A Study of the Linguistic Features of Medical Abstracts for Automated Information Retrieval

Report on Task 3 of the CMU/UMLS Project

Investigators:

Thomas N. Huckin
David A. Evans
Carnegie Mellon University

November 1987

Report No. CMU-LCL-87-7

A Study of the Linguistic Features of Medical Abstracts for Automated Information Retrieval

Thomas N. Huckin
David A. Evans

Carnegie Mellon University-
November 1987
Report No. CMU-LCL-87-7

Task 3 of the Carnegie Mellon University UMLS Effort has involved the analysis of a set of 90 abstracts from biomedical journal articles for linguistic features related to automatic information retrieval. The abstracts were selected as a representative subset of a larger corpus of 150 abstracts deemed 'good' by NLM staff and covering three broad domains: ***basic science, clinical medicine,*** **and** ***health-care delivery.*** **Our investigation involved three** principal phases: (1) a characterization of the lexical/semantic content of the abstracts, by domain (as determined by experts and comparison to standardized terminologies); (2) an investigation of the relation between abstract titles and 'topics' as identified by indexers, experts, and the gross structure of the abstracts themselves; and (3) a study of the rhetorical and linguistic patterns in the abstracts, with the goal of identifying possible recommendations for improvements in abstract writing and for use in designing possible semi-automated indexing systems.

1. Characterization of the Corpus

1.1. Domain experts

Since the abstracts selected by the NLM were taken from three rather differgent domains, it was important for us to identify a domain expert for each subset of abstracts, who could act as a consultant and evaluator of abstract content. Our consultants were:

- Hugh Curtin, M.D., Radiologist at Eye & Ear Hospital, Pittsburgh, and Associate Editor of *Radiology,* who served as consultant on the *clinical-medicine* domain;
- Pat Jamieson, M.D., Neurologist at the University of Pittsburgh, who served as consultant on the *basic-sciences* domain; and
- Michael Hodgson, M.D., Internist in Occupational Medicine at the University of Pittsburgh, who served as consultant on the *health-care delivery* domain.

1.2. Selection of study corpus

Selection was carried out according to the following criteria:

- Equal distribution across the domains of clinical medicine, basic science, and health-care delivery, i.e., 30 abstracts in each of these three domains. Identifying categorial membership proved to be occasionally problematic, as a number of the abstracts did not seem to fall neatly into any particular one of these domains. To make the necessary distinctions, we devised a set of schema-based heuristics serving to operationalize what we perceive to be degrees of *granularity*. In future work, we plan to test our heuristics on medical experts. (Cf. our comments below.)

- Appropriate distribution according to the expertise and interests of our three medical consultants. We wanted each consultant to be knowledgeable of and professionally interested in the subject matter covered in the 30 abstracts assigned to him. Therefore, prior to making the final selection, we had the consultants read through the entire set of 150 abstracts and identify those that lay within their expertise and interest.

- Reasonable length. We wanted to work with abstracts of fairly typical length, rather than with unusually long or short ones. Hence we set upper and lower bounds of 350 words and 70 words, respectively, not including the title.

1.2.1. 'Granularity' as a heuristic in classifying abstracts

In trying to classify the abstracts as pertaining to clinical medicine, basic science, or health-care delivery, we devised the following informal set of heuristics:

- Clinical medicine—Studies involving diagnosis, treatment, or therapy; case studies, especially small ones; reports of manifestations/complications; studies of drugs (rather than of chemicals); studies of associations (rather than of causal mechanisms).

- Basic science—Experimental studies, especially those involving animals rather than people; a concern with causal mechanisms, rather than with associations; studies of natural and synthetic chemicals (rather than of drugs).

- Health-care delivery—Concern with such matters as prevention and safety, compliance and access, training, and environmental factors.

Using these heuristics, we were able to examine the full set of 150 abstracts and sort them into the three domains. We then checked our judgments against the asterisked MeSH headings given in the citation record and against those of our three consultants and found that there was a high level of agreement.

Many abstracts did not fall cleanly into one particular domain, and in these cases we found ourselves applying a general principle of 'granularity'. By this we mean that the basic-level terms associated with basic science appear to be more fine-grained than the basic-level terms associated with clinical medicine, which in turn are more fine-grained

than those associated with health-care delivery. We suspect that granularity is an essential part of expert knowledge in information retrieval. It would be useful to explore the role of granularity further by eliciting basic-level terms from expert MEDLINE users and, if possible, expert indexers.

1.2.2. The resulting corpus

The resulting corpus consisted of 90 author-generated, biomedical-journal article abstracts ranging in length from 71 words to 340 words. Each abstract also had the title appended to it. The corpus can be described in general content terms as follows:

- 30 abstracts in clinical medicine, of professional interest to a clinical radiologist;
- 30 abstracts in basic biomedical science, of professional interest to a research neurologist;
- 30 abstracts in health-care delivery, of professional interest to an internist in occupational medicine.

1.3. Analysis of MeSH-related terminology

Having established our study corpus, we began analyzing its linguistic features. The first question we wanted to address was this: How well does the controlled MeSH vocabulary reflect actual usage by journal article authors? To try to answer this question, we used the 1987 MeSH Vocabulary File Data tapes as provided by NLM and the TEXTSEARCH program developed by F. Masarie and R. Miller.[1] We used a Lisp program to extract all the MeSH headings (element No. 251) with their backwards cross-reference terms (element No. 277) from the file tape. This yielded an augmented MeSH vocabulary of 28,073 unique 'words' (a word is defined here as a string of alphabetic characters delimited by spaces or by nonalphabetic characters) and 82,053 unique terms.[2]

We then ran the set of 90 abstracts (with titles) through Masarie and Miller's TEXTSEARCH program.[3] In each abstract, TEXTSEARCH identified and counted the content words, the MeSH words, the MeSH terms, and the MeSH words appearing in MeSH terms.

---

[1]The TEXTSEARCH program was developed by Masarie and Miller for the NLM under NIH contract NLM RFQ 85–178.

[2]This number of words and terms represents a major increase in size over the 1986 version, which had only 21,782 unique words and 31,675 unique terms (Masarie & Miller, 1987).

[3]This program, which is written in Turbo Pascal and runs on an IBM AT or XT, basically tries to match words and phrases from a source text with the study vocabulary, in simple character-by-character fashion. It looks only for nouns and other 'content words', ignoring prepositions, conjunctions, and other 'function words'. It is tolerant of singular–plural endings but not of other forms of morphological variation. It is tolerant of variation in word order within phrases.

Average no. of content words:	121.3 (range: 64-277)
Average no. of MeSH words:	75.3 (range: 31-180)
Average no. of MeSH terms:	26.9 (range: 4-68)

Figure 1: Analysis of the full set of abstracts

Domain	% total words which are MeSH words	% MeSH words in MeSH terms
Basic science (n = 30)	64.0	48.4
Clinical medicine (n = 30)	62.0	39.3
Health care (n = 30)	60.4	41.9

Figure 2: TEXTSEARCH analysis of journal abstracts by domain

1.3.1. Results of TEXTSEARCH analysis

The results of the TEXTSEARCH analysis on the full set of 90 abstracts are shown in Figures 1 and Figure 2.

Given these figures, it is possible to calculate the percentage of MeSH-related terminology in these abstracts. Dividing the number of MeSH words by the number of content words for each abstract, then averaging over the full set, yields:

Percent of total words which axe MeSH words: 62.1%

In their earlier applications, Masarie and Miller used TEXTSEARCH on hospital charts. Since they used an older and far less developed version of the MeSH vocabulary, we cannot make a direct comparison between the language of charts and the language of abstracts. Nonetheless, it is interesting to note that they found an average of 65.24% of the total content words in charts to be MeSH words. Given the 29% increase in number of unique MeSH words from 1986 to 1987, we suppose that if Masarie and Miller's charts were reanalyzed using the newer MeSH vocabulary, there would be a significant difference between the chart data and the abstract data. In short, hospital charts appear to contain more MeSH-related terminology than do biomedical journal abstracts.

Figure 2 shows the different percentages of MeSH words in these abstracts by domain. Analysis of variance using the SYSTAT program reveals that these differences axe significant at $p < .05$. This suggests that the MeSH vocabulary covers the language of basic science better than it does the language of clinical medicine or health care.

1.3.2. Methodological caveats

As Masarie and Miller have noted[4], the TEXTSEARCH program is not without bugs. The errors it makes are of several types:

[4]Cf. Masarie & Miller, 1987.

- Redundancy in identifying MeSH terms. Multiple-word MeSH terms axe often made up of smaller MeSH terms, so that a term like *positron emission tomography* triggers two **'hits'** ***(positron* and *positron emission tomography).*** **If TEXTSEARCH were concept-**driven rather than string-driven, it would identify *positron emission tomography* as designating a single concept and exclude *positron* from the count.

- False hits due to technical rather than semantic matches. The word *right* is used consistently in these abstracts as an adjective meaning *the opposite of left.* But TEXTSEARCH, which does not contain a true parser, identifies it as a noun (as in *right to die).* In some cases, such false hits did not affect the counts described above; for example, while *PET* is misidentified as referring to household animals, its designation as a MeSH term is not incorrect (since *positron emission tomography* is also a MeSH term). In other cases, though, false hits did affect the accuracy of the counts.

2. Relations of Titles and MeSH Headings

2.1. Representation of main topics

By definition, an abstract discusses those topics that are central to the article it represents. This allows readers to decide, early on, whether or not they are interested in reading the article as a whole. Knowing this, indexers at NLM try to assist the reader by identifying such topics in the MeSH-headings part of the citation record and marking them with an asterisk. To what extent does this identifying of main topics depend on domain knowledge, and to what extent does it depend on linguistic/rhetorical knowledge? That is, are there textual features and patterns in abstracts that might be exploitable for semi-automatic indexing, or are there not?

We attempted to address such questions by examining the relations between the content of titles and the topics actually encountered in the abstracts and their associated articles. We based our identification of 'topic' principally on the starred MeSH headings in the citation record of each abstract, on the assumption that accurate indexing is topic-driven and that expert indexers will reflect the conventions of judgment over the biomedical domain. We systematically searched for textual correlates of the starred MeSH-headings (both major descriptors and subheadings) in our corpus of 90 abstracts. In so doing, we deliberately simulated normal reading/scanning behavior by proceeding in top-down fashion, i.e., by looking first at titles.

2.1.1. Method of Analysis

Our procedure for each citation record was therefore as follows:

1. List all of the starred MeSH headings. In cases where a subheading is starred, list both the subheading and the main heading to which it is attached (on the assumption of an *implicational hierarchy).* In cases where a main heading is starred but a subheading is not, list only the main heading.

		Major Descriptors	Subheadings
yes	(=same)	125	34
(yes)	**(=variant)**	109	75
(yes?)	(=implicit)	42	69
no	(=not represented)	26	26
		302	***204***

Figure 3: Correlates of starred MeSH headings in the titles of 90 biomedical journal articles

2. For each MeSH term listed, search the title for the same term. If there is one, score **it as a *"yes".***

3. If there is no identical match but there is a *synonym*, *near-synonym* or *derivational* ***cognate,* score it as a *"(yes)".***

4. If there is no synonym, near-synonym, or derivational cognate of the MeSH term but there are words that *imply* (for a somewhat knowledgeable reader) the MeSH term, score it as a *"(yes?)".*

5. For all remaining MeSH terms listed, score *"no".* These terms are not *represented* in the title at all.

6. Double-check all questionable ratings with domain experts.

For example, an abstract titled *Growth Hormone Response to Sodium Valproate in Chronic Schizophrenia* lists *schizophrenia/*DIAGNOSIS, somatotropin/*BLOOD,* and *valproate/*DIAGNOSTIC USE as* its starred MeSH headings. Two of these, *schizophrenia* and *valproate,* are identical matches and so receive *yes* scores. *Somatotropin* is a synonym for *growth hormone* and so receives a *(yes)* score. *Blood* is not referred to directly but is implied by the first six words of the title; it receives a *(yes?)* score. *Diagnosis* and *diagnostic use* are not referred to either directly or indirectly with any certainty (our neurologist guessed that this article was not about diagnosis but about *complications* brought on by treatment with sodium valproate); they receive *no* scores.

2.1.2. Results

A compilation of the scores associated with the evaluations of the 506 starred MeSH headings is given in Figure 3. As can be seen from Figure 3, the vast majority of MeSH headings designating main topics in these abstracts (454/506, or 89.7%) have textual correlates in the title alone. If we exclude subheadings, the correlation is even higher (91.4%). This suggests that titles are fairly reliable indicators of main topics in abstracts and articles, and that professional indexers can and perhaps do rely heavily on them. It also suggests that attempts to semi-automate the indexing process should exploit the language of titles.

Since major descriptors greatly outnumber subheadings in the MeSH vocabulary, it comes as no surprise to see that there axe far more exact matches (125 vs. 34) with major descriptors than with subheadings.

184 of the 506 starred MeSH headings have synonymous, near-synonymous, or derivationally-related correlates in the title. In 79.1% of these cases, the alternative term is a MeSH variant. The remaining 38 cases involve non-MeSH variants, specifically 26 key words, most of which are derivational cognates like *thermogram*, *intracerebral*, and *dosimetric*.

111 of the 506 starred MeSH headings are in the *(yes?)* category, meaning that they are implied by certain words or phrases in the title. In many of these cases, a single word or phrase serves as the trigger. In 19 such cases, the trigger is a non-MeSH variant. There are 8 such non-MeSH variants, some of them being used more than once, e.g., *neuropathology* (6 cases), *cecostomy* (4 cases), and *immunodetection* (3 cases).

It seems clear from these facts that the current MeSH Vocabulary is sufficient to account for a majority of the key concepts in biomedical journal article titles. And most of the exceptions are cognate forms whose morphological features could be exploited by an intelligent parser.

2.1.3. Titles as frame-setting devices

There is obviously a close relationship between abstracts and their accompanying titles, a relationship that should be exploited for indexing purposes. Typically, biomedical-journal article titles axe highly informative; indeed, they appear to represent the author's best judgment as to the most important information in his or her article. However, taken by themselves, titles may be somewhat unclear as to what this *most important information* actually is. In a study under the MedSORT-I Project[5], for example, it was noted that a title **such as *Relationship between clinical synovitis and radiological destruction in rheumatoid arthritis*** would be appropriate for an article that was either a criticism of the drawbacks of common diagnostic practices or an etiological study.

As a check on this identification of topics, we interviewed physician users of MEDLINE who had not previously seen our abstracts to determine the prevalence of vagueness in interpreting article content from titles. We presented each of our 90 titles to at least two such users independently and had them try to guess the main point of the abstract. We recorded and compared their responses, then used the abstract itself (with our consultants' help) to determine the correct answer. Our interviews with these consultants indicate that readers can quite accurately set *mental frames* for an abstract on the basis of the title alone. It is likely that, in many cases at least, such frames serve to structure the way in which readers read abstracts. In fact, we found no bias for topics that had not been identified in our principal procedure. These observations have important implications for manual indexing, for the possible semi-automation of indexing, and for writing.

2.2. Representation of main concepts

One assumption in focusing on starred MeSH headings of the 90 citation records was that such headings represent central topics—topics which are most directly and fully addressed in their respective articles—and that one would presume that users of MEDLINE would, on average, derive maximal benefit (i.e., get maximal information per query) if they used

[5]**Cf. Carbonell et al., 1985.**

such query topics for these particular articles. This does not preclude, of course, the possibility that some individual users might be more interested in relatively minor topics for idiosyncratic reasons.

If MEDLINE users are allowed to do free-text searches on abstracts (including titles), most of them would presumably want to find articles where the answer to their particular query is of central concern, i.e., a central topic. By focusing on the starred MeSH headings, therefore, we have been concerned not only with the indexing process but also with information retrieval. Indeed, our research during this time can be seen as addressing a single question of relevance to both indexers and users: *To what extent can MEDLINE users and indexers* ***reliably use abstracts to find the central topics of biomedical journal articles?*** **If there are** linguistic/rhetorical patterns to the presentation of central concepts, then both MEDLINE users and indexers could take advantage of them.

Our first studies focused on main topics, i.e., starred MeSH headings and subheadings taken individually. In so doing, we treated headings with starred subheadings as two distinct topics, with the main heading inheriting a star from the subheading. For example, we analyzed the heading *Alzheimer's disease/*DIAGNOSIS* as representing two distinct topics (Alzheimer's disease and diagnosis), with each being a main topic. On this basis, we found that the 90 articles whose abstracts we have been analyzing contain 506 main topics.

To better emulate information retrieval processes, however, we decided to reanalyze our data by looking not at individual topics but at complete headings. On this basis, *Alzheimer's disease/*DIAGNOSIS* would be treated not as two topics but as a single, more complex one, viz. *the diagnosis of Alzheimer's disease.* This level of analysis, we felt, would allow for a more precise formulation of a topic and would thus be more likely to result in accurate information retrieval. It would also be more likely, we felt, to reflect medically meaningful *concepts.*

2.2.1. Method of analysis

We used basically our earlier procedure, but modified it to reflect our interest in *concepts* rather than *topics* and, in particular, our interest in those concepts not signaled in titles:

1. For each abstract, list the starred MeSH headings. In cases where the starred heading is actually a subheading, assume that the star applies to the heading as well and list both heading and subheading as a single entry.

2. For each of these headings, search the title for the same (identical or nearly identical) term. If there is one, write *"same in title".*

3. For each remaining heading, search the title for a synonymous, near-synonymous, or derivational variant. If there is one, write *"variant in title".*

4. For each remaining heading, search the title for words that imply the heading. If there are such words, write *"implied in title".* If they are only weakly implied (e.g., superordinates, hyponyms, complementaries), state the basis for the implication.

same in title	74
variant in title	107
implied in title	103
not in title	19
Total	*303*

Figure 4: Main-concept MeSH headings and their textual correlates in 90 abstracts (summary)

5. All remaining headings are "*not in title*". For each of these, search the body of the abstract for identical or equivalent terms. Note the location and incidence of these terms, paying special attention to:

   - rhetorical *moves*: statements of purpose, results, or conclusions are common features of abstracts, are known to draw the reader's attention, and are often readily identifiable through surface linguistic markings.
   - *sentence position*: sentences appearing early or late in an abstract (or, in general, in any block of unformated prose) are more prominent than those appearing in the middle.
   - *clause type*: thematic information is typically conveyed through main clauses, not subordinate clauses.

2.2.2. Results.

In our reanalysis, we found that the 90 citation records contain 303 main-concept MeSH headings. Of these, approximately 284, or 93.7%, are signaled in the title. In those few cases where the title does not provide a clear indication of the concept, there is usually a pattern of prominent linguistic/rhetorical clues in the body of the abstract. In fact, only two cases out of the 303 were found to lack adequate representation in either the title or the abstract proper. These findings suggest that with abstracts deemed 'good' by NLM, (a) indexers can rely on them for the indexing of main concepts without having to refer to the article itself, and (b) users of MEDLINE could use them for free-text searching (at least for main concepts), without necessarily searching the article as a whole or using the MeSH headings given in the citation record. Figure 4 is a capsule summary of the main findings.

Of the 19 starred MeSH-headings that do not have textual correlates in the titles of these abstracts, 11 have textual correlates that are prominently featured in the body of the abstract. For example, one of the starred MeSH headings for abstract #26 is *colonic neoplasms/*DIAGNOSIS*. Nowhere in the title (*Intraoperative probe-directed immunodetection using a monoclonal antibody*) is there any reference, direct or indirect, to colonic neoplasms. However, in the first sentence of the body of the abstract, where the purpose of the experiment is stated, explicit reference is made to *colorectal cancer*. Furthermore, this term appears at the end of the sentence and is part of the main clause. And a second reference to *colorectal cancer* is made in the concluding sentence of the abstract. In short, though this particular MeSH heading is not represented in the title, it is represented, and

good rhetorical/linguistic clues in abstract	11
fairly good rhetorical/linguistic clues in abstract	6
weak rhetorical/linguistic clues in abstract	2
Total	*19*

Figure 5: Textual correlates of the main-concept MeSH headings not represented in titles

prominently so, in the body of the abstract. A full recapitulation of these 19 outliers is given in Figure 5.

2.2.3. Conclusion.

According to our analysis, which was aided by expert judgment from our medical consultants, 284 of the 303 main-concept MeSH headings in our corpus of 90 abstracts have clear textual signals in the titles of these abstracts. Of the 19 remaining cases, 17 have clear textual signals in the body of these abstracts. This means that, in this set of abstracts at least, 99.3% of those MeSH headings deemed to be of central importance could be determined by looking at rhetorically-prominent positions (title, lead sentence, purpose statements, etc.) in the abstract alone, without having to consult the article itself. This finding supports the possibility of semi-automated indexing. At the same time, this finding has implications for information retrieval, for it suggests that users of MEDLINE could indeed engage in the free-text searching of abstracts (including titles) to find the main concepts of particular biomedical journal articles, at least in those cases where the abstract is a good representation of the article.

3. Rhetorical and Linguistic Patterns in the Abstracts

A second question we addressed involved the relationship between bibliographic retrieval questions and the corresponding rhetorical and linguistic features of abstracts. Since much of our early work concentrated on titles—titles being the single most important part of abstracts—we deliberately sought to broaden our investigations by concentrating on the body of the abstract, ignoring the title. Specifically, we

- compiled a set of bibliographic retrieval questions that could be answered by important concepts in the selected abstracts;
- determined the key words and phrases in those questions;
- determined the parts of the abstract which correspond to the key words and phrases;
- completed a linguistic/rhetorical analysis of those parts; and
- devised preliminary guidelines for authors of abstracts.

3.1. Compilation of bibliographic retrieval questions.

To gather a set of appropriate retrieval questions, we asked each of our medical consultants (Curtin, Jamieson, Hodgson) to read through each of the 30 abstracts in his domain (clinical medicine, basic sciences, or health care) and come up with at least two questions that axe addressed in the abstract (or more properly, that the abstract promises axe addressed in the article). The consultants had only the full abstract (i.e., title and body) to look at; they were not allowed to see the MeSH headings. They were under no time pressure. All sessions were tape-recorded and later transcribed.

To double-check the *authenticity* of the questions we had gathered, we had a medical librarian[6] examine a subset of them. She said that most of them were *authentic* in the sense that they were similar to those she routinely gets from physicians. A few, however, struck her as *textbook-type* questions, broad questions that medical students might ask but not trained specialists. An example of this latter type is *What are the metabolic changes* ***that might be seen in patients with Huntington's disease?*** **At first we were inclined to** eliminate these questions, but since MEDLINE is used by a wide variety of people, including nonspecialists, we decided to keep them.

We gathered an average of three bibliographical retrieval questions for each of our 90 abstracts, or about 270 in all. As we had anticipated, these queries generally conform to the starred MeSH headings given in the citation record—an interesting finding since, as mentioned above, these consultants did not have access to these MeSH headings. The questions almost always contain technical terms that can be used directly for information retrieval, but they also contain non-technical terms that are somewhat problematic. (We discuss this further below.)

3.2. Determination of key words and phrases in these questions.

Next, we examined two of the questions for each abstract, picking out those words and phrases that represent the most important concepts. For example, from the question ***What are the complication rates and the long-term outcome of tube cecostomies?*** **we selected** ***complication rates, long-term outcome,*** **and** ***tube cecostomies*** **as the most important** concepts. In most cases this was a fairly straightforward procedure. In many other cases, however, it was more difficult, particularly those cases involving topic-focus constructions where the focus is on a non-technical qualifier or predicate. For example, a question like ***How frequently can partially or completely amputated patients maintain some form of biped ambulation?*** **has two semantically-rich noun phrases** ***(partially or completely amputated patients*** and *biped ambulation)* linked by a more general predicate *(maintain).* For efficient information retrieval, the concepts encoded in these two noun phrases must be used. But they do not represent the primary focus of the user's question. Rather, the primary focus, as signaled by the fronting of the ^-phrase, is on the time adverbial *(how frequently).* Furthermore, this adverbial is most directly linked (semantically and syntactically and, by extension, pragmatically) to *maintain.* Further linguistic analysis would show that there is a hierarchy *of focus-hood* among the four major constituents of this sentence which looks

[6] **The medical librarian was Lisa Jamnback, who also works for Randy Miller and Chip Masarie.**

like this:

> ***how frequently < maintain < biped ambulation < partially or completely amputated patients***

As can be seen, there is an inverse relationship in this case between semantic richness and degree of focus. For accurate information retrieval, one would want to use not only the semantically-rich concepts but the focused ones as well. Otherwise, one would not be responding directly to the user's expressed interest. But which focused terms should be used? *How frequently? Maintain?* Both? Should they be treated separately, or joined together? Should *maintain* be treated as an isolated concept, or should it be joined to ***biped ambulation?***

Issues like this came up fairly often, and there did not seem to be a principled way to resolve them. We felt compelled to recognize the questioner's interest and mark *how frequently,* for example, as a *key phrase,* yet we knew that such a phrase or any of its imaginable synonyms would not be very useful for text word searching. (In fact, the abstract to which this question pertains alludes to frequency only via numbers.) This points up a general dilemma for traditional information retrieval: The technical concepts embedded in a user's query usually have restricted lexical realizations and fairly well-defined referents, and so axe more likely to yield positive results in text word searching than non-technical concepts; however, it is often the non-technical concepts that constitute the actual focus of a user's query. The technical concepts typically represent topics, while the non-technical concepts represent relationships between those topics. What is needed to capture both the topics and the relationships, we think, is a more pragmatic approach to the information contained in questions and abstracts, allowing a more natural question-and-answer fit. Such an approach would simultaneously exploit biomedical knowledge-structures and rhetorically-defined text structures. (See discussion below.)

3.3. Determination of the parts of the abstract corresponding to these key words and phrases.

Having isolated the key words and phrases in each of the 180 bibliographical retrieval questions, we then analyzed each of the corresponding abstracts (excluding titles) for language related to those key words and phrases. A total of 932 words, phrases, and sentences were so extracted. Repeated forms and simple anaphoric forms (e.g., pro-forms) were ignored. Syntactically, the items range from single words to full sentences and include some discontinuities.

3.4. Analysis of language in abstracts

3.4.1. MeSH language

To determine the amount of MeSH language that is used in these parts of the abstracts, we checked each of the 932 items against the 1987 MeSH Vocabulary Files. In so doing, we ignored stop words but required that all content words in a phrase be MeSH terms for that

Domain	No. of phrases or words in abstract that correspond to queries (A)	No. of A phrases or words that consist of MeSH language	No. of A phrases or words that consist of non-MeSH language	Percent MeSH language
Basic science	323	204	119	63%
Clinical medicine	260	141	119	54%
Health care	349	196	153	56%
Totals	*932*	*541*	*391*	*58% avg.*

It should be noted that there were differences across the three domains covered in our corpus, with the basic science abstracts containing significantly more MeSH language than the clinical medicine or health care abstracts. This difference is consistent with findings described in earlier reports, using Masarie and Miller's TEXTSEARCH program.

Figure 6: MeSH/non-MeSH analysis of parts of abstracts (excluding titles) corresponding to 180 bibliographic retrieval questions

phrase as a whole to be considered a MeSH *phrase*. For example, we counted *agenesis of the corpus callosum* as a non-MeSH phrase because even though *corpus callosum* is a MeSH term, *agenesis* is not. Presumably, someone who specificies the full term in his query, as this consultant did, wants information about that full term, not just the MeSH part of it.

The results of this analysis are given in Figure 6. It can be seen that only 58% of the key words and phrases in our 90 abstracts consist of MeSH language. This means that if a person used only MeSH words to scan abstracts for answers to these 180 questions, he or she would be losing a great deal of specificity. Even if one happened to use all the right MeSH words and refrained from *and*-ing them to any wrong ones, he or she would most certainly generate far too many false-positive retrievals.

3.4.2. Paraphrasing

Next, we analyzed the corpus to see if authors used keywords consistently or used paraphrases. Given the demanding nature of text word searching, requiring exact string matching, the use of paraphrases would increase the likelihood of retrieval. We again worked through the full set of bibliographic retrieval questions, noting all paraphrases (i.e., excluding *pro*-forms and inflectional variation). This gave us a raw count. To get a true picture of how often authors used paraphrases, however, we had to first calculate the number of opportunities they had. An author cannot use a paraphrase when first referring to a concept, but only on subsequent mentions. So we calculated the number of paraphrasing opportunities by subtracting the number of key words and phrases identified in the queries from the total number of corresponding words or phrases in the abstract. We then determined the actual frequency of paraphrasing by dividing the number of paraphrases by the number of paraphrasing opportunities. Results are given in Figure 7.

Domain	No. of paraphrases	No. of paraphrase opportunities	Paraphrase percentage
Basic science	34	323	11%
Clinical medicine	12	229	5%
Healthcare	38	292	13%
Totals	***84***	***844***	***10%***

Figure 7: Paraphrasing in the parts of abstracts corresponding to key words and phrases in 180 bibliographic retrieval questions

These figures show that there is relatively little use of paraphrasing in these abstracts. A good example of paraphrasing would be the following (Abstract #14):

> L-[18F] fluorodopa was administered in trace amounts intravenously to healthy control subjects and to patients with Parkinson's disease. Striatal uptake of radioactivity was measured using positron emission tomography. The capacity of the striatum to retain tracer was severely impaired in patients compared to controls. This may reflect a reduction of striatal dopamine storage in Parkinson's disease. Patients showing the *on/off* phenomenon had an even greater decrease of striatal storage capacity.

After reading this abstract, our neurologist posed the following bibliographic retrieval **question:** ***What are the changes in Parkinson's patients in terms of the binding of L-fluorodopa ?*** The key words and phrases in this question, in our view, are *changes, Parkinson's, binding,* and *L-fluorodopa.* The abstract itself does not use the words *changes* and *binding,* but it does use a variety of alternatives: *impaired, reduction,* and *decrease* for the former; *uptake, capacity to retain,* and *storage* for the latter. This enhances the possibility that someone using free-text searching would hit upon the right set of words.

In general, though, with only 10% of paraphrasing opportunities actually being taken advantage of, there is ample room for an increase of paraphrasing.

3.4.3. Rhetorical patterns

We have noticed a number of linguistic and rhetorical patterns in our set of abstracts that we feel could be exploited for both automatic indexing and information retrieval. Indeed, we feel that such patterns must be exploited if we are to resolve the problem of representing relationships and not just topics.

The abstracts in this study typically consist of three parts: a single statement of methodology, one or more statements of results, and a statement of conclusions or recommendations. All three parts are important to indexing and retrieval. Our bibliographic retrieval questions, however, seldom addressed anything other than the results and conclusions. Most—but not all—of our abstracts reflect this bias, in that they give major attention to results and conclusions and only minor attention to methodology. We think that all ab-

stracts reporting on a particular study (as opposed to review-type abstracts, for example) should be written so as to conform to this pattern.

The three text-parts mentioned above have distinct linguistic features that can be exploited for indexing and retrieval:

Methodology statements. 63% (57/90) of the abstracts in this study have statements of methodology. Review-type abstracts were the major exception. All but six of the abstracts with methodology statements use verbs in the past tense; the remaining six use verbs in the present perfect tense. The most common verbs found in these statements are *used* (13 times), *measured* (12), *evaluated* (9), and *performed* (8). Almost all of the other verbs are semantically related *(assessed, compared, administered,* etc.).

Statements of results. 88% (79/90) of the abstracts have statements of results. These statements almost always occur with past-tense verbs, the only exceptions being all 14 of the reviews and four of the case studies, which use either the present tense or present perfect tense. The most common lexical items found in these results statements are *showed* (27 times), *significantly* (19), and *significant* (18). Other commonly occurring lexical items in these statements are semantically related *(demonstrated, found, was observed,* etc.). Using linguistic features such as these, results statements can be readily identified. However, they cannot usually be *interpreted* by themselves, out of context. They typically use ellipsis and various anaphoric devices that depend for interpretation on previously mentioned lexical noun phrases.

Statements of conclusions. 78% (70/90) of the abstracts have statements of conclusions or recommendations. These statements almost always occur with present-tense verbs (65/70). Many begin with an explicit signal phrase such as *This study shows* or *We conclude that* (26/70). Many contain single-word hedges such as *suggests, appears,* and *may* (35/70) and many others contain more elaborate hedges. Statements of conclusions typically restate and add information to the topic given in the title. Unlike results statements, they can usually be interpreted out of context.

These linguistic features combine to form a virtual text-schema for biomedical journal article abstracts. The following abstract is an example:

> Carbamazepine lowering effect on CSF somatostatin-like immunoreactivity in temporal lobe epileptics (title)
>
> The effect of carbamazepine treatment on CSF-somatostatin-like immunoreactivity (SLI) in patients suffering from temporal lobe epilepsy was investigated. A baseline lumbar puncture was performed on 12 patients and 10 normal volunteers. A second tap was repeated only in patients when they were on peak of carbamazepine concentration for 10 days. Levels of CSF-SLI were measured by RIA. No significant differences were found in CSF-SLI basal concentrations between epileptics and controls, whereas a significant decrease (p less than .0002 Duncan's multiple range test) of CSF peptide levels occurred in 9 of 12

patients under medication. Although the neural mechanism through which carbamazepine lowers CSF-SLI is still unknown, the results of the present study suggest that the reported effect might be part of the apparatus by which carbamazepine exerts its anticonvulsant action.[7]

3.4.4. Implications of rhetorical patterning

For indexing. To the extent that abstracts conform to such a text-schema, indexers should find it easier to locate key information for indexing. If frame-based knowledge representations and parsers are used, even semi-automated indexing becomes a possibility. Using the linguistic features described above, a text-analyzing (TA) program could identify first the conclusion and/or title. An indexing program could then analyze these parts for disease names or other major frame headings and call up the appropriate knowledge frames. Next, the TA program would identify the results section, and the indexing program would analyze it according to appropriate slots in the knowledge frames. Names for major frame headings and key related terms could then be converted into MeSH headings and subheadings. A weighting scheme giving priority to titles and conclusions might account for the starred headings. Headings related to methodology could be handled in the same way, or, perhaps better, could be determined in a totally different, non-computational way, by having authors simply fill out a check-list before submitting their articles for publication.

For information retrieval. NLM's current mode of text word searching does not appear to be capable of exploiting this rhetorical patterning. Text-word searching on MEDLINE is based entirely on exact string matching. Therefore, improvements can be made in only two ways: (1) by getting more authors to include more key information in the abstract, and (2), as suggested above, by having them refer to this information with varied terminology.

If frame-based knowledge representations and parsers were used, however, retrieval could be enhanced significantly. After using MeSH headings to narrow their search, users could employ a second-stage search program (exploiting the TA and indexing programs) to zero in on just those abstracts where the target information appears in conclusions and results statements.

3.5. Some tentative abstracting guidelines for authors.

On the basis of our work so fax, we believe there are several changes editors could impose on the way biomedical journal article abstracts are written. These changes would (a) facilitate information retrieval, (b) facilitate indexing, (c) be easy to implement, and yet (d) not detract from the traditional role of the abstract as a previewing/overviewing device.

1. Write the abstract so that results and conclusions are emphasized and purpose and methodology are minimized.

[7] **Abstract #17 (UI 87044958)**

2. Make sure the abstract has a clear statement of conclusions, emphasizing the most significant point(s) of the article.

3. Use past tense verbs to describe specific results of the study. Use present tense verbs to state general conclusions.

4. Within reason, try to refer to important concepts with varied terminology. Avoid exact repetition.

Additionally, the NLM should consider implementing a system in which authors are asked to fill out a checklist of methodology-related headings before submitting a manuscript for publication.

References

Carbonell et al., 1985 J.G. Carbonell, D.A. Evans, D.S. Scott, and R.H. Thomason, ***Final Report on the Automated Classification and Retrieval Project (MedSORT-I).*** Technical Report No. CMU-LCL-85-1, Laboratory for Computational Linguistics, Carnegie Mellon University, 1985.

Masarie & **Miller, 1987** R.A. Miller and F.E. Masarie, Medical subject headings and medical terminology: An analysis of terminology used in hospital charts, *Bulletin of* ***the Medical Library Association,*** **Volume 75, No. 2, 1987.**