Mapping Vocabularies Using "Latent Semantics"

D.A. Evans, S.K. Handerson, I.A. Monarch
J. Pereiro, L. Delon, W.R. Hersh

July 1991

Report No. CMU-LCL-91-1

Laboratory for Computational Linguistics

139 Baker Hall
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Mapping Vocabularies Using "Latent Semantics"

D-A. Evanst, S.K. Handersont, LA- Monarcht, J. Pereirot, *L.* Delont, W.R- Hersh*

tLaboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh, Pennsylvania;

iBiomedical Information Communications Center, Oregon Health Sciences University, Portland, Oregon

June 8, 1991

Paper presented at *AMIA 91*, San Francisco, CA, June 8, 1991

Abstract. **Individual users of medical language manifest great variation in the expression of concepts and have difficulty in selecting appropriate terminology when confronted with systems that rely on standardized language, such as MeSH, SNOMED, or ICD, and the special terms sets of systems such as HELP, INTERNIST-I/QMR, and DXplain. Indeed, the need to map natural language into appropriate special terms—as well as the need to map one system's specialized terminology into another's—is one of the problems being addressed by the National Library of Medicine's UMLS System, with its associated information sources maps. The problem is extremely difficult, in part, because such mappings depend on *semantic* equivalences among terms, not merely the superficial matching of words or phrases.**

As a general and robust solution to the problem of mapping across vocabularies, we implemented a version of "latent semantic indexing", taking terms from different vocabularies as the 'documents' to be retrieved by natural-language expressions of concepts, taken as 'queries'. In one of several experiments testing our approach, for example, we selected approximately 225 terms each from the INTERNIST-I/QMR, PTXT, and META-1 vocabularies corresponding to clinical findings under the *physical exam.* We constructed a *source* matrix of associations between the findings and all the 'words' they contained, supplemented with word-level synonyms and related terms. The resulting source rectangular matrix was approximately 650x3000. Under singular value decomposition, this was reduced to a compressed space of at most 650 dimensions. The performance of the reduced space as a "latent semantic" map of the source domain was evaluated by processing phrases, intended to be interpreted as clinical findings, as term-vectors, projected into the reduced space. In calculating the projections, we considered only the 150 most significant dimensions. Mappings to concepts (i.e., standardized terms) were determined by taking the cosine-distance measure of the vector to all the term-points in the reduced space. As an example of the results, the phrase "stomach discomfort worse after eating" scored as follows (taking the cosine squared measure as a score of 'relevance', for the top four):

0.563391 [PTXT] ABD PAIN, AGGRAVATED BY EATING
0.529701 [QMR] DIET INTOLERANCE TO SPECIFIC FOOD <S>
0.499395 [PTXT] CHEST PAIN, MADE WORSE BY EATING
0.494474 [META-1] EATING DISORDERS

The power of the approach is that it does not depend on explicit, declarative semantic representations or on word-for-word correspondences among terms; and that multiple vocabularies can be represented side-by-side.

Communications should be directed to:
Dr. David A. Evans, Laboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890. (dae@ld.cmu.edu)

Mapping Vocabularies Using "Latent Semantics"

D.A. Evanst, S.K. Handersont, LA* Monarchy
J- Pereirot, L. Delont, W.R. Hersh*

t**Laboratory for Computational Linguistics,**
Carnegie Mellon University, Pittsburgh, Pennsylvania;
‡+Biomedical Information Communications Center,
Oregon Health Sciences University, Portland, Oregon

June 8, 1991

1 The Problem: Managing Semantics in Medical Vocabularies

Many systems that store information in natural language make that stored information available to users through language subsets. In the simplest case, the subsets are isolated strings of the language—access is afforded via inverted indexes. In more refined instances, the subsets are controlled vocabularies or encoding schemes—access is afforded via selection of terms. In fact, almost any process that succeeds in *indexing* natural language will generate a subset of words or phrases that are taken to represent *concepts* in the texts to which they are linked. Indexing and encoding schemes in biomedicine reflect such practice. The biomedical literature is indexed using MeSH terminology. Hospital records are often annotated with SNOMED^ or ICD^[18,] codes. Many medical institutions have 'home-grown' coding schemes to manage patient data. (Cf. [16] for one example at the Mayo Clinic.)

The problem is not confined to systems that provide access to textual materials. Varieties of computer systems that support medical decision making—such as HELP,I[26,] INTERNIST-I/QMR,'[22,23,24,] and DXplain'[1]!—utilize medical language subsets to indicate information (or concepts). Input to such systems often must be in the form of a restricted terminology; output is typically expressed in a restricted language (often the same set of terminology that is acceptable as input).

In the face of the common practice to restrict terminology, individual users of medical language manifest great variation in expressing concepts. In general, people can be expected to show very little agreement in preferred choices of language to make observations^ and will have difficulty in selecting appropriate terminology when confronted with systems that respond only to limited subsets of terms. (Cf. [3] for a study in which this effect is shown to contribute to poor (20%) retrieval of relevant information.)

Indeed, the need to map natural language into appropriate special terms—as well as the need to map one system's specialized terminology into another's—is one of the problems being addressed by the National Library of Medicine's (NLM) UMLS System, with its associated information sources mapJ[17,19]! The problem is extremely difficult, in part, because such mappings depend on *semantic* equivalences among terms, not merely on the superficial matching of words or phrases. The semantic typing of terms and the basic semantic network in the UMLS System represent a first step in the direction of developing an independent, semantic basis for associating terms from different vocabularies.^[20,21,] The problem is certainly not new—a number of studies have advocated semantic typing, decomposition, and networks to establish and control cross-vocabulary concept representations^[2,9,10,]—but it is becoming increasingly urgent in the context of efforts to unite multiple terminologies.

2 The Basis of a Solution: Latent Semantic Indexing

One approach to the problem of finding dependencies among 'information objects' involves a procedure called "Latent Semantic Indexing* (LSI)*W* In its applications to date, LSI has been used principally to improve document retrieval. In brief, the method takes advantage of the fact that words in documents do not co-occur randomly: there are implicit dependencies among them based on their meanings. LSI facilitates the construction of a high-dimensional space in which the words and documents are co-located. Through transformations of the space, it is possible to derive a similar space in which a subset of the dimensions maximally clusters the documents. The new, reduced set of dimensions is statistically the 'best' representation of the hidden dependencies among the words in the documents—the 'latent semantics'.

In the following sections we describe the basis of LSI as applied to document indexing and discuss the strengths and weaknesses of the approach. Given the space limitations of this paper, our discussion of these points is necessarily abbreviated.

2.1 Brief Characterization of **LSI**

While the basic idea of using word co-occurrence dependencies to define a semantic space is quite intuitive, the procedures associated with LSI are not. A sense of the LSI approach may be obtained by a description of LSI methodology. We offer a characterization of LSI from the point of view of its use in information retrieval, not a full exposition of the method. (The reader is referred to [7] for a detailed explanation of the latent semantic indexing process, with sample matrices and numbers.)

In general, documents can be regarded as collections of words. A correlation matrix (such as is given in Figure 1) can be constructed to make this correspondence explicit. In such a matrix, the columns represent documents and the rows represent words that appear in the documents. In practice, such correlation matrices are rectangular; the number of words is greater than the number of documents. A value can be assigned to a word (e.g., "0" or [a]l") based on whether it appears in a document or not. Each word-row defines an orthogonal dimension; documents are thus located in a high-ordered space. For virtually any actual collection of documents, the matrix will be 'sparse': there will be many zeros, effectively giving regions in which some subsets of documents are not found.

A correlation matrix of the sort in Figure 1 makes clear why word-based indexing and retrieval can lead to bad results. Traditional word-based retrieval depends on finding a match between a word and a document—a cell in the matrix containing a "1". When users use different words (perhaps meaning the same thing) they will retrieve different documents, since two different words will almost always have different patterns of "0"s and [a]l"s. LSI attempts to circumvent this problem by indexing documents based on secondary and tertiary associations of words—essentially, 'discovering' the semantic relations that discriminate among alternative word meanings, as revealed by the co-occurrence patterns of words in documents.

Beginning with a *word* x *document* source matrix, M, the LSI process derives a dimensionally 'reduced' space using a method called "singular value decomposition" (SVD)J[1W5j] The effect of the reduction is that previously orthogonal word-dimensions are coerced into derived, composite dimensions based on their ability to 'fit' documents into the reduced space. In particular, words and documents that are closely associated are placed near one another in the reprojected space. LSI actually approximates the source space using the *k* largest

	D_1	D_2	D_3	...	D_k
$word_1$	1	0	0	...	0
$word_2$	1	1	0	...	0
$word_3$	1	0	0	...	1
$word_4$	1	0	1	...	0
$word_5$	1	1	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮
$word_n$	0	0	0	...	0

Figure 1: Schematic Representation of a Source *Word*×*Document* Matrix

singular vectors of the word–document relevance matrix, where k is some user-selected number. (The number is typically in the range of 100–200—big enough to capture the semantic structure and small enough to avoid noise).

Decomposition allows one to reconstruct (reproject) a matrix—M′—having the full dimensionality of the source matrix M but with the effect of locating all the points as though they were in a reduced space. There are two ways to view the result. First, the original matrix has been transformed from a high-dimensional ($O(10^3)$) space into a space of a smaller number (e.g., 100–200) dimensions. Documents represented by vectors in the larger source space are now transformed into vectors in the smaller space, 'pushing them closer' to other documents. Second, the original matrix has been approximated (in a least-squares sense) with a smaller number (e.g., 100–200) vectors. When we reconstruct the original matrix (deriving M′), many of the entries that were zero will contain values other than zero. A synonym of a word is likely to have a non-zero value and to be taken into consideration in the retrieval process.

Queries can be interpreted as vectors of the words they contain and can be projected into the space defined by M′. A distance from the query-vector to all the points in the space can be calculated, for example by calculating a cosine distance measure—the cosine of the angle of the query vector and the vector to each document-point. Those points that are closest (geometrically) on this measure are the 'best' responses to the query.

2.2 Strengths & Weaknesses of LSI

LSI has a number of important strengths. In a variety of experiments and applications, it has been shown to lead to improved retrieval. It provides extremely good fits of queries to documents. Queries can be made in natural language and are virtually unrestricted in length: another document can be used as a query, for example. The method is general and robust.

However, LSI has several weaknesses that make it unsuitable for unrestricted information-retrieval applications. Most critically, LSI is computationally expensive (hence, impractical for use with large document collections) and the value of LSI in discovering "semantics" is weakened to the extent that polysemous words in any collection will lead to bad results of processing. We elaborate on these points briefly.

In practice, LSI/SVD is too computationally expensive to be applied to large document sets: currently, the processing of non-trivial matrices on the order of 10,000×10,000 can take significant time (e.g., days) on a supercomputer. Given these limitations, even with

reasonable increases in the efficiency of SVD algorithms and in the speed of workstations, it is not likely that the general method will apply to 'industrial-sized' indexing and retrieval problems.

Because the method exploits the co-occurrence of items in documents to derive pseudo-semantic associations, any words that are used with different senses will result in semantic distortions. As an illustration of this effect, consider the case where the polysemous word "digit" occurs in some documents with the sense *finger or toe* and in other documents in the same collection with the sense *number*. Presumably the two sets of documents will not occupy the same general space—the location for those documents having to do with fingers and toes will likely be different from the location of the documents having to do with numbers. The single dimension defined by the string "digit" is bound to 'split' the distance between the two clusters. Such a derived sense for "digit" will be true of neither of the original senses.

3 The Hypothesis: LSI Applied to Terms

In work under the CLARIT Project,'[1142]' a variation of the LSI approach described above is being used as a general and robust solution to the problem of mapping unrestricted language into sets of appropriate index terms. In particular, while the 'traditional' use of LSI is to relate words and documents, the CLARIT approach involves relating words (or language, more generally) to *terms*.

3.1 The Special Case of Relating Non-Homogeneous Terms

One special case involves the use of LSI to derive the common semantic space of terms from different vocabularies. Each term is treated as a 'document'; the information content of the term is given in natural language, which can be treated as the 'words' of the 'document'. An LSI space not only clusters terms according to their implicit semantics, but can also be used to map any natural-language variants of the terms to the set of 'best-matching' terms in the space. In effect, the variant expressions are taken as 'queries' and the retrieved terms are the 'documents' that best match the query.

The matrix in Figure 2 illustrates one possible realization of a *word* x *term* space. In this case, the terms are treated as *concepts* having sub-conceptual structure given by *lexical items*. Some of the lexical items actually occur in the term. These have the highest association values. Others do not occur but are similar in meaning to items that do occur. These have lower values. Except for the differential weighting of lexical items and the introduction of related items along with those that actually occur in a term, such a matrix is identical in form to the *word*document* matrix of Figure 1.

3.2 The Decomposition of Concepts into Lexical Components

The approach we take in building representations for medical-term space is a generalization and extension of the strategy illustrated in Figure 2. As summarized in Figure 3, we treat a term (or concept) as a collection of lexical items. The concept can be regarded as containing, as well, the *equivalence class* and a set of *related* lexical items. Indeed, the concept itself might have equivalent or related terms (expressible as lexical items), which can also be treated as part of its meaning. All such lexical items define a vector for the concept. Different

	C_1 Postprandial Abdominal Discomfort	C_2 Chest Pain Substernal At Rest	...	C^* Irregular Heart Beat
Postprandial	5	0	...	0
After Eating	4	0	...	0
Food	2	0	...	0
Dinner	2	0	...	0
⋮	⋮	⋮	⋮	
Abdominal	5	2	...	0
Stomach	4	1	...	0
Belly	4	2	...	0
Chest	2	5	...	2
⋮	⋮	⋮	⋮	
Discomfort	5	4	...	0
Pain	4	5	...	0
Distress	4	4	...	0
⋮	⋮	⋮	⋮	
< *Lexical-Item*$>_n$	0	0	...	0

Figure 2: Example of a Partially Completed Source Matrix

weights for different categories of lexical relations determine the relative magnitude of the vector in each of the many lexical dimensions that comprise it.

Since concepts will exist in a complex space containing many other concepts, we need to insure that appropriate contrasts are preserved. We would not want "high blood pressure" and "low blood pressure" to occupy the same space simply because they share so many common features through the sub-concept "blood pressure". In fact, we want them to contrast: on the 'blood-pressure' scale, they are opposites. Thus, in developing a concept vector, we include semantically-appropriate 'opposites' among the lexical items, with *negative* magnitudes to insure their separation in the term space. Note that the weights we give in Figure 3 are suggestions. At this stage in our evaluation we have no reason to nominate particular weighting strategies except in theory.

Figure 4 gives a concrete example for the term "postprandial abdominal discomfort". The sets of 'equivalent' and 'related' lexical items (and their opposites) are not intended to be exhaustive; they merely illustrate the types of lexical items that one would attempt to include under each category. Though there are many dimensions for this one term in isolation, the dimensionality of the space for a collection of terms from the same domain will not be excessively great, since we can expect many of the same lexical items to appear in other terms. Thus, the space in which terms are located can be kept relatively small, insuring computational tractability.

4 Experiments: **LSI Mappings of Medical 'Term Space'**

We have conducted a variety of experiments to refine our methods for developing latent semantic spaces for medical terms. In this section, results from several experiments are presented, principally to iUustrate our basic methodology and some of the properties of LSI

	C = *Lex-Itemi + Lex-Item? + ... -f Lex-Itemn*
Lex-Itemi	*k*
Lex-Item?	*k*
⋮	\
Lex-Itemn	*k*
Eq(C)	*k*
Eq(Lex-Itemi)	*k-m*
Eq(Lex-Item?)	*k-m*
⋮	⋮
Eq(Lex-Itemn)	i-m
Rel(C)	*k-m*
Rel(Lex-Item$_l$)	*k-2m*
Rtl(Lex-Item?)	*k-2m*
\	⋮
Rel(Lex-Item)*	*k-2m*
Eq-'(C)	*-k*
Eq-^{1}(Lex-Item$_x$)	*-(k-m)*
Eq"^{1} (Lex-Item?)	*-(k-m)*
\	⋮
Eq"^{1} (Lex-Itemn)	*-(k-m)*
Ret^1(C)	*-(k-m)*
Ref^i(Lex-Item$_l$)	*-(k-Sm)*
Ret^1 (Lex-Item?)	*-(k-£m)*
i	;
Ref^1 (Lex-Itemn)	*-(k-Sm)*

Key:

C = a concept; term-phrase

Lex-Item = a unit lexical item; word or (sub-)phrase

Eq(X) = the equivalence class of X; a set of sense-preserving variants and synonyms of X

Rel(X) = the non-synonymous terms (words and phrases) related to X

Eq~^{1}(X) = the inverse equivalence class of X; a set of sense-appropriate antonyms of X

Re\T^x(X) = the inverse set of terms related to X; the pragmatically appropriate contrasts to X

k = a constant rational in the range $1 \leq k \leq 5$

m = a constant rational in the range $0 \leq m \leq k/2$

Figure 3: Generalized Concept Vector Illustrating Weighting of Lexical Items

		C = Postprandial Abdominal Discomfort
Lex-Items:	postprandial	*4*
	abdominal	*4*
	discomfort	*4*
Eq(C):	—	—
Eq(postprandial):	after eating	*5*
Eqfabdominal):	stomach	*5*
	belly	*3*
	intestine	*3*
Eq(discomfort):	pain	*3*
	distress	*3*
	upset	*3*
Rel(C):	dyspepsia	*3*
	indigestion	*3*
Rel(postprandial):	food	*2*
	meal	*2*
	dinner	*2*
	lunch	*2*
	breakfast	*2*
	full	*2*
	eat	*2*
Rel(abdominal):	chest	*2*
	groin	*2*
	side	*2*
Rel(discomfort):	burping	*2*
	burning	*2*
	sharp	*2*
• *Eq->(C):*	—	—
Eq~[1] (postprandial):	before eating	*-3*
Eq"^(abdominal):	head	*-3*
	neck	*-3*
	extremity	*-3*
Eq"[1](discomfort):	relief	*-3*
Ref(C):	—	—
Ret~[1](postprandial):	hungry	*-2*
	empty	*-2*
Ret"[1](abdominal):	—	—
Ret[1] (discomfort):	mild	*-2*
	moderate	*-2*

Figure 4: Example of a Concept Vector for *Postprandial Abdominal Discomfort*

term spaces. All of the experiments involved the following procedure:

1. Terminology was selected corresponding to clinical findings under *physical exam* from the INTERNIST-I/QMR, HELP/PTXT, and NLM/UMLS META-1 vocabularies. Every experiment involved some terms from each; the numbers were roughly equal. An example of some of the terms chosen is given in Table 1.
2. Terms were decomposed lexically (and normalized morphologically). Decomposition was accomplished automatically, then checked by hand. Essentially, all terms were broken into their constituent words. Function words of English (such as "the", "a", "of, etc.) were discarded. Any cases where words should not have been split (as when a pair of words or a phrase formed a unit lexical item) were corrected.
3. Lexical items were supplemented with synonymous and related terms. A sample of the lexically related items is given in Table 2. The sets of synonyms and related terms were derived using two methods. First, medically-knowledgeable members of the team (Hersh and Pereiro) added lexical variants, synonyms, and related terms by hand for each lexical item derived from the terms. In practice, this was not a time-consuming task: most of the synonyms were produced by one person (Hersh) in approximately three hours of work. Second, the sets of associated terms were reviewed by other members of the team for consistency and completeness. At this stage, some additional terminology was derived from available medical dictionaries.[8,25]
4. A source *Lex* x *Term* matrix (M) was created with different values for lexical entries based on their status in terms.
5. SVD was performed. At present, we use a fairly standard, numerically stable algorithm due to Golub and Reinsch[15] Unfortunately, the algorithm has execution time on the order of n^3 for the smaller dimension (typically the number of documents) and n^2 for the larger dimension (typically, the number of terms). A complete decomposition for matrices of the size we are currently experimenting with requires up to 24 hours on a DECsystem 5820.
6. Natural-language statements (treated as 'queries') were decomposed into lexical vectors and projected into the compressed space, M', as determined for variable numbers of factors (typically in the range 50-300).
7. Terms in M' ('retrieved' by the query) were ranked based on their cosine distance from the query vector. (We actually use the square of the cosine.) In practice, retrieval takes approximately one second: all distances are calculated and the top n terms, typically 10-20, are displayed.

4.1 Experiment 1: 648x3891 Space

Experiment 1 involved a space of 648 terms and 3891 lexical items. Approximately 225 terms each were taken from the three source vocabularies. All lexical items were given equal values of "1" or "0", depending on whether they were associated with a term. SVD took approximately six hours on a DECsystem 5820.

Table 3 presents two sample results. The number of factors used (to give the reduced dimensionality of the space) is indicated before the 'query'. The examples illustrate one of the principal effects of the method: terms that are 'retrieved' do not have to share lexical items with the 'query'. In the case of the first example, the only common lexical item is

UMLS/META-1	HELP/PTXT	INTERNIST-I/QMR
Heartburn	Abd pain made worse with bending	Abdomen mass paraortic
Hemianopsia	Abd pain nocturnal	Abdomen mass periumbilical
Hemiplegia	Abd pain periumbilical	Abdomen mass right lover quadrant
Hiccup	Abd pain radiates to back	Abdomen pain present
Hoarseness	Abd pain radiates to left chest	Abdomen pain right upper quadrant
Hyperalgesia	Abd pain radiates under sternum	exertional hx
Hyperbilirubinemia	Abd pain recurring	Abdomen small bovel visible
Hyperbilirubinemia, hereditary	Abd pain recurring, duration	peristalsis
Hypercapnia	greater than tvo years	Abdomen tenderness generalized
Hyperesthesia	Abd pain resolved by vomiting	Abdomen tenderness hypogastrium
Hypersomnia	Abd pain rlq (right lover quadrant	Abdomen tenderness periumbilical
Hyperventilation	Abd pain ruq	Abdomen tenderness rebound
Hypesthesia	Abd pain severity causes	generalized
Hypothermia	diaphoresis	Abdomen tenderness rebound
Hypoventilation	Abd pain sharp	localized
Illusions	Abd pain sharp or cramping	Abdomen tenderness right lover
Insomnia	Abd pain vorse with movement or	quadrant
Jaundice	cough	Abdomen tenderness suprapubic
Jaundice, chronic idiopathic	Abdominal fullness, epigastric	Abdomen tympanites
Jaundice, neonatal	Abdominal fullness,	Abdomen urinary bladder palpable
Kernicterus	hypogastric/suprapubic	or percussable
Lameness, animal	Abdominal fullness, llq	Abdomen vail draining sinus <es>
Language development disorders	Abdominal fullness, luq	Abdomen vail fluctuant mass <es>
Language disorders	Abdominal fullness, periumbilical	Affect anxious and/or fearful
Liver cirrhosis, biliary	Abdominal fullness, rlq	Affect apprehensive
Halignant hyperthemia	Abdominal fullness, ruq	Affect blunted or flat
Heningism	Acid or food regurgitating up into	Affect depressed
Higraine	the pharynx	Affect depressed vorse in morning
Houth breathing	Acid or food regurgitation with	Affect euphoric
Hovement disorders	choking on fluid regurgitant	Affect labile
Husde hypertonia	Acute chest pain	Chest pain apical stabbing
Muscle hypotonia	Alternating constipation/diarrhea	Chest pain girdle distribution
Husde rigidity	Annual breast self examination	Chest pain lateral dull aching
Husde spasticity	Bloody diarrhea	Chest pain lateral sharp
Hutism	Bloody stool	Chest pain lateral sharp recurrent
Hyodonus	Breast dimpling	attack <•> hx
lausea	Breast discoloration	Chest pain substernal at rest
leuralgia	Burning chest pain	Chest pain substernal burning
leurologic Manifestations	Chest pain interferes with sleep	Chest pain substernal crushing
...	...	...

Table 1: Examples of the *Findings* Terminology from Different Systems

```
accoaaodate 1 contain I containing I bound I enclose I include I comprise I hold
accompanied I superpuesto | con I acoapana I acoapanado I with
accompany I with | along
ach« I problea I disease I discomfort | pain I difficulty I difficult
aching I pena I dolor I doloroso I hurting I tender I distressing
       I smarting I throbbing I sore I irritating I uncomfortable
       I pain I painful I calaa I irritativo I opresivo I dislacerante
       I agudo I pulsante | pupa I irritant* I disconfort
...
acrid I burning I acid I caustic I acute I sharp
act I behavior I Movement I »ove I action I do I conduct
action I behavior I move I act I do I conduct I movement
       I Motion I activity
activities I motion I movements I movement
activity I exertion I labor I work I constitutional I body I exertional
         I exercise I movement I stool I Motion I action
acute I pain I severity | strong I severe I harsh I burning
      I acid I caustic I acrid I sensitive I penetrating
      I shooting | high I annoying I threatening I stabbing
      I piercing I cutting I intense I peaked I pointed I sharp
      I sever I extreae I rapid I sudden I abrupt I painful
      I excruciating I dire I impending I iaainent I deep
      I serious I aajor I great I critical
...
beat | flap I tick I pulsate I pulse I throb I heartbeat
bed I lying I down I reclining I resting I recline
beef | steak I aeat
...
bleed I bleeding I hemorrhage I blood
bleeding I epistaxis I nosebleed I bloody I bleed I blood
blind I heaianopsia I heaianopia I blindness I half I vision I loss
...
colic I pain I spasa I colon
colicky I spasmodic I spasa I intermittent
...
diffuse I scatter I extend I scattered
digest I postprandial I ingestion I eating I meal | after I later
       I lunch I dinner I eat
digestive I intestine I tummy | food I eat I stomach
dilatation I dilate I dilation I opening I expansion I open I swell
           I swelling I widening | increase I enlargement
...
```

Table 2: Sample Sets of Related Terms

```
100: stomach discomfort worse alter eating

0.563391  [PTXT]    Abd pain, aggravated by eating
0.529701  [QMR]     Diet intolerance to specific food <s>
0.499395  [PTXT]    Chest pain, made worse by eating
0.494474  [META-1]  Eating disorders

100: coffee ground emesis

0.849161  [QMR]     Vomiting coffee ground
0.745221  [QMR]     Vomiting feculent
0.517106  [META-1]  Vomiting
0.428197  [PTXT]    Recent vomiting, hematemesis
```

Table 3: Sample Results on 648x3891 Space

"eating" and it appears only once in the top two terms. The second example shows similar effects. Another observation, of course, is that the 'queries'—though reasonably formulated as expressions of medical findings—do not have exact matches with any of the findings from the three vocabularies. The mapping, thus, performs the additional function of showing the user which of the available terms might best match the concept he or she is attempting to express.

4.2 Experiment 2: 822x3015 Space

Experiment 2 involved a space of 822 terms and 3015 lexical items. 179 terms were taken from META-1, 221 from PTXT, and 422 from INTERNIST-I/QMR. Exact-matching lexical items were given a value of "5". Equivalence-class and related lexical items were given a uniform value of "4". SVD took 22 hours on a DECsystem 5820.

Table 4 presents a number of results. As in the case of the examples from Experiment 1, exact lexical matching is not required to retrieve appropriate terms. The examples also reveal some of the properties of the semantic space of the terms. The numbers are ranked based on the square of the cosine of the angle separating the query vector from the term vectors. Terms that are 'close' to one another will have similar distance from the query vector. In general, the set of closest terms will define a location in semantic space—the region in which the corresponding concept is represented. Naturally, in practical applications of expression mapping, we would use only the highest-ranking terms and would discard terms that dropped off in distance from the highest ones.

One can also see in the examples that the similarities of individual terms in the three vocabularies is captured without having to establish a mapping from vocabulary to vocabulary or term to term. In particular, terms are located in the same homogeneous space; the distances between terms gives an absolute measure of similarity.

4.3 Experiment 3: 369x3084 Space—English and Spanish

Experiment 3 involved a space of 369 terms and 3084 lexical items. Approximately 125 terms each were taken from the three source vocabularies. In developing the source matrix, we included lexical items in Spanish as well as English. (Some examples of the Spanish

150: cough blood	0.620719	[PTXT]	Hemoptysis
	0.534173	[META-1]	Cough
	0.534173	[QMR]	Cough
	0.528974	[PTXT]	Non-productive cough
	0.450769	[PTXT]	Cough productive of blood-streaked sputum
	0.372125	[PTXT]	Productive cough
	0.314800	[PTXT]	Recurring cough and sputum production
	0.300045	[META-1]	Anoxemia
150: rough voice	0.770907	[QMR]	Hoarseness
	0.770907	[META-1]	Hoarseness
	0.507794	[PTXT]	Hoarseness or a change in the voice
	0.237814	[META-1]	Aphonia
	0.159180	[META-1]	Voice disorders
	0.136267	[META-1]	Vocal cord paralysis
	0.091883	[PTXT]	A dry throat
	0.040473	[QMR]	Dehydration
150: uncontrolled repeat speech	0.602760	[QMR]	Speech echolalia
	0.465534	[META-1]	Echolalia
	0.201566	[QMR]	Speech neologisms
	0.201566	[QMR]	Speech explosive
	0.201566	[QMR]	Speech perseveration
	0.201566	[QMR]	Speech monotonal
	0.201566	[QMR]	Speech scanning
	0.185416	[QMR]	Speech slow
150: muscle quiver	0.543594	[META-1]	Fasciculation
	0.525968	[META-1]	Tremor
	0.433057	[META-1]	Muscle rigidity
	0.394071	[QMR]	Myalgia
	0.324234	[QMR]	Muscle <s> cramp <s>
	0.319028	[META-1]	Muscle hypertonia
	0.302419	[META-1]	Muscle spasticity
	0.286339	[META-1]	Torticollis
150: decrease breathe	0.816871	[META-1]	Hypoventilation
	0.644709	[QMR]	Breathing biots
	0.644709	[QMR]	Breathing cheyne stokes
	0.391361	[META-1]	Hyperventilation
	0.294757	[META-1]	Mouth breathing
	0.158838	[META-1]	Hypothermia
	0.158838	[QMR]	Hypothermia
	0.113876	[META-1]	Apnea
150: cannot sleep	0.492858	[META-1]	Insomnia
	0.492858	[QMR]	Insomnia
	0.253118	[QMR]	Nocturia
	0.208416	[META-1]	Somnambulism
	0.208416	[QMR]	Somnambulism
	0.204001	[QMR]	Sleep paralysis
	0.177027	[META-1]	Hypersomnia
	0.154213	[QMR]	Sleeping excessive

Table 4: Sample Results on 822×3015 Space

```
ISO: dificultad para respirar acostado
     [ difficulty breathing lying down ]

0.636176 [PTXT]  Orthopnea
0.636176 [QMR]   Orthopnea
0.606431 [QMR]   Dyspnea paroxysmal nocturnal
0.565984 [PTXT]  Dyspnea
0.548591 [QMR]   Insomnia

150: emision heces negras con presencia sangre
     [ black stools "emission" with blood presence ]

0.824417 [PTXT]  Bloody stool
0.824417 [PTXT]  Bloody stools
0.734859 [PTXT]  Helena
0.676487 [PTXT]  Greasy stools
0.641968 [PTXT]  Bloody diarrhea
```

Table 5: Sample Results for Spanish on 369x3084 Space

lexical items are found in Table 2.) All lexical items received a "1" or "0" value. SVD took approximately ninety minutes on a DECsystem 5820.

Table 5 gives results. English glosses of the Spanish phrases are provided below each phrase (but were not used in the 'retrieval' process). We include these examples to illustrate another potential use of LSI-term mapping. Because the method operates only on strings (= lexical items) and is not sensitive to features of actual natural language, it is possible to decompose terms in one language with lexical items from another. Indeed, the lexical items of several languages can be used side by side to represent the conceptual content of terms. The resulting semantic space will locate terms on multi-lingual dimensions. 'Queries' in one language can be used to 'retrieve' terms in another—without overt translation. The examples in Table 5 show this effect.

5 **Conclusion: Developing General, Robust Methods for Medical** Semantics

The are several important features of our *Lex* x *Term-LSI* method. It does not depend on explicit, declarative semantic representations or on word-for-word correspondences among terms. Multiple—arbitrary—vocabularies can be represented side-by-side. The required resources—lexical-item correspondence sets—can be developed quickly and authoritatively. There is tolerance of noise; 'fuzzy' approximations are handled automatically. It is entirely algorithmic.

Some of the difficulties of the traditional *Word x Document*-LSI approach are effectively avoided. We circumvent the problem of polysemy because we work with terminology in a sub-language: there is little inherent polysemy and actual problems can be anticipated and treated as special cases. We work with spaces that remain computationally tractable, since we do not deal with more than several thousand terms at a time. (Indeed, for sub-domains of medicine, several thousand terms is quite sufficient.)

The methods we have described need further refinement. We clearly also need to experiment with different strategies for weighting lexical items and for decomposing terms.

Some LSI research must be devoted to developing or applying better SVD algorithms, especially to take advantage of the inherent sparseness in the input data. In addition, better decomposition algorithms could make use of an upper limit on the expected or approximate dimensionality of the latent semantic space, then calculate only to that size. Lanczos methods show some promise: since we are only interested in the 'most significant' aspects of the data, such a statistical approach seems appropriate. However, some implementations of Lanczos[5,6] have potentially significant drawbacks. Global orthogonality of the singular vectors is not assured; the method has difficulty resolving close singular values, and therefore vectors; and its operation is not as clearly automatic as others.

We have argued that the success of any attempt to unite multiple medical vocabularies or to link restricted medical terminology to natural language will depend on our ability to treat terms as semantic objects and to relate them based on their conceptual content. Traditional attempts to create explicit semantic networks for terms have had only limited success. In general, semantic networks are expensive to develop, controversial in their structure, and epistemologically—and computationally—problematic. We clearly need general, robust, and empirically sound methods for discovering and utilizing the semantics of terms. We believe that the approach we outline in this paper has promise as one such method.

Acknowledgements. The CLARIT has been supported by grants from the Digital Equipment Corporation. Mary Hart and Robert Lefferts provided valuable assistance in developing resources for the LSI experiments reported here. Chris Chute and Yiming Yang performed some of the first independent experiments confirming the utility of our approach. Their reports have been instructive.

References

[1] Barnett, CO., Cimino, J.J., Jupp, J.A., and Hoffer, E.P. DXplain: An evolving diagnostic decision-support system. *Journal of the American Medical Association, 258,* 1987, 67-74.

[2] Barr, C.E., Komorowsky, H.J., Pattison-Gordon, E., and Greenes, R.A. Conceptual modeling for the unified medical language system. In: R.A. Greenes (Editor), *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care,* Washington, DC: IEEE Computer Society Press, 1988, 148-151.

[3] Blair, D.C. and Maron, M.E., An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28,* 1985, 289-299.

[4] Cote, R.A. (Editor). *Systematized Nomenclature of Medicine (SNOMED).* Skokie, IL: College of American Pathologists, 1982.

[5] Cullum J.K., Willoughby R.A., and Lake, M. A Lanczos algorithm for computing singular values and vectors of large matrices. *SIAM Journal of Scientific and Statistical Computing, 4(2),* 1983, 197-215.

[6] Cullum, J.K. and Willoughby, R.A. *Lanczos algorithms for large symmetric eigenvalue computations, Vol. I Theory.* Boston, MA: Birkhauser, 1985.

[7] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6),* 1990, 391-407.

[8] *Dorland's Illustrated Medical Dictionary,* 26th *Edition.* Philadelphia, PA: W.B. Saundere Company, 1985.

[9] Evans, D.A. *Final Report on the MedSORT-II Project: Developing and Managing Medical Thesauri,* Technical Report No. CMU-LCL-87-3. Laboratory for Computational Linguistics, Carnegie Mellon University, 1987, vii-f113 pp., and appendices.

[10] Evans, D.A. Pragmatically-structured, lexical-semantic knowledge bases for unified medical language systems. In: R.A. Greenes (Editor), *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care.* Washington, DC: IEEE Computer Society Press, 1988, 169-173.

[11] Evans, D.A. Concept management in text via natural-language processing: the CLARIT approach. *Working Notes of the 1990 AAAI Symposium on "Text-Based Intelligent Systems'*,* Stanford University, March, 27-29, 1990, 93-95.

[12] Evans, D.A., Ginther-Webster, K., Hart, M., Lefferts, R.G., and Monarch, LA. Automatic indexing using selective NLP and first-order thesauri. RIAO '91, Autonoma University of Barcelona, Barcelona, Spain, April 2-5, 1991, 624-644.

[13] Forsythe, G.E., Malcolm, M.A., and Moler, C.B. *Computer Methods for Mathematical Computations.* Englewood Cliffs, N.J: Prentice-Hall, 1977.

[14] Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. The vocabulary problem in human-system communication. *Communications of the ACM, 30(11),* 1987, 964-971.

[15] Golub, G.H., and Reinsch, C. Singular value decomposition and least squares solutions. In: Wilkinson, J., and Reinsch, C. *Linear Algebra* New York, NY: Springer-Verlag, 1971.

[16] *HICDA-2, Hospital Adaptation of ICDA, 2?^{d} Edition.* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1968.

[17] Humphreys, B.L., and Lindberg, D.A.B. Building the unified medical language system. In: *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.* Washington, DC: IEEE Computer Society Press, 1989, 475-480.

[18] *International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM).* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1986.

[19] Lindberg D.A.B, and Humphreys, B.L. The UMLS knowledge sources: tools for building better user interfaces. In: R.A. Miller (Editor), *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care.* Washington, DC: IEEE Computer Society Press, 1990, 121-5.

[20] McCray, A.T. The UMLS semantic network. In: *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.* Washington, D.C.: IEEE Computer Society Press, 1989, 503-507.

[21] McCray, A.T., and Hole, W.T. The scope and structure of the first version of the UMLS semantic network. In: *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care.* Washington, D.C.: IEEE Computer Society Press, 1990, 126-130.

[22] Miller, R.A., Pople, H.E., and Myers, J.D. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine, 301,* 1982, 468-476.

[23] Miller, R.A., Masarie, F.E., and Myers, J.D. Quick medical reference for diagnostic assistance. *MD Computing, 3,* 1986, 34-48.

[24] Miller, R.A., McNeil, M.A., Callinor, S., Masarie, F.E., and Myers, J.D. Status report: The INTERNIST-I/quick medical reference project. *Western Journal of Medicine,* December, 1986.

[25] Thomas, C.L. (Editor), *Taber's Cyclopedic Medical Dictionary, 15th Edition.* Philadelphia, PA: F.A. Davis Company, 1985.

[26] Warner, H.R. HELP—An approach to hospital-wide artificial intelligence. In: G.E. Statland and S. Bower (Editors), *Computer Assisted Decision Making Using Clinical and Paraclinical (Laboratory) Data.* Tarrytown, NY: Mediad, Inc., 1980.