERMS. *Type frequencies are the numbers of distinct candidates recovered; instance frequencies are the total numbers of occurrences of these types. Different absolute numbers reflect article lengths, about 2300 for Nádas, 6300 for Pustejovsky and Boguraev, and 14,900 words for Cox.)*

Term candidates	Nádas		Pustejovsky & Boguraev		Cox	
recovered	types	instances	types	instances	types	instances
correct	36	146	72	350	195	834
incorrect	3	6	26	59	97	251
% correct	92%	96%	73%	86%	67%	77%

appears higher than suggested above: 8/13 for types, 28/33 for tokens. However, we showed above that coverage in fact relates very directly to the topicality of a particular dictionary's terms in a particular paper: in our case, Lapedes (1978) simply has a higher proportion of the topical than of the nontopical vocabulary that appears in Nádas's paper; in another paper, it could be just the opposite.

4.2 Quality

The quality of the candidates recovered by TERMS was evaluated by Nádas (for Nádas 1995), Boguraev (for Pustejovsky and Boguraev 1993), and Elena Katz (editor of Cox 1995). Table 2 summarizes their analyses. In all three texts, the lowest-frequency candidates are of lower quality than higher-frequency candidates. The most frequent (see Appendix) consist almost exclusively of valid terminological units or topical, term-like phrases.

The overall quality of the recovered candidates declines when prepositions are allowed; this is because NPs with prepositions are common constructs yet are rarely valid terms. If prepositions are allowed in candidates, the three texts together yield only 5 or 6 valid terms from among 58 recovered candidates having prepositions.[8]

Table 2 also shows that overall quality declines as the size of the text increases. This is an inherent characteristic of the algorithm; with sufficient distance, repetition of nonterminological NPs is no longer stylistically obtrusive or inappropriate. The effect can be controlled, largely or completely, by dynamically adjusting the frequency constraint so that larger texts have a higher thresholds. One approach, for example, is to suppose that the density of topical information is limited, so that the total number of topical terms will not exceed some fixed threshold per sentence. This would impose a limit on the number of candidate instances to be extracted, and have the effect of raising the frequency threshold for longer texts.

[8] In Nádas (1995), a single additional candidate is found, which is not a valid term. 20 are found in Pustejovsky and Boguraev (1993), one or perhaps two of which are valid. In Cox (1995), 37 additional candidates are recovered, only 4 of them valid.

The candidate strings that we count as errors are those that were not judged to be valid technical terms. Usually they are noun phrases; our texts yield, for example, *different senses, desired product,* and *important parameter.* Typically, an adjective of rather generic applicability appears as a modifier in these candidates. Another type is illustrated by noun–noun candidate sequences that are not full phrases in context; *scale separation,* for example, comes from phrases such as *larger scale separation* and *production scale separation.* A less common case is the recovery of word strings that are not noun phrases; this is due to part-of-speech filtering, as in the verb–adjective sequence *positing separate* or the noun–verb sequence *solutes move.*[9]

The results reported here are qualitatively comparable to those obtained during more than three years of exploratory work in which we applied TERMS to hundreds of documents from various domains: in the overall quality and coverage of the terms recovered; in the very high quality of the more frequent term candidates and in their topical relevance; in the level of coverage and quality provided by candidates having prepositions; and in the types of errors customarily observed.

5 Related work

The research reported here has several points of contact with other work on multi-word phrases that relate conceptually to a document's content. Most such work deals with indexing (e.g. Salton 1988; Salton, Zhao, and Buckley 1990; Cherry 1990), especially for information retrieval (e.g. Hamill and Zamora 1980; Jones, Gassie, and Radhakrishnan 1990) and for natural language database query systems (e.g. Damerau 1993). These applications are responsible for both the similarities and differences between our approach and those already in the literature.

Domain. We are working on extracting multi-word technical terms from an individual document, particularly those terms that are highly topical and provide a broad characterization of the content of the document. Our approach makes no reference to other documents. Work on indexing attempts to provide a narrower characterization of the content of a document, in order to distinguish it from the content of other documents. As a result, such work usually makes crucial reference to a collection of documents.

Structure. It is a commonplace observation that technical terminology consists mainly of noun phrases and that it is replete with noun–noun compounds. We propose a more specific set of constraints, targeting noun phrases consisting only of nouns, adjectives, and (optionally) prepositions (or specifically *of*). Such constraints do not appear to be widely used in work on multi-word terms and phrases; most of the works cited above, for example, seem to seek noun phrases of unrestricted structure, although those recovered by Salton, Zhao, and Buckley (1990), using

[9] In two cases, part-of-speech filtering led to the recovery of *valid* adjectival technical terms. *Explosion proof* is extracted from Cox (1995) and *standard normal* from Nádas (1995); these are modifiers, not NPs, recovered because *proof* and *normal* have noun as a possible part of speech, and are therefore assigned by part-of-speech filtering as head nouns of an NP.

Church's (1988) tagger, exhibit a similar but perhaps somewhat more restricted set of structural constraints. Bourigault (1992), however, does argue for the use of structural constraints to extract multi-word terms from French texts, and those he uses, though not explicitly described, seem to be quite similar to ours. Bourigault also appears to implement these constraints with a part-of-speech filter and not a parser, largely for processing efficiency.

Frequency. One of the key requirements of indexing for the purpose of retrieval is to find words or phrases that are both highly indicative of document content and highly distinctive within a text collection. In general, the frequency of words and phrases within a document relative to their frequency in a corpus, or relative to the proportion of documents including them, is used to distinguish a document from others in its domain; thus, frequency within a document contributes to the potential value of a phrase as an index term, with a candidate's evaluation being in direct proportion to some increasing function of its frequency. Such an approach is found in all the works cited above. There is no implication in these works that a document's topical phrases are necessarily frequent in it, or that its frequent phrases be topical, although we do report such a correlation (section 4.1); it is simply that terms that do not distinguish documents, however topical they may be, are not useful in such applications. In addition, other factors can override frequency to the degree that, in some systems, even a phrase that occurs but once in a document may be evaluated favorably and selected as an index.

Our goal is different, and admits the less complex frequency criterion of simple repetition: all candidate phrases must meet a frequency threshold of 2 or more. Unlike in all previous research, no further frequency gradient is used; unlike some of it, our approach does not permit selection of a candidate with frequency 1. Cherry (1990:609) also uses repeated noun phrases, along with certain specially tagged phrases, in constructing book indexes.

6 Conclusions

This paper has presented an algorithm which is effective in identifying novel and topical terminology, and yet is remarkably simple. It is effective for specific linguistic reasons. The great majority of technical terms are noun phrases, largely limited to those including adjectives and nouns only. In running text, most topically important technical terms are repeated; those noun phrases that are repeated are very likely to be technical terms. Apart from the least frequent of repeated noun phrases, almost all are technical terms; and the most frequent repeated noun phrases clearly point to the topics discussed.

The performance of the algorithm discussed here can surely be improved by applying a diverse set of special-purpose procedures at various stages of processing. What we have presented in this paper, however, is a solid core for terminology identification systems – a basic algorithm that produces good results on its own, and that can provide a basis for systematic and automatic identification of terminology from a variety of text types and domains.

While this paper has specifically addressed only English terminology and its uses, the linguistic issues that motivate the algorithm are quite general and are, to a great degree, language-independent. If so, the algorithm presented here should be adaptable to other languages. The prospects for French in particular seem promising, for example; though no performance evaluation is available, Bourigault (1992) has already implemented a French terminology identification system using structural constraints very similar to those proposed here for English.

Acknowledgments

Katz thanks Roy Byrd and Mary Neff for introducing him to the problem of automatic terminology identification. Katz also thanks Steve Richardson for contributing the text preprocessor to TERMS, and Len Cantor for providing valuable feedback from professional terminologist users of the program. Bran Boguraev, Geoffrey Cox, and Arthur Nádas gave us access to machine-readable forms of the papers analyzed in section 4. For their work evaluating the output of TERMS, we are grateful to Boguraev, to Elena Katz, and especially to Nádas. Thanks are due to the Department of Artificial Intelligence and the Natural Language Group at IBM Research for supporting Justeson's continuing participation in this research. We also thank Branimir Boguraev, David Lewis, and John Tait, for very useful comments on earlier versions of this paper.

Appendix

This appendix displays the most frequent terminology candidates recovered by TERMS. The number to the left of a candidate is the number of occurrences of that string in the text from which it comes. The total number of candidates presented is as close to 20 as is possible. In Nádas, these are the candidates occurring 3 or more times; in Pustejovsky and Boguraev, 5 or more; and in Cox, 8 or more.

Statistical pattern classification (from Nádas 1995):

15 neural net
13 stochastic neural net
9 em algorithm
9 joint distribution
8 feature vector
6 complete data
6 covariance matrix
5 data model
4 conditional expectation
4 incomplete data
4 linear function
4 training algorithm
3 binary classification
3 class probability
3 classification problem
3 classifying bit
3 class index
3 complete data model
3 continuous function
3 gaussian mixture
3 mean vector
3 random variable
3 standard normal

Lexical semantics (from Pustejovsky and Boguraev 1993):

33 word sense
14 qualia structure
13 lexical knowledge
12 lexical ambiguity
12 lexical item
11 lexical entry
10 ambiguity resolution
10 language processing
10 lexical structure
10 natural language
10 telic role
9 lexical semantics
9 word meaning
8 natural language processing
7 semantic interpretation
6 lexical meaning
6 lexical representation
5 lexical ambiguity resolution
5 selectional restrictions
5 syntactic realization

Liquid chromatography (from Cox 1995):

26 mobile phase
25 surface area
21 packing material
18 preparative separation
16 particle size
15 preparative chromatography
14 column efficiency
14 flow rate
14 production rate
13 loading capacity
11 pore diameter
11 volume overload
9 displacement chromatography
9 displacement effect
9 gradient slope
9 injection volume
9 preparative hplc
9 preparative lc
8 computer simulation
8 explosion proof
8 gradient elution
8 mass overload
8 operating pressure

References

Akmajian, A., and Lehrer, A. (1976) NP-like quantifiers and the problem of determining the head of an NP. *Linguistic Analysis* **2**: 295–313.

Berlin, Brent, Breedlove, Dennis, and Raven, Peter (1973) General principles of classification and nomenclature in folk biology. *American Anthropologist* **75**: 214–42.

Blakiston's Gould Medical Dictionary (1984) 2nd edition. New York: McGraw-Hill Book Co.

Bourigault, Didier (1992) Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of COLING-92*. Nantes, France.

Cherry, Lorinda L. (1990) Index. *UNIX Research System Papers, Tenth Edition*, volume 2, pp. 609–10. Murray Hill, NJ: Computing Science Research Center, AT&T Bell Laboratories.

Church, Kenneth W. (1988) Stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

Cox, Geoffrey B. (1995) Preparative HPLC of biomolecules. To appear in *HPLC: Principles and Methods in Biotechnology* ed. by Elena Katz. Chichester, England: Wiley.

Dagan, Ido, and Church, Ken (1994) Termight: identifying and translating technical terminology. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Stuttgart.

Damerau, Fred J. (1993) Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management* **29**(4): 433–47.

Ellis, Stephen R., and Hitchcock, R. J. (1986) The emergence of Zipf's law: spontaneous encoding optimization by users of a command language. *IEEE Trans. Syst., Man and Cybern.* **16**(3): 423–27.

English, Horace B., and English, Ava Champney (1958) *A Comprehensive Dictionary of Psychological and Psychoanalytical Terms.* New York: Longmans, Green and Co.

Hamill, Karen A., and Zamora, Antonio (1980) The use of titles for automatic document classification. *JASIS* **31**(6): 396–402.

Huddleston, Rodney (1984) *Introduction to the Grammar of English.* Cambridge: Cambridge University Press.

Jones, Leslie P., Gassie, Edward W., and Radhakrishnan, Sridhar (1990) INDEX: the statistical basis for an automatic conceptual phrase-indexing system. *JASIS* **41**(2): 87–97.

Lapedes, Daniel N. (editor-in-chief) (1978) *McGraw-Hill Dictionary of Physics and Mathematics.* New York: McGraw-Hill.

McCord, Michael C. (1990) Slot grammar: a system for simpler construction of practical natural language grammars. In R. Studer (ed.), *Natural Language and Logic: International Scientific Symposium,* Lecture Notes in Computer Science, Berlin: Springer Verlag. pp. 118–45.

Mueller, Patrick (1990) Optimized dictionary (OD) user's manual. Unpublished paper. Bethesda, MD: IBM.

Nádas, Arthur (1995) Binary classification by stochastic neural nets. *IEEE Transactions on Neural Networks* **6**(2): 488–91.

Pustejovsky, James, and Boguraev, Branimir (1993) Lexical knowledge representation and natural language processing. *Artificial Intelligence* **63**: 193–223.

Salton, Gerald (1988) Syntactic approaches to automatic book indexing. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics,* Buffalo, New York.

Salton, Gerald, Zhao, Zhongnan, and Buckley, Chris (1990) A simple syntactic approach for the generation of indexing phrases. Technical Report 90–1137. Department of Computer Science, Cornell University.

Shepard, Roger N., and Romney, A. Kimball (1972) *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences.* 2 volumes. New York: Seminar Press.

Weik, Martin H. (1989) *Fiber Optics Standard Dictionary.* New York: Van Nostrand Reinhold.