# Talking about AI: Socially-Defined Linguistic Subcontexts in AI

**Amy M. Steier** and **Richard K. Belew**
Cognitive Computer Science Research Group
Computer Science & Engr. Dept. (0114)
University of California - San Diego
La Jolla, CA 92093
{steier,rik}@cs.ucsd.edu

## Abstract

This paper describes experiments documenting significant variations in word usage patterns within social subgroups of AI researchers. As some phrases have very different collocational patterns than their constituent words, we look beyond occurrences of individual words, to consider word phrases. The mutual information statistic is used to measure the information content of phrases beyond that of their constituent words. Previous research has shown that some phrases are much more informative as word pairs outside *topically* defined subsets of a document corpus than within it. In this paper we show that individual universities provide an analogous, *socially* defined context in which locally-used phrases are "exported" into general AI vocabulary.

## Introduction

An increasing body of research relating free-text information retrieval (IR) techniques with methods in "corpus-based" computational linguistics is providing AI with a new approach to natural language understanding and simultaneously with an important new problem domain. Rather than attempting to syntactically analyze and then derive a deeper semantic understanding of each and every sentence in a collection of natural language documents, these methods use statistical characteristics of word token occurrences to form (at least gross) characterizations of what the documents are "about." Further, it appears that that such statistical methods may be *complimentary* with more traditional NLP methods. For example, morphological and syntactic methods can provide more appropriate elements for statistical analysis; conversely, the resulting statistics can help to guide semantic interpretation.

There are currently a number of major simplifications typical of the IR approach that must be refined, however. This paper reports on attempts to address two of these. First, we extend the basic IR statistical methods beyond single word tokens to consider multiword phrases. In this respect, we are consistent with a number of others in corpus-based linguistics. Our next section discusses some of the important issues involved in using phrases rather than single words.

Our second variation is less common. Rather than treating the entire textual corpus as a single, homogeneous collection we consider variations in the statistics in smaller, related subsets of documents. It is a truism that the *context* of a any linguistic utterance obviously has great influence on its interpretation. This is especially true in IR, where particular "relevant" keywords are selected from a text by virtue of differences in their statistics in that text, relative to the *context* of corpus-wide norms. The larger a collection, the more likely it is to have been generated by many authors, addressing diverging audiences, using very different vocabularies, and the less likely it is to be topically cohesive. We hypothesize that if smaller subsets of documents within the larger corpus are analyzed separately, these will reveal useful "local" variations in word usage patterns that will be lost when statistics are gathered globally across the entire corpus.

In previous work, we have used *a priori* taxonomic classifications of the texts' topics as the basis for partitioning the corpus. For example, in a collection of legal texts (judicial opinions), human editors have manually classified each document according to a extremely refined taxonomy describing the full range of legal topics. Based on these characterizations we were able to separate documents about (for example) "Labor Law" from others about "Constitutional Law", etc. and find that, in fact, word usage patterns vary significantly from one topical area to another. Judges and lawyers working within a specialized area of the law use language among themselves in consistent, reliable but specialized ways, relative to "general" legal parlance. In particular, these variations can be used to identify *phrases* whose utility in topical sub-context is much different than it is across the entire collection. This work is summarized in the section on Previous Research.

Of course variations in the topical domain of discourse is only one dimension of linguistic context that

might interest us. In this paper, we attempt to consider *socially* defined sub-contexts within the larger context of artificial intelligence (AI). Using as our corpus a set of thesis abstracts all nominally about AI, we consider variations in single word and "bigram" (sequential word pairs) statistics arising within the context of particular *universities* at which the theses were written. We hypothesize that research groups at universities also define a useful linguistic sub-context. That is, they too use language among themselves in consistent, reliable ways that may not – yet – be shared with the rest of AI. The details and results of our experiment are described in the section titled Recent Experiments. We conclude with comments about relating this socially-defined notion of context with others.

## Phrasal Semantics

Simple phrases provide an attractive first step beyond the simple "bag of words" techniques generally associated with IR. Phrases are interesting in part because of the complex way in which meanings associated with constituent words are combined to form more elaborate semantic expressions. D.A. Cruse has defined a "transparent" expression to be one whose meaning is derived directly from that of its constituent words; and "opaque" phrases as those whose semantics cannot be attributed to the simple composition of its constituent terms (Cruse 1986). For example, whereas the meaning of **red herring** or **red carpet** is opaque, the meaning of **red paint** is much more transparent. Halliday has made similar distinctions between "simple," "compound" and "phrasal" lexical items (Halliday 1966). He writes that often a phrase can act very much like a lexical unit in and of itself, and that often a phrase can have very different collocational patterns than the sum of its parts. From a practical IR perspective, these linguistic issues become the question of just when indexing word compounds offers advantages over simpler indexing of individual words.

We follow other recent work that considers a minimalist notion of phrase, based on simple collocation (Maarek & Smadja 1989)(Fagan 1989)(Church & Hanks 1990). As we are most interested in simple word compounds (such as **social security**), we use "bigrams" (sequential word pairs) to identify potentially interesting word co-occurrences. Church and Hanks have shown that compounds have a very fixed word order and that the average separation is one word (Church & Hanks 1990). We therefore restrict ourselves to a window of one when parsing for bigrams.

Words have meanings derived from their use both independent of, and with respect to, use within a phrase. Conversely, the phrasal meaning is drawn from the meaning of the constituents as well as from the "use" of the phrase. As a phrase becomes more frequently used, we hypothesize that a phrase's meaning draws less on the experience of the constituents' uses in other contexts and more on the experience of the phrase in its particular context. In other words, less meaning is drawn from the "transparent" semantics of the constituents, while more meaning becomes related to the direct experience of using the phrase. *Mutual information* therefore becomes a very natural measure of the disparity between a phrase's use and the independent use of its constituent words:

$$MI(w_1, w_2) = log \frac{Prob(w_1, w_2)}{Prob(w_1)Prob(w_2)}$$

Here, $Prob(w_i)$ is the frequency of word $w_i$ divided by the size of the corpus $N_W$ and $Prob(w_1, w_2)$ is the frequency of bigram $(w_1, w_2)$ divided by $N_W$. $MI$ has often been used to measure lexical cohesiveness or strength of association between two words (Magerman & Marcus 1990; Hindle 1990; Church *et al.* 1991).

Because we will be considering sub-contexts of a large corpus, we will compute a phrase's mutual information with respect to both the entire corpus and a particular subset of documents. We will show that across the difference sub-contexts of a large corpus, the extent to which a phrase acts like a lexical unit varies, and in a way: if a phrase has high mutual information with respect to the entire collection, then it will have a depressed mutual information value with respect to a specific context if and only if at least one of the phrase's constituent words is independently very descriptive of the subcontext.

## Previous Research

This paper is a continuation of earlier research where we looked beyond occurrences of individual words, to consider word pairs or phrases (Steier & Belew 1993). The focus of this earlier research was to study how the informational content of phrases, beyond that of their constituent words, can change within topically restricted areas. The mutual information measure was used as the statistic best reflecting this value.

This previous research was performed using a large collection of judicial opinions, covering virtually all topics of U.S. case law, provided by West Publishing Company. West uses a rich hierarchical indexing scheme (the Key number system) to topically organize all case digests. We were thus able to define topical sub-contexts within the collection through the use of this indexing scheme. The headnotes [1] of these cases were grouped together by topic numbers to get a total of 339 topical subcontexts.

[1] Headnotes are precis generated by West's editors to capture the central points of each judicial opinion.

After some simple stemming, [2] all bigrams within each topical subcontext were extracted. Any bigram crossing punctuation marks, as well as any bigrams containing "noise words" from a list of common English words were not considered. It is important to note that our noise word list not only contained non-content words such as articles and prepositions, but words that are extremely common in a legal context. [3] Since the mutual information measure can become unstable when counts are very low, any bigram occurring less than three times within a topical subcontext or outside of it was also filtered.

For each bigram within a topical subcontext, its mutual information (cf. Phrasal Semantics section) with respect to the subcontext, $MI_t$, as well as a value with respect to the entire collection $MI_c$ was computed. In computing $MI_c$, $N_W$ is the total number of words in the entire collection. In computing $MI_t$, $N_W$ is the number of words in that particular topic area. The consequence of this measure is if the constituents of a phrase occur together much more often than chance, the phrase will have a high mutual information value. The higher the mutual information value is between a pair of words, the more informative that pair is as a phrase. It is not that the phrase is necessarily more contentful but that our interpretation of the phrase is less easily derived from the typical meaning associated with its constituents.

Next, each individual *word* within a topic area was given an index term weight based on a variant of the term frequency × inverse document frequency weighting scheme devised by Salton and Buckley [4] (Salton & Buckley 1988). If $F_{ij}$ represents the frequency of term $j$ in document $i$, $DF_j$, the document frequency of term $j$, $N_D$, the total number of documents, and $N_i$, the number of words in document i, then $TW_{ij}$, the term weight of term $j$ in document $i$, is given by the following formula:

$$TW_{ij} = \frac{F_{ij} \times \log(N_D/DF_j)}{\sqrt{\sum_{k=1}^{N_i}(F_{ik} \times \log(N_D/DF_k))^2}}$$

In studying how the mutual information of bigrams changes across different subcontexts, our most central finding was that a decrease in mutual information within a topic area turns out to by highly correlated to the maximum of the term weights associated with the bigrams constituents. In other words, bigrams with high informational content across the entire collection have depressed informational content (i.e. convey less information) – as bigrams – within their semantically related topical area *if, and only if* the bigram's constituents are good descriptors of that topic area.

| *PHRASE* | $TW_M$ | $MI_t$ | $MI_c$ |
|---|---|---|---|
| MINIMUM WAGE | .057 | 9.48 | 11.50 |
| WORKER COMPENSATION | .043 | 5.98 | 10.37 |
| COLLECTIVE BARGAINING | .185 | 7.80 | 12.08 |
| UNION MEMBER | .449 | 6.31 | 8.91 |
| OCCUPATIONAL SAFETY | .050 | 10.51 | 12.56 |
| OVERTIME PAY | .062 | 5.40 | 8.95 |
| LOCAL UNION | .449 | 6.03 | 8.17 |
| PENSION PLAN | .076 | 8.42 | 10.40 |
| LABOR RELATION | .495 | 4.99 | 8.84 |
| LABOR STANDARD | .495 | 5.78 | 9.20 |

Table 1: Example phrases where the mutual information within the collection ($MI_c$) exceeds that within the topic Labor Relations ($MI_t$.)

Table 1 shows some examples where $MI_c$ exceeds $MI_t$ for the topic Labor Relations. In studying these phrases, what stands out is how apropos to the Labor Relations topic the phrases in Table 1 are. $TW_M$ (maximum term weight) is the larger of the two term weights associated with the phrases constituents with respect to the topic Labor Relations. All term weights range between 0 and 1, and the median term weight for a topic is about .03. There is a strong tendency for phrases with high $TW_M$ to have a depressed mutual information value within the topic. The correlation between $TW_M$ and difference in MI ($MI_t - MI_c$) in the Labor Relations topic was -.62. This was found to be very typical across the other topics within the collection. This correlation has also been independently reported by Damerau (Damerau 1993).

The conclusion we draw from this phenomenon is that the more descriptive the constituents of a phrase are with respect to a specific topic, the more transparent that phrase is within the topic and the more opaque it is outside the topic. Take, for example, the phrase PENSION PLAN. This phrase seems to be much more opaque (i.e. have a much higher phrasal information content) across the entire collection. Yet within the Labor Relations topic, the constituents PENSION and PLAN are very descriptive of the central issues. Not only do these words occur together as PENSION PLAN, but they often occur separately (e.g., PENSION FUND, PENSION BENEFIT, PENSION BOARD, INSURANCE PLAN, WELFARE PLAN.)

Our interpretation of these results was that technical phrases are often "deconstructed" into their constituents within a particular topical sublanguage. The semantic nuances that are explored in detail within a topical area are then left behind as this phrase is "exported" into general vocabulary, where dominant use

[2] Our stemming consisted only of converting all plural nouns to their singular form.

[3] An example of this would be legal abbreviations of statute sections.

[4] This term weighting technique is currently the most widely used in I.R.

of the constituent words are as part of the phrase. Further, we found evidence of an intriguing "self-similar" regularity in this exporting relation. When we repeated the experiments using a topic subcontext as the "collection", and that topic's different sub-topics as the different subcontexts, we again found this exporting phenomenon to exist. This led us to conclude that "opacity" is a relative concept. An opaque phrase whose meaning appears to have little relation to that of its constituent words may in fact be transparent within some restricted sublanguage, perhaps associated with its genesis. Of course, this sublanguage may no longer exist, or it may simply require closer analysis of restricted portions of the general textual corpus.

## Recent Experiments

The textual corpus used in our current is a set of masters and Ph.D. thesis dissertations, collected via questionaires but especially from the University Microfilms database during the past 5 years. Title and abstract text for each of these approx. 2600 theses was combined, with each thesis comprising approx. 2000 bytes. [5] It is important to recognize that while this corpus is of obvious interest, it is very small by IR standards and we have had to tailor many of our statistics somewhat in order to handle the very small sample sizes involved. Subsets of the corpus were formed for each of the 100 universities most well-represented in the collection. This ranged from Stanford with 79 theses to Yale with only seven.

To begin our experiments, the noise word list is first updated to include terms that are extremely common in the context of AI. A noise term is a low information word or phrase within a particular context. In Information Theory, the information encoded in a symbol is relative to the uncertainty associated with that symbol occurring. In language, very frequent words (or phrases) provide us with little information. Less frequent words provide us with more information. To put it another way, information is a deviation from what you expect. It is a common practice within IR to maintain a "negative dictionary" of frequently occurring "noise words" in any language.

When a textual corpus is further restricted to a particular topic, additional words also become effectively noise. For example, in a collection of computer industry magazine articles, the word COMPUTER becomes almost (statistically) meaningless. It is therefore appropriate to augment the standard negative dictionary with those additional words that are effectively noise in a particular corpus. The top noise words with respect to our AI collection are listed in Table 2.

Next, all bigrams from each of university subcontexts are extracted. Any bigram that contains a noise word or occurs more than 100 times in the collection is discarded. These very frequently occurring phrases are also effectively "noise" when computing collection statistics. These phrases are listed in Table 3. Note how much more contentful these phrases are than the single words. This additional specificity in meaning is the central reason why phrases are being used more and more in document indexing.

A mutual information value is computed with respect to each university subcontext, $MI_u$, as well as with respect to the entire collection, $MI_c$. Each individual word within a subcontext is assigned an index term weight, using the term weighting formula mentioned earlier.

| *WORD* | $TF_c$ | *WORD* | $TF_c$ |
|---|---|---|---|
| SYSTEM | 6991 | BASED | 1357 |
| MODEL | 3328 | TECHNIQUE | 1319 |
| PROBLEM | 2754 | DATA | 1319 |
| KNOWLEDGE | 2734 | RESEARCH | 1293 |
| NETWORK | 2267 | USE | 1257 |
| EXPERT | 2062 | INFORMATION | 1253 |
| DESIGN | 2046 | NEURAL | 1231 |
| PROCESS | 1815 | DEVELOPED | 1188 |
| USED | 1693 | RESULT | 1175 |
| APPROACH | 1652 | ANALYSIS | 1130 |
| METHOD | 1605 | TWO | 1049 |
| LEARNING | 1527 | NEW | 1013 |
| USING | 1501 | COMPUTER | 982 |
| ALGORITHM | 1480 | APPLICATION | 979 |
| CONTROL | 1394 | PERFORMANCE | 963 |

Table 2: Top noise words in our AI collection. $TF_c$ is the term frequency within the collection.

| *PHRASE* | $TF_c$ |
|---|---|
| EXPERT SYSTEM | 1456 |
| NEURAL NETWORK | 1075 |
| ARTIFICIAL INTELLIGENCE | 632 |
| KNOWLEDGE BASE | 394 |
| PROBLEM SOLVING | 247 |
| ARTIFICIAL NEURAL | 175 |
| KNOWLEDGE-BASED SYSTEM | 160 |
| KNOWLEDGE ACQUISITION | 159 |
| CONTROL SYSTEM | 158 |
| KNOWLEDGE REPRESENTATION | 154 |
| LEARNING ALGORITHM | 145 |
| DECISION SUPPORT | 144 |
| NATURAL LANGUAGE | 142 |
| DECISION MAKING | 131 |
| PATTERN RECOGNITION | 126 |
| SUPPORT SYSTEM | 124 |
| MACHINE LEARNING | 120 |
| DISSERTATION PRESENT | 103 |
| DESIGN PROCESS | 102 |

Table 3: Most frequent phrases in our AI collection. $TF_c$ is the term frequency within the collection.

[5] Total number of words in our collection is 535804.

Our analysis is then to study how the mutual information ($MI$) of phrases changes within the individual university subcontexts. As in our previous experiments, we look at the correlation between the maximum term weight of a phrase's constituents and the difference in collection and university $MI$ ($MI_u - MI_c$). Once again, we find the "export" relation to exist. The average correlation between difference in $MI$ and maximum term weight is -.61 within the individual university subcontexts.

Figure 1 shows this relationship for all university subcontexts combined. Since our term weighting algorithm assigns to the majority of terms a very low weight, and then progressively less terms a higher and higher weight, we use the log of the maximum term weight to more aptly show its relationship to the difference in mutual information. The correlation of the two parameters in this graph is -.61.

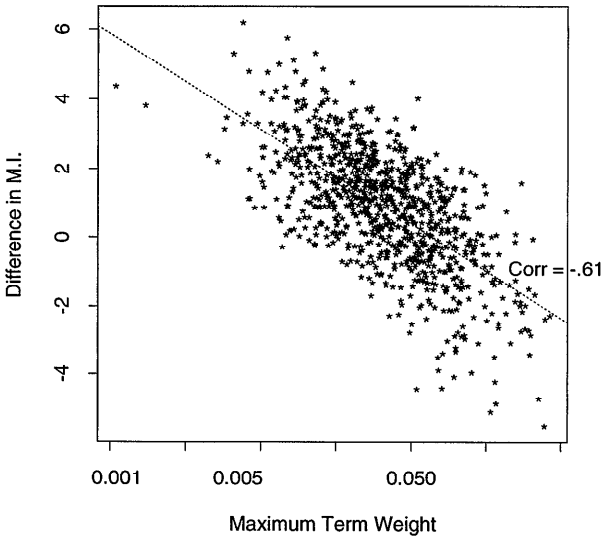


Figure 1: Log of the Maximum Term Weight to difference in $MI$ ($MI_u - MI_c$) for all university subcontexts. The dotted line in the graph maps the least squared fit.

Table 4 lists example bigrams which exhibit the export relation. The 35 universities from which examples are taken are those with the highest number of AI theses associated with them.

The conclusion we draw from this analysis, is that individual universities do indeed represent distinct linguistic subcontexts. More specifically, we find that the more descriptive the constituents of a phrase are with respect to a specific university's AI thesis research, the more transparent that phrase is within that university's context and the more opaque it is outside that context. Take, for example, the phrase **CASE-BASED REASONING**. This phrase seems to be very opaque (i.e. have a high mutual information value) across the entire context of AI. Yet within the a few of the individual university subcontexts (e.g. Yale, Georgia Tech. and the University of Massachusetts) the phrase's meaning is much more transparent. Within these particular subcontexts, the terms **CASE-BASED** and **REASONING** are highly weighted as being descriptive of their AI thesis research. The constituent concepts of **CASE-BASED** and **REASONING** are examined in detail and independently. Not only do these words occur together as **CASE-BASED REASONING**, but they often occur separately (e.g. **CASE-BASED PLANNING, CASE-BASED ARGUMENT, CASE-BASED PROBLEM, SCHEMA-BASED REASONING, ANALOGICAL REASONING, COMMONSENSE REASONING**). As this phrase is "exported" into the general AI vocabulary, the semantic nuances are left behind, and the dominant use of the constituent words is as part of the phrase.

## Conclusion

In studying the phrases that exhibit the export relation, we find occasional instances where the same phrases is exported from different universities. For example, the phrase **CASE-BASED REASONING** is shown to be an export phrase from the universities of Yale, Georgia Tech. and the University of Massachusetts. It is interesting to note that some of the students Schank advised at Yale, graduated and then took on teaching positions at both Georgia Tech. and the University of Massachusetts. We hypothesize that this type of AI lineage (or *genealogy*) information might be used to create yet another socially-defined set of linguistic subcontexts. We are currently in the process of gathering this type of AI genealogy information.

We interpret the export phenomena of important word combinations as a migration, from a context in which the nuances of each words' meaning is considered independently out into a broader context in which only the central and now "opaque" interpretation of the phrase is useful. If true, there must be a *time delay* to this progression, and a particular textual corpus may or may not capture it. In our previous experiments with legal texts, the opinions cover a very long time period, with ample opportunity for technical jargon (e.g., **PENSION PLAN** or **SOCIAL SECURITY**) to migrate from use by a restricted group of specialists to lawyers and judges generally.

It is somewhat curious, then, that even in the very short five-year window spanned by our AI theses should be sufficient time to also give evidence of phrase

| | | | |
|---|---|---|---|
| **STANFORD** | **UNIV. MASSACHUSETTS** | **UNIV. PENNSYLVANIA** | **UNIV. MISSOURI (ROLLA)** |
| BELIEF REVISION | PLAN RECOGNITION | HORN CLAUSE | DEVELOPING COUNTRY |
| INFLUENCE DIAGRAM | ROBOT ARM | END EFFECTOR | IMAGE SEGMENTATION |
| NONMONO. REAS. | CASE-BASED REAS. | **PENN. STATE** | **UNIV. CINCINNATI** |
| **UNIV. ILL. (URB-CHAMP)** | **UNIV.MARYLAND** | ARC WELDING | LP FORMULATION |
| MASSIVELY PARALLEL | HIDDEN UNIT | MOMENT INVARIANT | MOMENT INVARIANT |
| **PURDUE** | LOAD BALANCING | GARBAGE COLL. | **N.C. STATE (RALEIGH)** |
| COLLISION-FREE PATH | GRADIENT DESCENT | **UNIV. WISC. (MADISON)** | SIMULATED ANNEALING |
| MATERIAL HANDLING | **COUN. NAT. ACA. AWRDS** | ARC WELDING | IMAGE SEGMENTATION |
| MOBILE ROBOT | SPEECH RECOGNITION | WELDING ROBOT | CONSTRAINT SATISF. |
| **UNIV. TEXAS (AUSTIN)** | COLLISION AVOIDANCE | **UNIV. FLORIDA** | **UNIV. COLORADO (BOULDER)** |
| INFLUENCE DIAGRAM | FLEXIBLE MANUF. | OBJ.-ORIENTED D.B. | RIVER BASIN |
| TEXT UNDERSTANDING | **UNIV. MINNESOTA** | D.B. MANAGEMENT | BIOLOGICAL NERVOUS |
| **CARNEGIE-MELLON** | ASSOCIATIVE MEMORY | **UNIV. ARIZONA** | **UNIV. CALIF. (LOS ANGELES)** |
| CHESS PROGRAM | ANALOGICAL REAS. | EQUILIBRIUM POINT | NUCLEAR REACTOR |
| ABSTRACTION HIER. | **UNIV. MICHIGAN** | **UNIV. PITTSBURGH** | FAULT TOLERANCE |
| MOBILE ROBOT | BODY MOTION | NONE | **RUTGERS (NEW BRUNSWICK)** |
| **OHIO STATE** | RIGID BODY | **CASE WESTERN** | DEDUCTIVE DATABASE |
| MALF. DIAGNOSIS | DOCUMENT RETR. | OBJECT RECOGNITION | **UNIV. WASHINGTON** |
| CREDIT ASSIGNMENT | **NORTHWESTERN** | **UNIV. TEXAS (ARLING.)** | STATE-SPACE SEARCH |
| **UNIV. CALIF. (BERK.)** | DEDUCTIVE DATABASE | BUILDING BLOCK | MULTI-LAYER PERC. |
| SEMICOND. MANUF. | THEOREM PROVER | **RENSSELAER POLY. INST.** | STEADY STATE |
| TOOL WEAR | **ARIZONA STATE** | ROBOTIC ASSEMBLY | **MIT** |
| FUZZY LOGIC | SYSTOLIC ARRAY | ERROR RECOVERY | TOOL WEAR |
| **TEXAS A&M** | LOAD FORECASTING | **GEORGIA TECH.** | LIMIT CYCLE |
| OBSTACLE AVOIDANCE | LP FORMULATION | SENSOR FUSION | THESIS PRESENT |
| CIRCUIT BOARD | **ILLINOIS INST. TECH.** | CASE-BASED REAS. | **YALE** |
| | INTELL. TUTORING | | INTENTIONAL STATE |
| | | | CASE-BASED REAS. |

Table 4: Example bigrams where the mutual information in the university file ($MI_u$) exceeds that within the collection ($MI_c$). Universities are listed in order of number of AI theses associated with them.

exporting. Our explanation is that this is due to the speed with which technical, scientific language adapts, relative to more common forms. It does not seem unreasonable that through conference and journal papers, let alone "directly" through influential theses, language used successfully within one research group (e.g., **CASE-BASED REASONING**) can come to influence the AI community generally in just a few years. If true, it suggests that scientific (and other professional) communication might provide an especially illuminating perspective into fundamental properties of linguistic evolution.

## References


Church, K., and Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1).

Church, K.; Gale, W.; Hanks, P.; and Hindle, D. 1991. *Using Statistics in Lexical Analysis.* New Jersey and London: Lawrence Erlbaum Assoc. 115–164.

Cruse, D. 1986. *Lexical Semantics.* New York: Cambridge University Press.

Damerau, F. J. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management* 29(4):433–447.

Fagan, J. 1989. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science* 40(2):115–132.

Halliday, M. 1966. *Lexis as a Linguistic Level.* London: Longmans.

Hindle, D. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics.*

Maarek, Y., and Smadja, F. 1989. Full text indexing based on lexical relations. In *Proceedings of the 12th International Conference on Research and Development in Information Retrieval.*

Magerman, D., and Marcus, M. 1990. Parsing a natural language using mutual information statistics. In *Proccedings of AAAI '90.*

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5):513–523.

Steier, A., and Belew, R. 1993. Exporting phrases: A statistical analysis of topical language. In *Second Annual Symposium on Document Analysis and Information Retrieval.*