Procedures for Mapping Vocabularies from Non-Professional Discourse A Case Study: "Consumer Medical Vocabulary"

Tony Tse
College of Information Studies, University of Maryland, College Park, MD. 20742
Email: tsetony@oriole.umd.edu

Dagobert Soergel
College of Information Studies, University of Maryland, College Park, MD. 20742
Email: dsoergel@umd.edu

Non-professionals are often unable to access or comprehend information in specialized domains because of technical terminology. Understanding domain-specific vocabulary commonly used by laypersons will help bridge this gap. This paper describes in detail a methodology for collecting and extracting linguistic forms from documents authored by healthcare consumers. The forms were mapped to concepts and the resulting terms were analyzed. Although the case study uses examples from a "consumer medical vocabulary," the procedures are applicable to investigating non-professional vocabularies in other domains.

Introduction

Increasingly, members of the public, called "healthcare consumers" ("consumers"), seek information about medical topics online. For example, a Harris Poll® nationwide survey (Taylor, 2002) found that 80% of 707 adult respondents who use online resources seek health information on the Internet:

> …the Internet continues to be used by huge, and growing, numbers of the public interested in getting information about particular diseases or treatments or about staying healthy.

In response to consumer needs, healthcare organizations provide information through publications, the mass media, and the Web. But, even though more data are available on demand than ever before, the gap in consumers' knowledge of health topics still thwarts their attempts to find information or make sense of what information they do receive (Stavri, 2001). Terminology mismatch is one barrier:

> …little systematic effort has been made to develop a complete consumer-oriented medical terminology… [even though] patients often present quite a different perspective from that of healthcare professionals on what is important in the health care encounter [and] …the vocabulary used by laypeople can differ significantly from that of healthcare professionals… (Rose et al., 2001, p. 328).

Exploring how consumers refer to medical concepts by collecting and generating a "consumer medical vocabulary" (CMV) and comparing it with a "professional medical vocabulary" (PMV) will support improvements in the design of consumer health information systems that are both accessible and understandable.

The purpose of this study is two-fold:

(1) to investigate methods for identifying lay terms and
(2) to arrive at an overall characterization of a CMV and compare the ways non-professionals and professionals express medical concepts.

Our ultimate goal is to gain insight into how consumers understand medical concepts and to bridge communication between laypersons and domain specialists.

Background

Communication among domain specialists is facilitated by terminology because terms, which consist of surface structures (linguistic forms) and associated meanings (concepts), are derived from shared viewpoints and common knowledge. Non-specialists may not fully share in this consensus; their understanding of medical concepts and use of linguistic forms is influenced by a range of knowledge and experiences perceived through socioeconomic, cultural, and other perspectives. Thus, miscommunication or no communication may result from several kinds of term "mismatches": not knowing a form, not understanding a concept, or differences in usage or interpretation (Table 1).

Table 1. Effects of term familiarity (knowledge of forms and understanding of concepts) on communication

Form	**Concept**	
	Understood	**Not understood**
Known	Communication	Miscommunication
Not Known	Miscommunication / No communication	No communication

A layperson may know a form but not understand the technical concept. For example, *asthma* "may cause great anxiety to parents whose experience of that disease is of sudden, unexpected death. If, however, there is familiarity with the condition … there may be less alarm" (Wood, 1991, p. 536). Conversely, when a concept is understood, but the technical form is not known, a layperson may "fill in the blank" by using less precise, general language expressions, such as a euphemism (*go to the bathroom* for *urination*), definition (*slowed heartbeat* for *bradycardia*), or even another technical form (*depressed* for *sad*).

Hence, the goal of a consumer medical vocabulary (CMV) is to bridge technical and lay terminologies. However, "In order to be effective, a standardized consumer health vocabulary will need to consist of 'normal standard ways of expressing things' (in everyday life), and will also likely need to contain 'informal' terminology" (Lewis, Brennan, McCray, Tuttle, & Bachman, 2001, p. 1530). Systematically collecting "authentic" forms used by consumers and mapping them to appropriate medical concepts poses a methodological challenge.[1]

Previous studies in different domains have used different techniques for collecting and mapping lay terms. For example, Haas and Hert (2002) used several methods to collect the language of "communities of users" in the statistical data domain for building the LABSTAT crosswalk. These included:

(1) Examples of terminology mismatches from interviews with specialists;
(2) Words from search log queries and user email; and
(3) Language used by journalists writing about the domain for the public, the most fruitful approach that resulted in many "more casual, or slangy [words] other than the 'official' BLS terminology…" (p. 44)

In the medical domain, Patrick, Monga, Sievert, Hall, and Longo (2001) extracted diabetes-related terms from email and consumer health Web search logs, using substring probes such as *diab* and *gluco*. Two of the authors matched the lay terms with terms extracted from physicians' progress notes, organizing the terms into semantic neighborhoods.

Our study adapted a corpus-based terminographical approach for exploring and collecting terms used by nonprofessionals. The procedures are first described, followed by the results of a case study on CMV.

[1] A mediator medical vocabulary (MMV) was also collected and analyzed in the original study (Tse, 2003), based on the hypothesis that the MMV might serve as a "natural" bridge between CMV and PMV. Terms were extracted from health-related documents authored by professional information intermediaries (e.g., health communicators, journalists, and librarians). However, MMV will not be covered in this paper.

Corpus Generation and Vocabulary Term Collection Procedures

We describe a semi-automated procedure for identifying and processing non-professional terms from a specific domain, using a corpus-based approach in three steps:

Step 1. Collect documents from Web-based discussion forums on medical topics and extract health-related forms.

Step 2. Process the extracted forms (e.g., spell checking and expanding abbreviations, acronyms, and clippings), normalize, and map to concepts to the U.S. National Library of Medicine's (NLM) Unified Medical Language System® (UMLS®) Metathesaurus® (UMLS, 2002), using automated and manual mechanisms.

Step 3. Analyze the forms and mapped-to UMLS concepts in aggregate and individually. Compare consumer forms and concepts to a "professional medical vocabulary" (PMV), using one-sided overlap analyses at the concept and form levels.

In this paper, terms are represented as form-concept pairs (i.e., <form, concept>). Forms are strings representing concepts. As concepts are abstractions, they are expressed using UMLS "concept unique identifiers" or CUIs with Metathesaurus preferred terms in parentheses:

<allergy, C0020517 (Hypersensitivity)>
<antacid, C0003138 (Antacids)>
<heart attack, C0027051 (Myocardial Infarction)>

Corpus Generation

Documents for the consumer corpus were selected based on criteria such as authorship and extensiveness of topics. Online discussion forum postings were used to represent consumer utterances.

Document Sources. Postings from online health forums between 1999-2000 served as sources of consumer documents. The postings, written by and intended for consumers, reflect non-professional discourse health. For reasons of privacy, we collected and "deidentified" archival postings, which were uniquely identified by forum name and subject heading. Because authorship by actual consumers could not be verified, forum moderators and self-identified healthcare professionals were excluded.

Document Collection. Documents were collected using guidelines developed by the researchers to ensure a broad representation of perspectives, such as including different disease topics and limiting the number of documents by a single author from a particular source. We continuously monitored the collection process to control for document duplication and to ensure that documents came from a

Doc ID	Source ID	Forum Name	Document Title
CMV-15.1	MMD	Prostate Cancer	What a strange battle we fight
CMV-16.1	MMD	Heart Disease	trying to learn
CMV-16.2	MMD	Heart Disease	trying to learn

Figure 1. Sample consumer corpus records. The shaded record identifies document CMV-15, post 1, from a prostate cancer forum on the MMD Web site.

variety of sources and represented different disease topics. A unique identifier was assigned to each document and document attributes were recorded, as seen in Figure 1. The goal was to collect a corpus that would yield a sufficient number of terms used by consumers to allow terminological patterns to emerge.

Vocabulary Term Collection

Because the terminological properties of consumer discourse on medical topics are not well characterized, lay forms were extracted manually. We subsequently mapped the extracted forms to medical concepts from the UMLS, based on context.

Form Extraction. Fourteen extractors, mostly college students with a general interest in health matters, recruited and paid for this purpose, identified terms — a word or multiple words representing a medical concept. Although general guidelines were provided, extractors were instructed to highlight terms, using their personal experience, knowledge, and judgment. Each document was independently reviewed by two extractors to assure complete detection of all health-related terms and also to derive some measure of agreement. The marked-up documents with extractor-identified lay medical forms (Figure 2) were returned to the researchers for quality review, coding, and entry into an electronic format. The forms were classified by the degree of extractor agreement:

- Complete Overlap: Identical strings highlighted by both extractors, appearing at the same location (e.g., "clinic" was selected by both extractors in Figure 2)
- Partial Overlap: Different strings sharing a common substring highlighted by extractors (e.g., 'urologist specialist" was highlighted as a single form by extractor 2, but as separate forms—"urologist" and "specialist"—by extractor 1)
- No Overlap: Strings highlighted by only one extractor (e.g., "Achy" was identified by extractor 2 only)

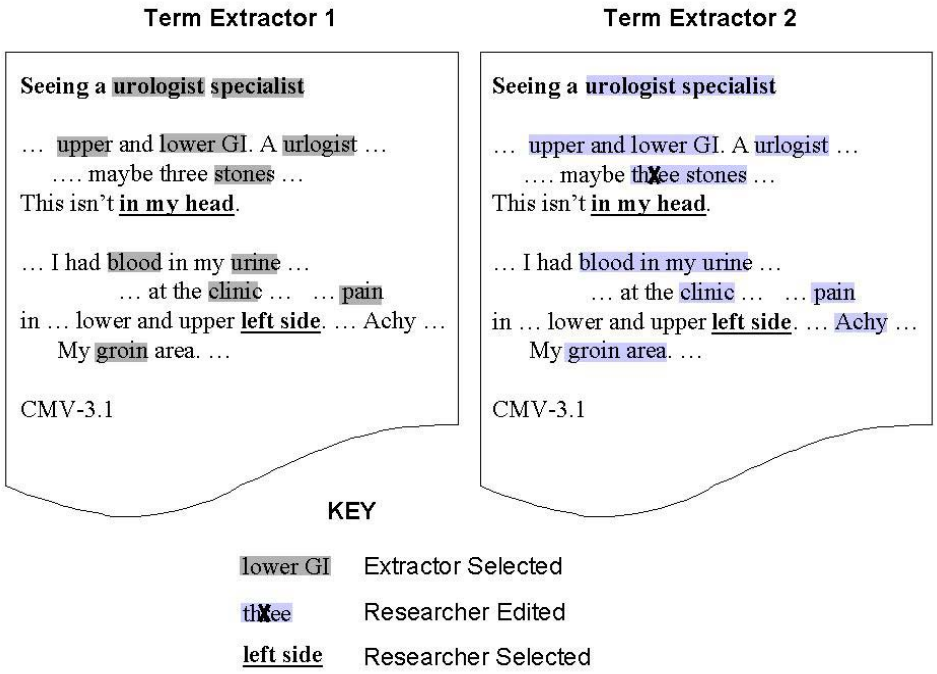

Figure 2. Sample marked-up documents (CMV-3.1) from two extractors

Doc ID	Loc ID	Extracted Form	#1	#2	Code	Annotation
CMV-3.1	1.1	urologist	1			
CMV-3.1	1.2	urologist specialist		2	Cl	urologist; urology specialist
CMV-3.1	1.1	specialist	1			
CMV-3.1	2.1	upper	1		CC	upper GI
CMV-3.1	2.1	lower GI	1			
CMV-3.1	2.2	upper and lower GI		2		
CMV-3.1	3.0	urlogist	1	2	Sp	urologist
CMV-3.1	4.0	stones	1	2	Tr	kidney stones
CMV-3.1	5.9	in my head			M	imagined

Figure 3. Sample records of extracted forms and information from extractions in Figure 2. Fields (left to right): document identifier, location identifier, extracted form, first extractor identifier, second extractor identifier, annotation code (Cl = clarification; CC = coordinated construct; Sp = spelling; Tr = truncation/ellipsis; M = metaphorical), and annotation text.

Form Processing. The researchers recorded extracted forms and related contextual information useful for interpreting meaning, such as document source and location (Figure 3). All extracted forms were entered into electronic files as they appeared in the original documents and in the same order as the text would be read — top to bottom and left to right. Forms extracted from table and figure headings and graphical elements containing labels were inserted into the sequence, preserving the flow of the text as much as possible. Extracted forms in all overlap categories were entered, with completely overlapping forms entered only once.

For example, *urologist*, *urologist specialist*, and *specialist* are all at position 1 (first extracted form) and were selected by extractor 1 or 2, as indicated in Figure 3. *Stones* is at position 4 and was selected by both extractors. The shaded record in Figure 3 shows that the form *in my head* was extracted from location 5 by the researchers (".9") from document CMV-3.1 and interpreted to be a metaphor for "imagined," based on context. Note the modification of extractor 2's term *three stones* to *stones* in location 4. Because removing the *three* from *stones* transforms that form into a complete overlap with extractor 1's form, *stones*, the location designation was changed to "4.0", where ".0" represents forms selected by both extractors.

We made minor modifications to the extracted forms when they deviated too far from terminographical standards or did not follow our extraction guidelines. The most common modifications were deletions of numeric values, units of measurement, or a combination such as "50 mg." Long forms (i.e., phrases) were divided into constituent forms. A few forms not identified by the extractors but consistent with the form extraction guidelines, as judged by us, were added. We also annotated the forms, as appropriate (e.g., imprecise term, metaphor, and truncation/ellipsis). Overall, about 6% of extracted forms were modified to conform to our guidelines.

The final processing step was form normalization, a semi-automated process to standardize forms (Figure 4). Removal of various linguistic and form-based constructions greatly facilitated both analyzing forms (e.g., aggregation of inflectional variants for frequency counts) and mapping forms to the Metathesaurus concepts. Different "levels" of form normalization were used: moderate for analyzing forms by type (e.g., Diet Rich in Saturated Fat → diet rich in saturated fat) and aggressive normalization, including reordering words within lexical items and removing stop words, for automated mapping (e.g., Diet Rich in Saturated Fat → diet fat rich saturate). The UMLS lexical tools, Norm and LVG, were used to abstract away case, inflection, punctuation, stop words, and other morphological properties to reveal canonical (normalized) forms (McCray, Srinivasan, & Browne, 1994).

Note that collecting documents and processing extracted forms was an iterative process: consumer corpus processing stopped after about 25,000 form types (over 55,000 form tokens)[2] had been collected.

Mapping Forms to Concepts. The extracted forms were mapped to concepts in the 2002 Metathesaurus, an integrated database of concepts and terms from over 60 medical source vocabularies and classifications (UMLS, 2002). Part of the mapping was done with assistance from the MetaMap program. In brief, MetaMap tokenizes input text, generates lexical variants for noun phrases, matches tokens with generated variants against UMLS terms, and presents ranked candidate UMLS concepts based on a quality score for the closeness of match; 1000 is the highest score (Aronson, 1996).

[2] *Token* refers to any occurrence of an item. T*ype* refers to a unique item. For example, six occurrences of "Bell palsy" in a document would be counted as six tokens and one type.

Operation		Mechanism	Examples
1. Manual Review of Extracted Forms in the Context of the Text			
1.1	Acronym and abbreviation expansion	Manual	PSA → Prostate Specific Antigen DRE → digital rectal exam
1.2	Resolution of anaphoric reference, "inferred" terms, etc.	Manual	Suburban → Suburban Hospital digital results → digital rectal exam results
1.3	Spell-checking	semi-automated (e.g., MS-Word, *Dorland's*)	hepatitus → hepatitis
2. Automated String-Based Normalization Process			
2.1	Removal of punctuation (replaced by blanks)	automated (UMLS lexical program)	"mild" infection → mild infection Alzheimer's → alzheimers
2.2	Conversion to all lower-case characters	automated (UMLS lexical program)	Stomachache → stomachache
2.3	Conversion of lexical, derivational, and inflectional variants; change word order	automated; normalized forms suggested (UMLS lexical programs)	syndromes → syndrome foci → focus infected → infect cancers of the breast → breast cancer
3. Manual Review of the Normalization Results			
3.1	Selection of the appropriate normalized form	Manual	leave → leaf leave → left

Figure 4. Steps used for form normalization. Steps 1.1-1.3 were executed during extracted form processing. Step 3 was necessary because multiple normalized forms may result from one input form in step 2.3.

Initially, we assessed candidate quality scores from MetaMap and conducted different levels of manual review based on the following criteria, derived from a pilot study:

- Exact Match (Score = 1000): Manual review of the mapping was still necessary to check word sense. If no exact matches were appropriate, near matches were assessed.
- Near Match (1000 > Score ≥ 700): Manual review of candidate list for an appropriate (i.e., closely related) concept. Annotations ("Imprecise Term" and "Homonym") were consulted to disambiguate senses. If no candidates were appropriate, the unmatched form procedure was used.
- Unmatched (Score < 700): Manual analysis and manipulation of form to locate appropriate concepts interactively, using the Knowledge Source Server (KSS) tool (McCray, Razi, Bangalore, Browne, & Stavri, 1996), a labor-intensive process.

Each extracted form was mapped to the concept that, in our judgment, most closely matched the meaning of the form in context; mappings were labeled close or approximate (less/more specific or otherwise related). If none was suitable as a close or approximate mapping, modifiers were removed from the form and resubmitted to MetaMap or an interactive search was initiated in the KSS. Forms for which not even approximate UMLS concepts could be found remained unmapped.

However, part-way through this process, it became clear that MetaMap was not as effective for non-professional terms as hoped. Processing MetaMap output was more work than manual mapping, using the KSS and its approximate match capability. Perhaps the results of this study can be used to improve MetaMap's performance for non-professional terms. The Semantic Navigator tool (Bodenreider, 2000) was used for resolving mapping ambiguities. This tool provides a plethora of contextual information on particular concepts within the Metathesaurus. For example, the relative positions of concepts in the UMLS hierarchy, including their ancestors, siblings, and children concepts, are displayed. Such relationships are important because the vocabularies in the UMLS come from different health professions and are designed for specific groups of users with different tasks. Therefore, the intension of a particular concept is not always readily apparent.

The UMLS concept was added to the record for each form in the data file (Figure 5). After extracted forms are mapped to UMLS concepts (meanings), the <form, concept> pairs represent bona fide terms.

LocID	Extracted Form	R	Concept	CUI	Sem Type
1.1	urologist	S	Urologist	C0260214	professional or occu. group
1.2	urologist specialist	S	Urologist	C0260214	professional or occu. group
1.1	specialist	S	Specialists	C0087009	professional or occu. group
2.1	upper	N	Radiologic examination of upper GI	C0203057	diagnostic procedure
2.1	lower GI	N	Barium enema, NOS	C0203075	diagnostic procedure
2.2	upper and lower GI	N	Gastrointestinal tract function test	C0430148	diagnostic procedure
3.0	urlogist	S	Urologist	C0260214	professional or occu. group
4.0	stones	S	Kidney Calculi	C0022650	body substance
5.9	in my head	S	Imagination	C0020913	mental process

Figure 5. Sample records including mapped-to-UMLS concepts with the relationship R indicated as S = Same; B = mapped to Broader UMLS concept; N = mapped to Narrower UMLS concept. B and N are approximate mappings. The shaded record shows that the extracted form *urologist specialist* was closely mapped (same term, synonym, or quasi-synonym) to the UMLS concept "C0260214 (Urologist)" of the semantic type "professional or occupational group."

A medical expert was contracted to review a representative sample of problem or ambiguous cases. In addition, experts reviewed sample mappings for validation and correction. Mapping problems were discussed and resolved by consensus. While the expert was consulted on the most difficult cases, the majority of the mappings were based entirely on the researchers' judgments.

Data Analysis and Results

Corpus Generation

Overall, 1,936 electronic postings from 12 Web sites hosting medical discussion forums were collected.

Vocabulary Term Collection

The 14 extractors identified over 55,000 tokens. Two extractors reviewed each document independently. The inter-extractor agreement was 77% (55% complete, 22% partial). Approximately 13% of the form tokens were observed to be non-regular and were annotated:

- 5% abbreviations or acronyms (*Dr*. for *doctor*)
- 3% misspellings or typographic errors (*lupis* for *lupus*)
- 5% clippings (*doc*), metaphors or idioms (*plumbing*), and definitions (*gallbladder removal* for *cholecystectomy*).

Of the 35,326 form tokens reviewed by the researchers, 93% were closely mapped to a UMLS concept, representing 5,323 UMLS unique concepts and 116 semantic types. Despite our attempts to use MetaMap, most of the consumer forms were assigned to UMLS concepts manually.

To explore the extent of the problem using automated mapping techniques, two small-scale evaluations were conducted on a sample of 5,000 CMV forms. First, MetaMap quality scores for the sample were categorized using the score-based criteria provided previously, but taking only the top-ranked candidate concept without human review: 60% exact matches, 15% near matches, and 25% unmatched. Despite the large number of matches, only 10% of top concept candidates suggested by MetaMap matched the manually reviewed/hand-tagged concepts ultimately mapped to these forms. Second, more stringent criteria were tested to determine the feasibility of precision-enhanced automated mapping:

- Quality score of the top candidate concept provided by MetaMap ≥ 700
- No top candidate concepts with tied quality scores
- At least 200 quality points difference between the top and next candidate concepts

We found that of the 16% of mappings that met these criteria, 61% of the MetaMap candidate concepts matched the human assigned UMLS concepts. We speculate that characteristics of consumer terms, such as context dependency (non-specificity), form "irregularity," and form informality (idioms and metaphors), primarily account for these results and are investigating this finding further.

Vocabulary Characterization

This study used both quantitative analysis (counts, frequency distributions, and other descriptive statistics) and qualitative analysis (text analysis and interpretation of meaning).

Underlying these analyses is the distinction between form, concept, and term: A term consists of a form (a tangible surface-level structure) and a concept (an abstraction representing meaning) — represented as <form, concept>. Other facets include context (syntactic, pragmatic, and social), denotative and connotative senses, and relatedness. Our study addresses only a few terminological properties: form and concept separately and together, degree of overlap for clusters of terms (vocabularies), and observed usage patterns.

*Analysis of Forms***.** Counts, percentages, and lengths in characters and words (Table 2) were determined for each of these units. In addition, at the vocabulary level, the frequency or distribution of form occurrence was

calculated. Normalized forms were used in most of the analyses because minor form variations are not only of no consequence, but hinder form-based analysis.

Table 2. CMV form length

	Token	Type	Norm. Type
Mean Chars per Form	13.3	21.5	16.8†
Mean Words per Form	1.6	3.1	2.2‡

†23.5 in PMV ‡2.4 in PMV

Analysis of Concepts. Forms need to be arranged into synonym sets representing concepts. Since it was infeasible to do this manually, we mapped forms to concepts in the Metathesaurus; non-mapped forms are not considered in this analysis.

Because concepts represent ideas or abstractions, they are much more difficult to characterize than forms. In this study, approximately 740,000 concepts in the UMLS from a variety of medical and health-related professions provided a semantic reference point. Concept-based analyses include determining the number and percentage of forms in a vocabulary that can be mapped to UMLS concepts (Table 3). In addition, the number and percentage of closely or approximately mapped forms were analyzed.

"Semantic profiles" depicting the relative representation of different concept categories, by type and by token, were created to explore the nature of the CMV (Figure 6).

Figure 6. Semantic profile of CMV by semantic type group

Each UMLS concept is described by one or more of the 134 semantic types or categories of entities or events in the UMLS Semantic Network (UMLS, 2002). McCray, Burgun, and Bodenreider (2001) created 15 high-level semantic groups from the 134 types to simplify classification of UMLS concepts, facilitating comparisons.

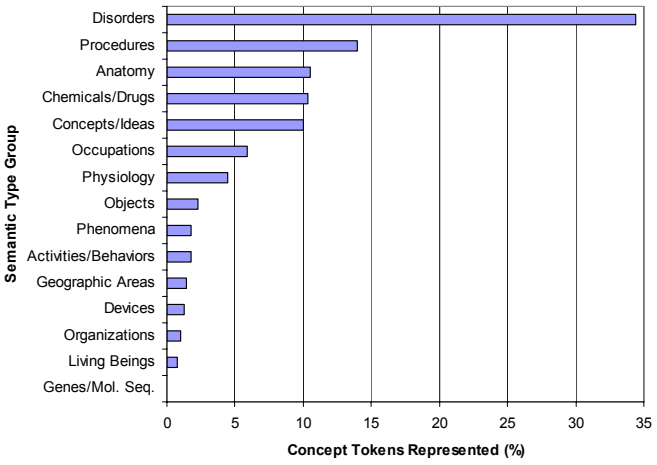

Table 3. Ten most frequently occurring CMV normalized forms (token), mapped-to UMLS concepts, and frequency

Rank	Form (Token)	Mapped-to UMLS Concept (Preferred Term)	Frequency (per 10,000)
1	doctor	Physicians	213
2	pain	Pain	108
3	diagnose	Diagnosis	88
4	test	Tests, Diagnostic	82
5	symptom	symptoms <1>	80
6	surgery	Surgery	67
7	cause	Causing	66
8	problem	Problem, NOS	61
9	treatment	Therapeutic procedure	46
10	drug	Pharmaceutical Preparations	44

Analysis of Form-Concept Combinations. Two primary relationships were explored. "Expressive variability" was defined operationally as the number of forms per concept; it represents the number of different ways people refer to the same concept. As seen in Table 4, some concepts had larger numbers of forms than others. Also of interest is how often these forms are used (token count). Is there one form, a "consensus form," that is used more frequently than other forms for a given concept? For example, a concept might have 5 forms, 4 with one token each and one with 16 tokens.

Table 4. Mapped-to UMLS concepts with the greatest expressive variability (greatest number of form types), rank by concept frequency, and consensus form

Mapped-to UMLS Concept	Form Types	Concept Freq. Rank	Consensus Form
Severe pain	36	53	severe pain
good health	20	94	healthy
Diagnosis	19	4	diagnosis
Dyspnea	19	123	breathing difficulty
Decrease	17	118	lower
Fatigue <1>	17	13	fatigue
Feels unwell	16	166	sick
Increased	15	150	raise
Lassitude	15	162	weakness
Therapeutic procedure	15	6	treatment

Table 5. Pair-wise concept (type) overlap of CMV (source, S) to MMV (reference, R) vocabulary

	Closely Mapped Concepts					
		(CMV and PMV) / CMV				(CMV not PMV) / CMV
			Degree of Form Commonality within Each Category			
Vocabulary Pairs	**Total** (n)	**Common** (% Total)	Complete (% of All Common)	Partial (% of All Common)	None (% of All Common)	**Distinct** (% Total)
CMV → PMV	4,830	81	49	19	32	19

Analysis of Relationships at the Individual Term Level. To get a sense of the difficulty consumers face when confronted with technical terms, we examined:

(1) how many of the consumer concepts occurred in the PMV and
(2) if a concept was found, to what extent consumer forms and professional forms for the concept overlap.

To determine whether a term belongs to PMV, we initially used two professional medical vocabularies: MeSH® (NLM, 2002) and SNOMED International® (Côte, 1998) to represent "professional" medical terms. Subsequent review of CMV terms not found in either MeSH or SNOMED, but judged to be within the "professional" medical domain (i.e., PMV concepts), were included in the form commonality calculation, as described below.

One-sided overlap between CMV and PMV was assessed to determine the degree of conceptual and terminological overlap (Table 5). The analysis was limited to forms closely mapped to UMLS concepts (i.e., identical terms, synonyms, and quasi-synonyms). This analysis was conducted in two steps:

- Conceptual Overlap: Determining the concepts that are common to both vocabularies and those that are distinct to one vocabulary
- Form Overlap: For common concepts, forms appearing in both vocabularies

Mapped-to concepts in CMV (the "source" vocabulary) were compared to those in PMV (the "reference" vocabulary). Common concepts, those appearing in both CMV and PMV, represent the conceptual overlap between these vocabularies. Concepts in CMV, but not in PMV, are distinct to CMV. In this study, the convention for denoting vocabulary pairs is CMV → PMV. The degree of conceptual overlap, CMV concepts in PMV, was expressed as one-sided overlap, the percentage of CMV concepts that are also found in PMV:

```
(Number of concepts in CMV and PMV / Number
       of concepts in CMV) x 100%
```

Non-overlapping concepts, those that appeared only in CMV and not in PMV, were referred to as "distinct" concepts and also expressed as a percentage:

```
(Number of concepts in CMV and not in PMV /
    Number of concepts in CMV) x 100%
```

Form commonality measures the degree of form overlap between the CMV (source vocabulary) and PMV (reference vocabulary) for a given common concept. There are three categories of form commonality (Table 6):

- Complete: For a given concept, all normalized forms in CMV are found in PMV
- Partial: For a given concept, at least one normalized form in CMV is found in PMV
- None: For a given concept, none of the normalized forms in CMV is found in PMV

Calculating form commonality for a given concept was a semi-automated process. Although string matching was used as a "first pass" to assess form overlap by common concept, the results were reviewed manually. To compensate for differences between forms extracted from natural language text and those in PMV from controlled vocabularies, the criterion that matching forms must be identical strings was relaxed in the following ways:

- Selected substrings, based on our judgment ("childbirth" and "birth")
- Lexical variants: "determine blood pressure" and "blood pressure determination"; "brush teeth" and "tooth brushing"
- Acronyms: "hrt" and "hormone replacement therapy"

Table 6. Examples of form commonality categories

UMLS Concept	CMV Form	PMV Form
Complete		
C0003842	artery	Arteries
Partial		
C0042963	vomit	Vomiting
C0042963	throw up	Emesis
None		
C0003449	cough medicine	Antitussive Agent
C0003449	suppressant	Antitussive Drug

Common forms for common concepts are shaded.

Table 7. Comparison of the corpus-based and "interactive" approaches of vocabulary term collection along several factors

Factor	Corpus-Based Approach	Interactive Approach
Resource	Documents	Participants
Linguistic Forms per Resource	Many forms	Few forms
Meaning	Implicit; General inferences	Explicit; Detailed structures
Suitability for Automation	High	Low

Discussion

We believe that the procedures developed and tested through this case study in the medical domain might be generalized to other domains. In this section we discuss a number of issues to be considered.

Other sources of terms are available and should be considered, namely direct interaction with members of the discourse group, such as semi-structured interviews and concept maps. Such methods elicit rich data and provide explicit access to participants' understanding of medical concepts (not possible with the corpus-based approach). Table 7 compares the approaches. We used a corpus-based approach to gain an overview of non-professional forms in the biomedical domain as a basis for more detailed studies. We intend to apply our results, using interactive approaches, to investigate how non-professionals think about technical concepts.

The corpus-based approach has several limitations. The major limitation is accurately assigning meaning to forms. Because only artifacts (text) and not the authors are available, understanding the forms is limited to the researchers' interpretation of each author's intent. Haas and Hert (2002) noted the problem as follows: "although the users' words can be seen, the intention behind the words, or what the individual actually wanted (the users' content and context), cannot be known" (p. 44). We attempted to decrease the bias by using full text documents, rather than query strings, to provide additional context. Nevertheless, the original intent can be understood best through using interactive approaches. As in the *asthma* example discussed earlier (Wood, 1991), the connotations associated with the term (deadly illness versus controllable condition) cannot always be discerned by inspection.

Another limitation is identifying "lay terms." The term extractors were instructed to select words or phrases that describe medical concepts. Some consistently identified the smallest lexical items while others preferred longer phrases (e.g., *heart attack* versus *having an acute heart attack*). In developing professional terminologies, there would generally be a policy on handling the degree of concept combination. However, the "appropriate" policy for non-professional terms is not clear. For example, whether *operation of the appendix* should be considered one term or two (*operation* and *appendix*) is not clear.

Finally, in the medical domain we were fortunate to have a comprehensive set of professional vocabularies in the form of the UMLS. In other domains, a large individual thesaurus may serve in the same capacity. If no such vocabulary exists, standardizing concepts would be difficult. On the other hand, the lexical normalization tools used in this study for standardizing forms should be applicable in other domains.

Conclusion

Non-professionals increasingly want to access information in specialized domains, but several significant barriers exist: identifying information needs, information seeking, and comprehension of domain-specific information. A common thread is terminology. Without an understanding of either the technical forms or the domain-specific concepts, non-professionals are not able to break through these barriers. Identifying all of the terms non-professionals use to describe domain-specific concepts is a challenge:

> Although there is merit as well as the need to comprehend the meaning of the language of consumers, to presume that it contains a knowable, stable vocabulary and grammar similar in structure to that of the formal languages of health care imposes a professional structure on a very personal experience. (Lewis et al., 2001)

Tracking and interpreting consumer terms is a "moving target," not only on a population level, taking into account culture, education, social status, and other factors, but also at the individual level, such as personal experience and learning through exposure to professional concepts. However, the need to help bridge non-professional and professional vocabularies clearly exists. "With the new Internet-enabled e-health environment… to enable patient participation, however, the words of the patient must be treated with as much respect as the words of the healthcare professional" (Rose et al., 2001). The procedures described in this paper are initial steps towards learning about the "words of the patient."

References

Aronson, A.R. (1996). MetaMap: Mapping text to the UMLS Metathesaurus. Bethesda, MD: NLM, NIH, DHHS.

Bodenreider, O. (2000). A semantic navigator tool for the UMLS. In J.M. Overhage (Ed.), Proceedings of AMIA 2000 Annual Symposium (p. 971). Philadelphia: Hanley & Belfus.

Côte, R.A. (Ed.) (1998). Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International (Version 3.5). Northfield, IL: College of American Pathologists; American Veterinary Medical Association.

Haas, S.W., & Hert, C.A. (2002) Finding information at the U.S. Bureau of Labor Statistics: overcoming the barriers of scope, concept, and language mismatch. Terminology, 8(1): 31-56.

Lewis, D., Brennan, P.F., McCray, A.T., Tuttle, M., & Bachman, J. (2001) If we build it, they will come: standardized consumer vocabularies. In V.L. Patel, R. Rogers, R. Haux (Eds.), Proceedings of MEDINFO 2001 (p. 1530). London: IOS Press.

McCray, A.T., Burgun, A., & Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In V.L. Patel, R. Rogers, R. Haux (Eds.), Proceedings of MEDINFO 2001 (pp. 216-220). London: IOS Press.

McCray A.T., Razi, A.M., Bangalore, A.K., Browne, A.C., & Stavri, P.Z. (1996). The UMLS Knowledge Source Server: A versatile Internet-based research tool. In Proceedings of AMIA 1996 Annual Symposium (pp. 164-168). Philadelphia: Hanley & Belfus.

McCray, A.T., Srinivasan, S. & Brown, A.C. (1994). Lexical methods for managing variation in biomedical technologies. In Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care (pp. 235-239).

NLM. (2002). Medical Subject Headings (MeSH). Bethesda, MD: NLM, NIH, DHHS.

Patrick, T.B., Monga, H.K., Sievert, M.C., Hall, J. H., & Longo, D.R. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. Journal of Medical Internet Research, 3, e24; Retrieved December 28, 2002 from http://www.jmir.org/2001/3/e24/.

Rose, J.S., Fisch, B.J., Hogan, W.R., Levy, B., Marshall, P., Thomas, D.R., & Kirkley, D. (2001). Common medical terminology comes of age, part two: current code and terminology sets—strengths and weaknesses. Journal of Healthcare Information Management, 15(3): 319-330.

Stavri, P.Z. (2001). Personal health information-seeking: A qualitative review of the literature. In V.L. Patel, R. Rogers, R. Haux (Eds.), Proceedings of MEDINFO 2001 (pp. 1484-1488). London: IOS Press.

Taylor, H. (2002). Cyberchondriacs Update: 110 million people sometimes look for health information online, up from 97 million a year ago. The Harris Poll®, #21. Retrieved May 10, 2003 from http://www.harrisinteractive.com/harris_poll/index. asp ?PID=299.

Tse, A.Y. (2003). Identifying and characterizing a "consumer medical vocabulary." Unpublished doctoral dissertation, College of Information Studies, University of Maryland, College Park.

UMLS. (2002). UMLS Knowledge Sources (13th ed.-January Release: 2002AA). Bethesda, MD: NLM, NIH, DHHS.

Wood, M.L. (1991). Naming the illness: the power of words. Family Medicine, 23: 534-538.