Cognitive Science Online, Vol.1, pp.46–57, 2003 http://cogsci-online.ucsd.edu

Lexical dynamics and conceptual change: Analyses and implications for information retrieval

Robert Liebscher & Richard K. Belew
Department of Cognitive Science
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0515
rliebsch|rik@cogsci.ucsd.edu

Abstract

One important aspect of a document's context is the time at which it was written. We report here on analyses of formal (dissertation abstracts) and informal (discussion board postings) communications among academics within two separate disciplines. We focus on academic communications because these especially must be understood within the context of what has been said before, together with what is considered relevant and worth saying at the time of publication. All corpora include time-stamp information that allows temporal analysis of changing lexical frequencies across decades. Using techniques borrowed from time series analysis, we find distinct patterns of "rising" and "falling" bigram frequencies in both domains, and argue that this information can be exploited to improve retrieval of relevant documents.

1 Introduction

A common idealization of many information retrieval (IR) tasks is to reduce retrievable documents and queries to simply the sets of lexical keywords they contain. But it is becoming increasingly acknowledged that if IR systems are ever to qualitatively improve their ability to help users seek relevant documents, increased attention must be paid to the *context* shaping use of these keywords by both the browsing user and the documents' original authors. In many situations, one obvious feature of both users' and authors' contexts is their respective places in time. As typically addressed, the retrieval task makes almost no assumptions about time. In general, the time frame in which an author commits her thoughts to writing, as well as the time frame within which the querying user operates, are either assumed to be irrelevant to effective retrieval, or tacitly assumed to be roughly contemporaneous with one another.

But the conceptual frameworks within which authors write can play an enormous role in what they choose to say, and how they choose to say it. This is especially true in scientific writings, where acceptance by peer reviewers is essential to success and recognition. By the same token, scientists and students searching historically through the body of work of a science will often be familiar with a different vocabulary as they pursue a more modern agenda.

Within the field of IR system design, our reflection of the changing semantic understanding

of a science is potentially informed by observation of word frequency statistics combined with metadata capturing the document's publication time. Words and phrases change in both frequency of use and meaning through time. For example, the token WEB is found in many areas of discourse today, though little more than a decade ago, prior to the advent of the World Wide Web, its use was much more circumscribed.

In this paper, it is assumed that within a given domain, the overall frequency of a term k at time t is proportional to a community's collective "interest" in k at t. As interest in a topic waxes or wanes, the raw *number* of documents containing information about that topic rises or falls through time. These arguments, along with observations about the nature of scientific discourse, are used to construct a temporal weighting scheme that places a document in an appropriate historical context.

Consider a "rising" term k, which moves from obscurity to immense popularity over the duration of a corpus. In an academic context, it is not unreasonable to assume that the term was once part of a small group of "seminal papers" that helped to launch a field of inquiry, a technology, a methodology, etc. (To be concrete, imagine a search for the now ubiquitous term DNA. We would surely want Watson & Crick's one-page paper of 1953 to be deemed relevant!) Under an atemporal paradigm, a query for k will return a temporally random subset of documents in the corpus, leaving our user in the dark with regard to any notion of the conceptual development of her query term. The seminal papers have a chance of being deemed relevant that is proportional only to their length.

But with the temporal weighting scheme introduced in Section 3, when the frequency of this rising term k is initially low (i.e. used in very few documents), its weight will be amplified. At a later point in time, when its use is much more common, its weight will be dampened so as not to over-emphasize the many documents about k that exist at that time.

Alternatively, imagine a "dying" term, where k is omnipresent at one point in time, then falls in frequency until it is no longer used. Under a temporal weighting scheme, its initial use is dampened, and its later use is amplified. One consequence of this would be to emphasize *historical* documents that are written retrospectively about the term in question. These provide a good starting place for someone who wishes, as above, to gain an historical perspective on the development of k.

This paper reports attempts to formalize these arguments and improve information retrieval by incorporating the time at which a document was written into the retrieval process.

2 Methods

2.1 Corpora

The analysis will concentrate on abstracts from Doctoral and Masters' dissertations because these are available across decades in a relatively consistent format. Further, the focus is placed on a particular discipline, artificial intelligence, in order to relate the analysis of changing frequency statistics to the semantics of the evolving science that generates them.

While this is terrifically rich data, the amount of text provided by only dissertations' titles and abstracts is not great. This is especially unfortunate, because standard time series analysis demands large data volumes. However, as will be demonstrated, much information can be gained from an analysis that employs the most straightforward linear models of trend.

Three corpora were used in these studies. The first, *AIT*, contains approximately 5,000 abstracts from Ph.D. and Masters theses in artificial intelligence, collected by University Microfilms, Inc. from 1986 to 1997 (Belew, 2000). Each document is labeled with its year

of publication.

The second corpus, *CommDis*, also from University Microfilms, Inc. contains approximately 4,000 abstracts from Ph.D. and Masters theses in language and communicative disorders. The abstracts run from 1980 to 2002, and each is labeled with its year of publication.

The third corpus, *AIList Digest* (hereafter AIList), is a subscription-based electronic newsletter that contains over 10,000 discussion board postings, conference announcements, and essays on artificial intelligence that were collected and distributed weekly from 1983 to 1988. Each document is labeled with the exact time and date on which it was written. As the intent of this study was to treat AIList as a record of informal academic communication, some documents, such as bibliographies and subscription statistics, needed to be removed. A very simple type/token ratio filter worked well to preserve relevant articles of discourse while filtering out the unwanted documents, which generally contained few grammatical terms relative to the total number of tokens.

2.2 Tracking lexical frequency change

Bigram frequency counts were made for each of the 12 years in AIT, for each of 12 bins of roughly equal size (approximately 110,000 tokens each) in AIList, and for each of the 23 years of CommDis. Statistics were also collected on unigrams, but for the purposes of this work, bigrams provide a much greater level of detail and are considered better descriptors (and therefore more likely to serve as query terms) in domain-specific corpora (Damerau, 1993). Unigrams of particular interest are abbreviations and acronyms, which are discussed in Section 5.

The task of modeling the frequencies of terms that have consistently increased or decreased through time lends itself well to formal methods in time series analysis (Box et al., 1994). However, the 12 data points associated with AIT and AIList are too sparse to allow accurate modeling, and even the 23 points associated with CommDis only approach data sufficiency. For this reason, the analysis here is restricted to simple *linear* models of trend.

Mean smoothing was performed over the temporal frequency plot of each term that met a minimum frequency requirement, each bin being averaged with its two neighbors. Smoothing is especially necessary for AIList, as an extended thread of conversation or long essay might cause a spike in the frequency of a particular term that does not accurately reflect the level of community interest at that time.

Each smoothed frequency plot was then fit with a regression line, and the adjusted correlation coefficient (between time and frequency) r and slope of the line b were measured. The slope b of a term's temporal frequency plot is only a meaningful number when compared to the slopes of other terms. Terms which steadily increased in frequency through time will have positive slopes; those which steadily decreased will have negative slopes. Terms with a slope of $2s$ rose (or fell) twice as quickly as those with a slope of s.

Figure 1 shows the temporal frequency plot of two examples drawn from each corpus. Terms that met a minimum threshold for r (0.70) and absolute value of b (corpus-specific)[1] were extracted for further study. These included 49%, 42%, and 28% of the bigrams, respectively, for AIT, AIList, and CommDis, and represent the terms that might benefit from a temporal analysis of frequency.

Table 1 depicts for the top ten "rising" and "falling" terms extracted from the AIT corpus, along with their b and r values. While the frequencies of many terms within a domain

[1] Careful manual examination of the data led to the choices of 5.0, 4.0, and 7.0 for threshold values of b for AIT, AIList, and CommDis, respectively.

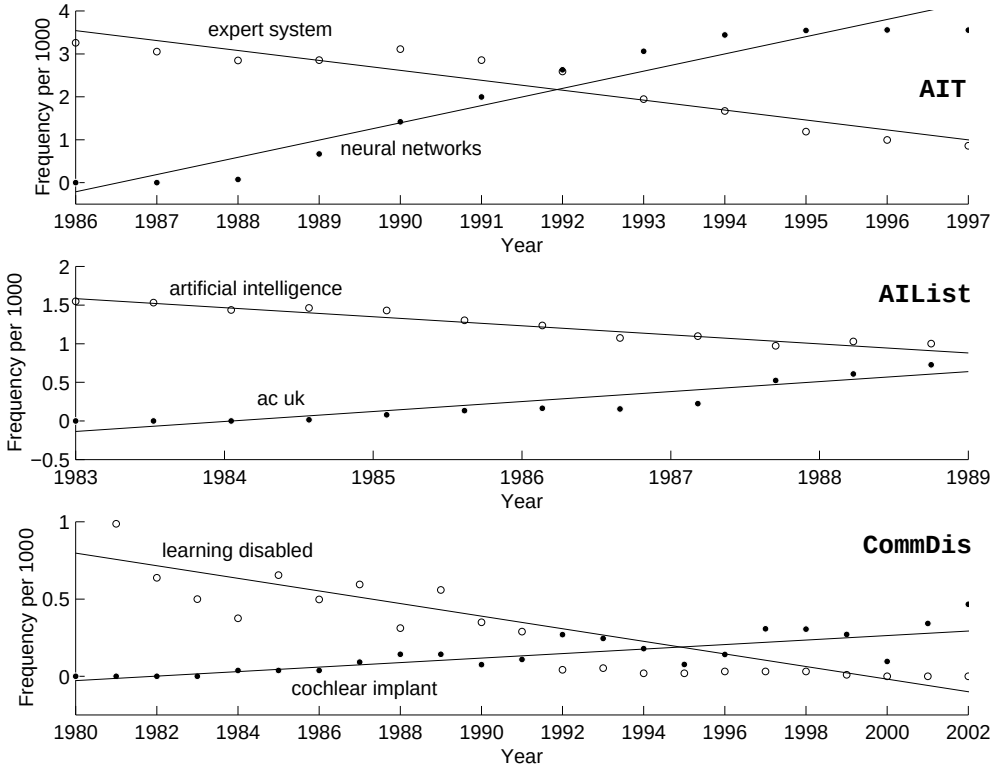

Figure 1: Examples of rising and falling terms from the three different corpora, with regression lines. Top to bottom: AIT, AIList, CommDis.

are temporally invariant or random, there are many others that undergo large frequency changes over time. In the traditional retrieval task, this information is not exploited.

Appendix A contains similar tables for AIList and CommDis. The results from CommDis dissertations provide strong confirmation that the analysis of conceptual change within the domain of AI transfers well to this new domain, despite large differences in the time frame and rate of dissertation publication across the two corpora. The results of the AIList newsgroup analysis are more mixed. While some of AIList's changing bigrams do indeed overlap semantically with those found in AIT, others appear to be "noise." These are most likely a consequence of both small AIList corpus size and difficulties in cleanly parsing the highly variable news postings (e.g., identifying common "signature lines" used by frequent posters). These anomolies may, however, also point to qualitative features of informal language use within such groups that limit the utility of the methods being used.

3 Temporal term weighting

The simplest form of **TF-IDF** weighting multiplies the raw **term frequency** (TF) of a term in a document by the term's **inverse document frequency** (IDF) weight:

$$idf_k = \log\left(\frac{NDoc}{D_k}\right) \quad (1)$$

Term	Slope (*b*)	*r* value
NEURAL NETWORK	474.32	0.9283
NEURAL NETWORKS	384.24	0.9505
FUZZY LOGIC	120.88	0.9035
ARTIFICIAL NEURAL	95.15	0.8990
GENETIC ALGORITHMS	55.14	0.9624
REAL WORLD	42.31	0.8509
REINFORCEMENT LEARNING	36.20	0.8447
PATTERN RECOGNITION	35.71	0.9314
NONLINEAR SYSTEMS	32.50	0.9511
LEARNING ALGORITHM	32.17	0.8313
ARTIFICIAL INTELLIGENCE	-248.07	-0.9309
EXPERT SYSTEM	-222.33	-0.9241
EXPERT SYSTEMS	-194.42	-0.9769
KNOWLEDGE BASED	-125.99	-0.9144
PROBLEM SOLVING	-90.48	-0.9490
KNOWLEDGE BASE	-73.65	-0.9281
KNOWLEDGE REPRESENTATION	-61.18	-0.9603
DECISION MAKING	-51.31	-0.9521
BASED EXPERT	-43.08	-0.9846
RULE BASED	-41.74	-0.9636

Table 1: Top ten rising and falling bigrams from AIT (1986-1997). Informal queries of AI practitioners revealed that the terms in these lists matched well with their memory of developments in the field over the period in question.

$$w_{kd} \quad = \quad f_{kd} \cdot idf_k \tag{2}$$

where f_{kd} is the frequency with which keyword k occurs in document d, $NDoc$ is the total number of documents in the corpus, and D_k is the number of documents containing keyword k.

In a temporal model, the corpus is divided into a series of independent sub-corpora, each associated with documents occurring within a particular time slice. IDF weights can then be computed independently for each sub-corpus.

One way to characterize the change is to contrast the temporally changing weights we propose as a *difference* with the traditional IDF weighting:

$$\Delta idf \quad = \quad \log\left(\frac{NDoc_k(t)}{D_k(t)}\right) - \log\left(\frac{NDoc}{D_k}\right) \tag{3}$$

$$= \quad \log\left(\frac{\frac{NDoc_k(t)}{D_k(t)}}{\frac{NDoc}{D_k}}\right) \tag{4}$$

$$\simeq \quad \log\left(D_k/D_k(t)\right) \tag{5}$$

where $D_k(t)$ is the number of documents in time slice t that contain k. Figure 2 shows how a keyword that changes in frequency over time can effect weighting. Following the assumption of a linear model in Section 2.2, a linearly increasing keyword is also assumed here. Note that the changes to the multiplicative IDF factor swing dramatically. With a keyword that increases in frequency over time, early uses of the term cause positive contributions, while later uses produce decreased contributions. The same argument holds for a term that decreases in frequency over time.

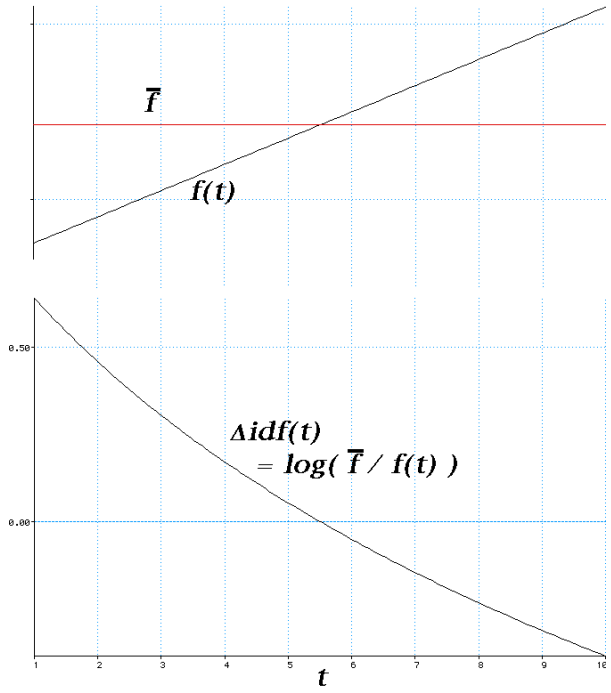

Figure 2: Dynamic IDF

Corpus	$\overline{r}(TF(t), D(t))$	$\overline{r}(TF(t), f(t))$
AIT	0.8899	0.3038
AIList	0.8885	0.3383
CommDis	0.8175	0.4007

Table 2: Correlation coefficent ($\overline{r}$) values for the three corpora.

Two assumptions must be made explicit before proceeding. First, f_{kd}, the number of times a term k occurs in document d, is only expected to correlate with the document's length. The average number of times that a document of length l mentions a term k, then, does not vary with respect to time. Second, documents of the same type within a domain do not become longer or shorter, on average, over time.

If these assumptions are true, then $TF_k(t)$, the total frequency of k in time slice t, should be proportional to $D_k(t)$, a rough measure of collective community interest in k at time t. This means that the following should hold:

$$log(D_k/D_k(t)) \simeq log(TF_k/TF_k(t)) \tag{6}$$

Table 2 shows that the number of documents in which a term appears at time t can be approximated by the total term frequency, as measured by the correlation coefficient $\overline{r}$ averaged over all terms. Note the striking similarity between $TF_k(t)$ and $D_k(t)$. Furthermore, the correlation between total term frequency and within-document frequency is low[2], supporting our first assumption above. f_{kd} is a constant over all t, as the burden of "temporal

[2] $TF(t)$ and $f(t)$ are not entirely *un*correlated ($\overline{r} = 0.0$) because there are some very infrequent terms in each corpus that have one document written specifically about them. These terms spike in

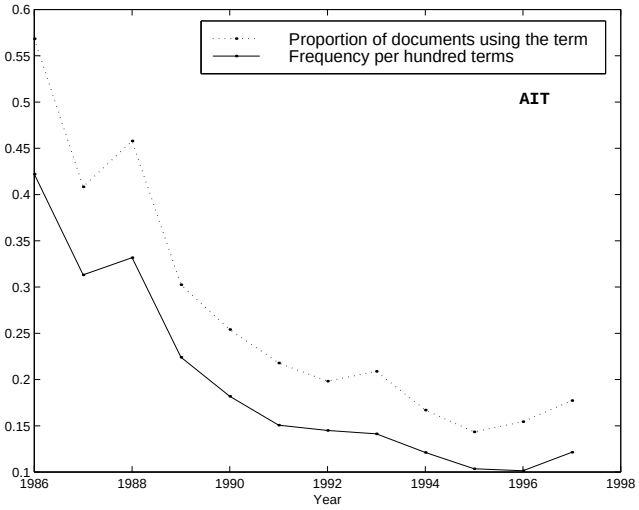

Figure 3: $TF_k(t)$ and $D_k(t)$ for the term `ARTIFICIAL INTELLIGENCE` over the duration of the AIT corpus. $TF_k(t)$ is frequency per hundred terms for scaling purposes.

culpability" in determining $TF_k(t)$ falls squarely on the weighting factor it approximates, $D_k(t)$. Figure 3 demonstrates this for the term `ARTIFICIAL INTELLIGENCE`; this example is discussed in greater depth in Section 5.

One virtue of using the techniques of time series analysis is that an estimate of what temporal perturbation is appropriate with respect to a particular document benefits from the more stable statistics of the entire series. Further, the fact that simple parametric models are fit means that these new weighting factors can be computed sufficiently at run/retrieval time. Linear models may be most appropriate for the corpora used in this study, but the same arguments apply to more elaborate (e.g. exponential; see Section 4.2) models of lexical frequency change as well.

4 Related work

4.1 Topic Detection and Tracking

Within the IR community, Topic Detection and Tracking (TDT) represents the most focused attention to special features in dynamic textual corpora. Research in TDT has focused primarily on the identification of *events* as reported in the news. Authors (e.g., newswire reporters) generate text capturing their reactions to, and interpretations of, "external" events arising in the world. There are obviously many commercial and security applications which motivate the specific TDT applications (e.g. story segmentation, new event detection, and event monitoring).

In part, the present work is motivated by the recognition within the TDT community that these methods must be extended:

> The results that we have presented [in this review paper] on the three

within-document frequency in the bin containing their signature document, while remaining flat over the rest of the corpus, and this is reflected in r. This is not very useful here, as the aim is to find terms that follow nonrandom temporal trends.

> detection tasks were acceptable, but not as high a quality that we would have liked. We believe that we have hit the limits of the effectiveness that can be reached with simple IR based approaches to story/topic comparison (Allan et al., 2002).

The approach of the current study has been to attempt to model patterns of change in the lexicon across time, and then more directly introduce the consequences of these changes. To this end, it is reasonable to believe that the methodology of TDT can be extended by considering the sociocultural contexts in which documents are written.

There are several ways in which the data used in TDT can be expected to differ from that in the corpora introduced in this study. Most obviously, the time scale differs dramatically. The AIT, AIList, and CommDis corpora contain documents written over the course of decades, while (for example) the TDT-2 corpus is collected over the first six months of 1998. Also, the TDT-2 corpus benefits from relevance assessment labels from 200 different topics on individual stories, whereas no such labels exist for the individual documents in the artificial intelligence and communication disorders corpora.

But there are more subtle factors influencing changes in the corpora as well. The *conceptual movement* of central interest in scientific discourse can be argued to be much more internally driven by the literature itself. Certainly there are "events" that ripple through scientific disciplines, but the amount of analysis, interpretation and theorizing is considerably greater. Scientific discoveries, when transformed into text, help to establish the next line of inquiry (Latour, 1986). In this way, the real world and the corpora continually influence one another. This places limits on a purely statistical approach to studying the corpora.

4.2 Recent models of trend

An interesting attempt within the TDT community to examine this dialectic is a study of the relations between financial news stories and stock prices by Lavrenko et al. (2000). They used piecewise linear regression to identify trends in a stock's price, then constructed language models from stories about the stock that preceeded gains or declines. The models were used to train a system to predict changes in price based upon the words used in an incoming stream of news stories. As investors spend a significant amount of time researching potential trades, as well as watching the fluctuations in a stock's price, the mutual influence of text and the outside world is obvious. Their system significantly outperformed a random trading simulation.

A key difference between Lavrenko et al. (2000) and the present work is, once again, time scale. As Web-based financial publishing and stock trading happen on the scale of minutes and hours, compared to months and years for scientific publishing and experimentation, there may not be reason to expect similar methods to apply in discovering trends. Piecewise linear segmentation is necessary for the financial study, and will likely play a role in the future work discussed in Section 5. In the present work, many terms followed a fairly linear increasing or decreasing trend in frequency, and these trends are likely a by-product of the particular time periods covered by the corpora. Many terms rose and fell over the duration of a corpus (particularly CommDis, which spans 23 years), and as such were accorded a poor goodness-of-fit to a single line spanning the corpus' length.

A recent attempt to formulate a methodology that works on any time scale is that of Kleinberg (2002). His model posits an infinite state automaton that takes advantage of the observation that, independent of context or type of discourse, terms generally occur in "bursts". That is, when a term is seen, its likelihood to appear again soon increases. Informally, the model can be described as performing piecewise exponential curve fitting. In one of several examples tested with the model, terms from conference paper titles spanning several decades were identified during periods of exponential growth in frequency as a way to find

the terms "... that exhibit the most prominent rising and falling pattern over a limited period of time."

Other cases examined in Kleinberg (2002) include modeling arrival times of e-mails containing a particular term concerned with an event in the world (such as the deadline for a grant proposal) and modeling term usage in U.S. Presidential State of the Union addresses over the course of two centuries. Besides spanning different time frames, another advantage of Kleinberg's model is its independence from requirements of formal time series analysis. It was mentioned earlier that the corpora used in the present study reveal a data sufficiency problem, when viewed as a problem for standard ARIMA-style modeling (Box et al., 1994).

Given this, a comparison of piecewise linear and exponential approaches to discovering trends in term usage seems in order. The main argument against a linear approach is data sufficiency, while a potential argument against an exponential model is that, while it is able to handle bursts of increasing intensity, it is difficult to imagine (outside of news corpora) the possibility of "reverse burstiness", i.e. a term exhibiting exponentially decreasing frequency. These are empirical matters to be investigated, along with the others described in the following section.

5 Conclusions and future work

Benefits to information retrieval, particularly for a user who is inexperienced in a particular domain, can result from paying attention to changes in term frequencies through time. In both formal and informal scientific communications, simple linear trends fit the temporal frequency curves for many terms, some of which showed dramatic rises or falls. Based upon the assertion that an academic community's "collective interest" in a topic at a given time is proportional to the frequency at which the topic is mentioned, a temporally-sensitive alteration to the standard (atemporal) TF-IDF weighting scheme was suggested. This allows documents, particularly within academic communication, to be placed in their proper historical context. The recent availability of collections of conference proceedings such as ACL and ACM-SIGIR promise to provide much more complete evidence, and allow more elaborate models.

Kleinberg has commented with regard to the burstiness of terms, and indeed most events, that our memory is structured with respect to these bursts in such a way that "... particular events are signaled by a compression of the time-sense" (Kleinberg, 2002). One interpretation of this statement is that readers recall historical periods in terms of a *Zeitgeist*, not a fluid timeline. For example, 1982 is sometimes referred to as the "year of the computer", when it was named Time Magazine's "Person of the Year"; similarly, the 1990s are sometimes described retrospectively as the "decade of the brain". Short of capturing features of popular culture's *Zeitgeist*, a potentially interesting way to create a *gold standard* for studies such as these is to collect data from scientific participants in the fields of artificial intelligence, communicative disorders, and (after the ACL corpus has been analyzed) computational linguistics. Future experiments will investigate the extent to which the present analyses are consistent with their historical recollections of the fields.

The behavior of the term `ARTIFICIAL INTELLIGENCE` in the AIT and AIList corpora suggests other linguistic and sociocultural interpretations of the data. One contributing factor seems to be the "substitution" phenomenon of the full phrase `ARTIFICIAL INTELLIGENCE` with its *acronym* `AI`. Within the AIList corpus, the decrease in frequency of `ARTIFICIAL INTELLIGENCE` is accompanied by a rise of the acronym `AI`, which was the sixth fastest rising unigram. Similarly, `COMPUTER SCIENCE` experienced the sixth fastest decrease, while `CS` was the thirteenth fastest riser.

The same patterns do not hold true in AIT, perhaps suggesting an important difference between informal (AIList discussion board postings) and formal (AIT dissertation abstracts) as channels of academic communication. Schwartz and Hearst (2003) have recently developed a fast algorithm for identifying abbreviations and acronyms and their full-phrase referents in the biomedical literature, which is presently experiencing explosive growth both in volume and coinages of new terms and abbreviations. A temporal examination of the rise of different acronyms and fall of their referents in different domains is an obvious next step.

Another contributing factor may involve the phenomenon of "internal vs. external keywords." Speaking of a token frequency analysis of the same AIT corpus:

> ... the statistics for stemmed, non-noise word tokens, noise words (e.g. `the`) are [combined]. As expected, the noise words are very frequent. But it is interesting to contrast those very frequent words defined *a priori* in the negative dictionary with those that are especially frequent in this particular [AIT] corpus. In many ways these are excellent candidates for *external keywords*: characteristizations of this corpus' content, from the "external" perspective of general language use. That is, these are exactly the words (cf. `NEURAL NETWORK, BASE, LEARN, WORLD, KNOWLEDGE`) that could suggest to a browsing WWW user that the AIT corpus might be worth visiting. Once "inside" the topical domain of AI, however, these same words become as ineffective as other noise-words, as *internal keywords*, discriminating the contents of one AI thesis dissertation from the next (cf. `SYSTEM, MODEL, PROCESS, DESIGN`). (Belew, 2000, pp. 71,72)

Considered from a temporal perspective, it seems that after a period of using the term, the practitioners of artificial intelligence began to assume its use to be *tacit* at least in some cases, and as such did not make explicit reference to it. Just how long this process takes, which factors influence this rate, etc., all become potentially interesting questions.

Also at work is the phenomenon of lexical replacement not by an acronym but by another term that, in some situations, becomes interchangeable with the original term (Bauer, 1994; Howard, 1977). As `ARTIFICIAL INTELLIGENCE` fell in AIList and AIT, `MACHINE LEARNING` rose in both corpora, and in fact was the eleventh fastest rising term in AIT. A more conspicuous example comes from the CommDis corpus, where the fastest falling unigram was `SUBJECT`, while the fastest riser was `PARTICIPANT`.

One can also imagine this analysis being extended to consider the differential flow of conceptual markers through various publication forums. For example, many of the same rising and falling bigrams identified within the more formal and chronologically later AIT dissertation corpus seem to be *anticipated* by similar usage patterns within the more informal and earlier AIList newsgroups. While it seems reasonable to expect such linguistic patterns to be "worked out" within informal contexts prior to more formal adoption by a community, insufficient data with respect to the AIT and AIList corpora make it premature to reach this conclusion.

Looking ahead still further, investigations of *groups* of terms that rise and fall together in time, as well as conceptual movements among such systems of vocabulary, are in order. Combined with holistic analyses of entire vocabularies (e.g., eigen-structure analyses such as Latent Semantic Indexing (Deerwester et al., 1990)), such methods may provide some of the most solid empirical foundations for an understanding of conceptual change within scientific communities. Almost certainly, knowledge about the nature of language change should support improved performance by temporally-informed retrieval and classification systems.

References

Allan, J., Lavrenko, V., & Swan, R. (2002). Explorations within topic tracking and detection. In J. Allan (Ed.), *Topic detection and tracking: Event-based information organization* (p. 197-224). Kluwer Press.

Bauer, L. (1994). *Watching english change.* London: Longman Press.

Belew, R. K. (2000). *Finding out about: A cognitive perspective on search engine technology and the world wide web.* Cambridge University Press.

Box, G., Jenkins, G., & Reinsel, G. (1994). *Time series analysis: Forecasting and control.* Prentice Hall.

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

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

Howard, P. (1977). *New words for old.* Hamish Hamilton.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *8th ACM SIGKDD international conference on knowledge discovery and data mining.*

Latour, B. (1986). Visualization and cognition: Thinking with eyes and hands. In *Knowledge and society: Studies in the sociology of culture past and present* (Vol. 6, p. 1-40). JAI Press, Inc. (transformation of rats and chemicals into paper.)

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). Mining of concurrent text and time series. In *6th ACM SIGKDD international conference on knowledge discovery and data mining, text mining workshop* (p. 37-44).

Schwartz, A., & Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing.*

Term	Slope (*b*)	*r* value
AILIST DIGEST	206.65	0.9535
VOLUME ISSUE	142.95	0.9762
AC UK	92.82	0.9371
MIT EDU	81.99	0.9470
NEURAL NETWORKS	52.33	0.8482
STANFORD EDU	42.35	0.9798
NEURAL NETWORK	41.54	0.8879
NEURAL NETS	33.52	0.8580
AI AI	32.50	0.8302
AI MIT	32.07	0.8971
NATURAL LANGUAGE	-85.41	-0.9402
EXPERT SYSTEMS	-71.76	-0.6261
ARTIFICIAL INTELLIGENCE	-71.34	-0.8766
EXPERT SYSTEM	-58.61	-0.3878
LOGIC PROGRAMMING	-54.87	-0.5980
COMPUTER SCIENCE	-52.78	-0.8242
SRI AI	-51.80	-0.9624
TURBO PROLOG	-37.48	-0.8752
PROGRAMMING LANGUAGE	-36.12	-0.7959
CSNET RELAY	-35.88	-0.9629

Table 3: Top 10 rising and falling terms in the AIList corpus.

Term	Slope (*b*)	*r* value
HEARING LOSS	60.5875	0.849
PHONOLOGICAL AWARENESS	43.2302	0.865
WORKING MEMORY	38.9826	0.951
HEARING AID	38.8868	0.456
SPEECH LANGUAGE	25.4638	0.490
CHILDREN SLI	23.1939	0.573
FORMANT TRANSITIONS	20.4096	0.613
HEALTH CARE	19.7127	0.611
SPEECH RECOGNITION	18.5384	0.651
COCHLEAR IMPLANTS	18.1522	0.831
HEARING IMPAIRED	-56.6631	-0.837
SB RM	-46.9355	-0.711
LANGUAGE IMPAIRED	-43.5814	-0.896
LEARNING DISABLED	-43.2694	-0.882
NORMAL HEARING	-40.7301	-0.881
CLOSURE DURATION	-34.1147	-0.940
IMPAIRED CHILDREN	-31.0231	-0.856
FUNDAMENTAL FREQUENCY	-30.6432	-0.771
ACOUSTIC REFLEX	-30.0184	-0.797
DISABLED CHILDREN	-29.3471	-0.716

Table 4: Top 10 rising and falling terms in the CommDis corpus. Note that some of the falling terms are as much a product of “political correctness” as they are of the nature of inquiry in the field.