[French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society.* NJ:LEA.]

Four Problems with Extracting Human Semantics from Large Text Corpora

Robert M. French and Christophe Labiouse
Quantitative Psychology and Cognitive Science
Psychology Department, University of Liege, Belgium
{rfrench, clabiouse}@ulg.ac.be

Abstract

We present four problems that will have to be overcome by text co-occurrence programs in order for them to be able to capture human-like semantics. These problems are: the intrinsic deformability of semantic space, the inability to detect co-occurrences of (esp. distal) abstract structures, their lack of essential world knowledge, which humans acquire through learning or direct experience with the world and their assumption of the atomic nature of words. By looking at a number of very simple questions, based in part on how humans do analogy-making, we show just how far one of the best of these programs is from being able to capture real semantics.

Introduction

"You shall know a word," wrote J. R. Firth in 1957, "by the company it keeps." This idea, in one form or another, underlies the statistical study of the co-occurrence of lexical items in large text corpora. This burgeoning field of research has been made possible to a large extent by the ready availability of vast databases of text that can be automatically scanned by computer. While we certainly do not dispute the value of the statistical study of large text corpora, we take issue with the claim that lexical co-occurrence alone can capture real-world semantics. We focus on four main problems with co-occurrence analysis programs:

- they do not take into account the intrinsic deformability of semantic space due to context-dependence;
- the cannot detect co-occurrences of abstract structures, especially when they are highly distal;
- they lack of essential world knowledge, which humans acquire through learning or direct experience with the world;
- they assume that words are "atomic" entities.

These issues are ones that will have to be effectively dealt with by text analysis techniques in order for them to capture even elementary human semantics. These points will be examined in the context of human analogy-making.

The remainder of this paper is organized as follows. We begin by discussing the relationship between analogy-making and concept meaning and show why current co-occurrence programs would have so much difficulty with this broader view of concept meaning. We then consider one of the best recent co-occurrence programs, PMI-IR (Turney, 2001a) and show just how far this program is from being able to plausibly respond (i.e., in a human-like manner) to even the simplest possible analogies.

The intrinsic deformability of semantic space

We will take issue with one of the main principles underlying LSA (Landauer and Dumais, 1997), HAL (Lund & Burgess, 1996) and other programs based on lexical analysis of large corpora – namely, that "The meaning of a word can be thought of as a location in semantic space and the dimensionality of that space and the location of any word within it can be recovered from estimates of the distance between word pairs." (Fletcher & Linzie, 1998). The implication is that words have stable, fixed locations in semantic space. While this is obviously not entirely false, this principle overlooks the fact that these locations in semantic space are *highly context dependent*. They not only can, but *must* be able to move considerably in semantic space depending on the context in which they are to be used.

Consider a very simple example. A "claw hammer" would, under most circumstances, be close in semantic space to terms like "ball-peen hammer," "hit," "pound," "nail," "saw" and, even, "club." However, if, while nailing a floor, you suddenly have a back itch, the "claw" part of the hammer will likely become much more salient as a back-scratcher, rather than a nail-remover. Your realization that you can use the hammer as a back-scratcher temporarily moves the object in semantic space much closer to "back-scratcher," "itch," etc., than when it is perceived only as an object with which one can drive in nails. This "relocation" of the meaning of a word/concept in semantic space based on context is at the very heart of analogy-making, of perceiving one object as an instance of another class of objects (Chalmers, *et al*, 1992; Hofstadter, 1995). It is therefore essential to any algorithm that claims to be able to automatically extract word meaning from very large text corpora And it is precisely this ability to relocate in semantic space in a context-dependent manner that is currently beyond the reach of all co-occurrence techniques.

In short, while co-occurrence techniques may plausibly situate a word in semantic space with respect to its *average* usage, this is not sufficient to capture the context-dependent shifts in word meaning required to understand even the simplest analogies on which so much of our cognition is based. For this we need to somehow extract, at the very least, abstract relational information concerning the word.

[French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. NJ:LEA.]

Detection of (distal) abstract structures

Acquiring the semantics of a particular word allows us to rate the quality of associations between that word and other words. To plausibly claim that a program has acquired, or even partially acquired, the semantics of a word means that it should give word-association ratings that are at least approximately similar to those given by humans (French, 1990). We will use this rating technique to judge the performance of text co-occurrence programs.

There are (at least) two different bases for these associations, even if this distinction is not always easy to characterize (Chalmers, French, & Hofstadter, 1992). To say, "John is a real beanstalk," refers to largely "surface" attributes of John and beanstalks — namely, they are both tall and thin. On the other hand, when we say "John is a real wolf with the ladies," we don't mean John grows long gray hair around women and bites them, but rather that his *relation* with women is socially predatory, analogous to a wolf's relation of physical predation with its prey. The first analogy is largely *attributional*, based essentially on common surface features (in this case, the attributes "tall" and "thin") of John and beanstalks, whereas the latter analogy is primarily *relational*, based on a mapping between John's behavioral interactions with women and wolves behavioral interactions with prey. The first kind of association can be captured by co-occurrence techniques, whereas the latter — the basis of virtually all deep analogy-making (Gentner, 1983) — is still well beyond the reach of these techniques.

Incorporating semantic information

An equally important difficulty involves the unavailability to these programs' of crucial semantic information that cannot be acquired merely by examining word co-occurrences. In two of the examples below this lack of crucial contextual knowledge — that fathers are always male in one example, and the fact that there is an undeclared war going on between the Israelis and the Palestinians in the other — causes the particular text co-occurrence analysis program under consideration to fail completely in responding to the simplest questions involving word meaning. Humans acquire this information through direct experience with the world or through explicit learning, whereas these programs currently have no way of acquiring it. The point is that, when making judgments about word meaning, people — unlike co-occurrence programs — make use of a wealth of relational and semantic information that is unrelated to word co-occurrence.

Words are not atomic entities

Consider an example of a "subcognitive" question from French (1988, 1990) involving the rating of a neologism. "On a scale of 1 (awful) to 10 (excellent) please rate:

- *Flugly* as the name of glamorous Hollywood actress,
- *Flugly* as the name of an accountant in a W. C. Fields movie."

Humans, of course, can do these two ratings without difficulty: *Flugly* is a decidedly lousy name for a glamorous Hollywood actress and a fine name for an accountant in a W. C. Fields movie or a teddy bear. But how do we "know" this, since you have never seen the word *Flugly* before? You know, at least in part, that *Flugly* doesn't work for the name of a glamorous actress because of its *component parts* (French, 1990). In particular, it contains an unpleasing-to-the-ear guttural "g," to say nothing of the syllable "ug" or the entire word "ugly." Similarly, we rate it as a good name for an accountant in a W. C. Fields' movie because, in our mind's ear, we hear him pronouncing the name as "Flugleeee." This requires phonetic information acquired by having heard W. C. Fields' unique manner of speaking (or having heard others imitating this manner) and by the fact that various components of *Flugly,* namely, "ly," can be transformed into a drawling "leeee."

The point is that words contain parts that contain crucial information that contributes to the overall meaning of the word. Co-occurrence programs are currently insensitive to this information. And it is not clear that by extending their analyses to the letter or syllable level that i) there would not be a problem of combinatorial explosion and ii) that this would be an appropriate way to acquire this information.

PMI-IR

In the examples that follow, we will consider the performance of one recent program, PMI-IR (Turney, 2001a, b), that, according to its author, outperforms all other current programs on the most widely used benchmark for programs that attempt to extract word meaning from large text corpora. This benchmark is their performance on the standard synonym selection tasks that are part of the Test Of English as a Foreign Language (TOEFL) and the test of English as a Second Language (ESL).

The co-occurrence technique used by PMI-IR is one of a family of "Pointwise Mutual Information" (PMI) techniques developed by Church & Hanks (1989) and Church *et al.* (1991). In order to calculate the conditional probability scores on which it bases its choice of the correct synonym, the program queries 350 million pages of English text indexed by the AltaVista search engine. The most sophisticated version of PMI-IR is able to make use of local (proximal) context in order to correctly answer questions such as, "Every year in the early spring farmers [tap] maple syrup from their trees (drain; boil; knock; rap)." As Peter Turney, the author of PMI-IR, points out, "the problem word *tap*, out of context, might seem to best match the choice words *knock* or *rap*, but the context *maple syrup* makes *drain* a better match for *tap*" (Turney, 2001b). The program factors in the context provided by "maple syrup" to correctly answer this question.

The program does, indeed, perform impressively on the synonym recognition task. According to Turney, the

[French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. NJ:LEA.]

program produced the following results on the standard TOEFL and ESL synonym recognition task:

> "The task of synonym recognition is, given a problem word and a set of alternative words, choose the member from the set of alternative words that is most similar in meaning to the problem word. PMI-IR has been evaluated using 80 synonym recognition questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym recognition questions from a collection of tests for students of English as a Second Language (ESL). On both tests, PMI-IR scores 74% . . . For comparison, the average score on the 80 TOEFL questions, for a large sample of applicants to US colleges from non-English speaking countries, was 64.5% (Landauer and Dumais, 1997). . . . Latent Semantic Analysis (LSA), another statistical technique, scores 64.4% on the 80 TOEFL questions (Landauer and Dumais, 1997)."

Three examples

In what follows we use a word-rating technique from French (1988, 1990) and similar to standard similarity judgment techniques used to study how word meanings are represented (see, for example, Rips, Shoben, & Smith, 1973). The key idea is that these simple questions require non-local context for their answers (French & Labiouse, 2001).

Rating lawyers

Our first example involves the rating of *lawyers* as various other entities.

> "Rate on a scale of 1 (terrible) to 10 (excellent) rate *lawyers* as: horses, fish, telephones, stones, sharks, cats, flies, birds, slimeballs, kangaroos, robins, dogs, and bastards."

We applied the PMI-IR search technique described in Turney (2001b) using the Alta-Vista search engine and found that it gave the *lowest* (i.e., poorest) ratings to "Lawyers as slimeballs" (1.06) and "Lawyers as bastards" (1.15), the latter being roughly equivalent PMI-IR's rating of "Lawyers as kangaroos" (1.17)! We then asked a group of 26 undergraduates at Willamette University (Oregon) to also do these ratings. These results (Figure 1) are much more in line with one might expect for humans with a clear understanding of the semantics of the word "lawyer" — namely, lawyers are judged (fairly or unfairly) to be most like slimeballs, bastards, dogs and sharks, and least like telephones, kangaroos, and birds. PMI-IR, on the other hand, judges lawyers to be most like computers, cats, and telephones and least like slimeballs, bastards, kangaroos and robins. Lawyers as sharks or fish are judged to be equally bad. A comparison of human vs. PMI-IR results can be seen in Figure 1. In short, it is amply clear that even for this straightforward question about lawyers, the human semantics of "lawyer" does not even vaguely resemble the semantics extracted by PMI-IR.

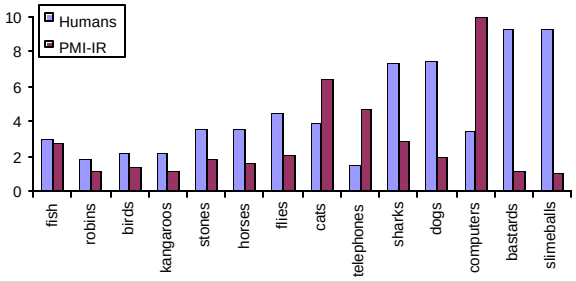

Figure 1. A comparison of PMI-IR and human data. The two profiles are clearly very different.

We also found that PMI-IR gave an extremely high rating to "Lawyers as children," higher, in fact, than any of the choices tested in Figure 1. Clearly, something is wrong here: first, lawyers cannot even *be* children (something which PMI-IR has no way of knowing) and, even metaphorically, it just doesn't seem right to us.

Rating the plausibility of Jewish/Palestinian ministers' names

Next we used PMI-IR to judge how good various first names would be for an Israeli or a Palestinian minister. We chose ten traditional Jewish names (Uri, Ariel, Moshe, Yitzhak, Yehudi, David, Samuel, Benjamin, Shimon, and Zeev) and nine traditional Arab names (Saddam, Usama, Ahmed, Mohammed, Salah, Amin, Khalil, Ashrawi, and Yasser). We asked two separate questions, each processed independently by the program. The first was "How good is X [one of the names, e.g., *Ahmed*) as the name of an Israeli minister?" All nineteen names were rated for this question. Then a second question was asked: "How good is X [again, one of the 19 names] as the name of a Palestinian minister?" All 19 names were rated for this second question. We then compared the ratings for each name for the two questions to determine their degree of correlation.

Once again, PMI-IR fails rather spectacularly: for example, it considers *Yasser* to be almost as good a first name for an Israeli minister as for a Palestinian minister! Similarly, *Ariel* is judged to be the best name, out of all ten Jewish names and all nine Arab names, for either an Israeli minister or a Palestinian minister . The results for the other names are shown in Figure 2.

Why does the program rate *Yasser* as a highly probable name for an Israeli minister and *Ariel* as highly probable for a Palestinian minister? The reason is simple: Because the program is concerned *only* with the co-occurrence of words, in this case the words *Yasser*, *Ariel*, *Israeli*, *Palestinian* and *minister*. The fact that Israel and Palestine are currently waging an undeclared war is known to PMI-IR only through higher than normal co-occurrences of war-related words and words like *Israel, Palestine, intifada*, etc. It knows nothing *about* wars, about their causes and effects, about their effects on societies and individuals in those societies, about hatred, about destruction, about refugees, about Israel, about Palestine, etc. *ad infinitum*. It knows only that sometimes these words co-occur with higher

frequency than others. The complete absence in PMI-IR of this deep relational structure in which the words that it encounters (and concepts these words represent) are embedded is precisely why PMI-IR fails to convincingly answer even the simplest questions that require deeper relational structure and knowledge to be answered plausibly.

So, to return to our example, in the context of the current crisis in the Middle East and of cultural specificities of first-names, good names for Palestinian ministers should be perceived as bad names for Israeli ministers and vice-versa. PMI-IR is, as we have said, unaware of the cultural context surrounding these questions. Specifically, PMI-IR is ignorant of the obvious (to us) cultural fact that some first names are typically Jewish while others are typically Arab and the relation of that cultural fact to the currently perceived inappropriateness of Palestinian ministers with Jewish names and vice-versa. So, according to PMI-IR, the appropriateness of a name for a Palestinian minister correlates almost perfectly (+0.98) with the appropriateness of the same name for an Israeli minister (See Figure 2).

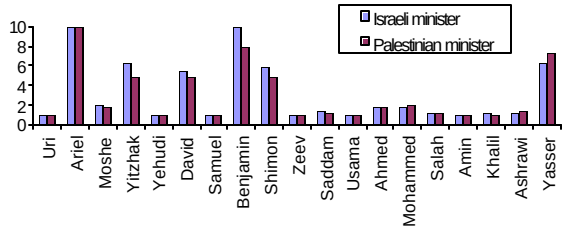

Figure 2 For two separate questions: "How good is X as the name of an Israeli minister?" and "How good is X as the name of a Palestinian minister?" PMI-IR produces an almost a perfect correlation between the appropriateness of a given name as either that of an Israeli or a Palestinian minister.

Rating names of mothers and fathers

Finally, we decided to pick an example, simple in the extreme and far removed from politics and current events, that relied on a very specific piece of contextual information that would be available to all humans but not to a word co-occurrence analysis. We compared PMI-IR's answers to the following two questions: How good is X [a first name] as the name of a father?" and "How good is X [the same first name as in the first question] as the name of a mother?" For each question we asked PMI-IR to rate ten very common men's names (John, William, Stuart, Peter, Robert, Jack, Gary, Steve, Albert, and Michael) and ten very common women's names (Barbara, Mary, Patricia, Linda, Susan, Jennifer, Karen, Nancy, Elizabeth, and Dorothy).

When judging the appropriateness of a particular name as the name of a father (or mother), humans partly rely on a simple fact that the program does not have — namely, that fathers are invariably men, while mothers are invariably women. Consequently, humans will *necessarily* rate women's names lower than men's names for the question: "How good is X as the name of a father?" Not so PMI-IR. The program concludes that "John" is the best name out of all twenty names for a father *and for a mother*. It rates "Mary" as being a very good name for a father or for a mother. Ditto for the name "William." As in the above example, the appropriateness of a particular name for a father correlates essentially perfectly (+0.99) with the appropriateness of that same name for a mother! (Figure 3)

Once again, the program fails because extracting co-occurrences of words in a large corpus of text is simply not good enough to answer questions that require abstract contextual knowledge or experience. Again, the problem is that PMI-IR has neither abstract rules nor world experience that it can rely on. And since, in any text where the word "father" occurs, the word "mother" will generally not be far away, PMI-IR fails completely on this simple rating task.

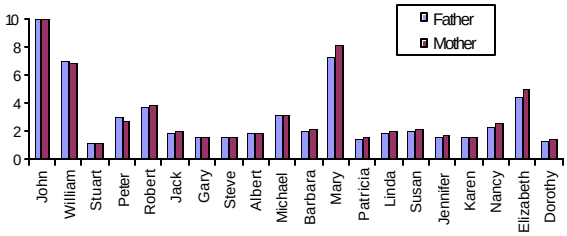

Figure 3 Two questions were asked: "How good is X as the name of a father?" and "How good is X as the name of a mother?" Lacking all context about what "fathers" and "mothers" actually are, PMI-IR produces an almost a perfect correlation between the appropriateness of the names, male or female, for a father or for a mother!

Why PMI-IR works so well on the synonym selection task

Given PMI-IR's poor performance on the simple examples above, how could it be so good in the synonym selection task, a task that would seem to require a relatively sophisticated semantic understanding of words in order to be done successfully? In what follows we will briefly examine why this program, and other similar programs, most notably LSA, are able to perform so well on this task, in spite of their inability to do the examples above.

The author of PMI-IR claims that his program can do better on the TOEFL and ESL synonym tests than any other current computer program (Turney, 2001a,b). This is believable and reasonable. Turney illustrates PMI-IR's performance on the synonym-finding task with the word *levy* (as in "to levy taxes"). Four choices are proposed — *imposed, believed, requested, correlated* — and the program chooses one of them as the best synonym based on how often that word is close to "levy" in many Web pages. The reason for PMI-IR's success does, indeed, reflect the semantics of the word under consideration, but is tied most directly to the stylistic reasons for which we use synonyms — viz., so

[French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. NJ:LEA.]

as not to repeat the same word too often in a given text or, especially, in the same paragraph. This purely *stylistic* constraint imposes the proximity of synonyms, which is detected by PMI-IR.

Assume you are writing an article to be put on a Web page about some *blunder* that occurred. In describing this blunder, you are aware that it is bad style to repeat the word *blunder* over and over again in your text, so you resort to synonyms, such as *failure*, *mishap*, *mistake*, *slip*, *bungle*, *mess*, and so on. This obviously produces co-occurrences of *blunder* and *mistake*, of *blunder* and *slip*, etc., and this is precisely what PMI-IR detects. A *blunder* IS (to a first approximation) a *mistake*, which IS a *slip*, etc. These words all have approximately the same dictionary definitions. In other words, the features that describe them are largely identical. This is what we called above *attributional similarity*. The point is that we can expect attributionally similar words, if only for stylistic reasons, to occur close to one another in a text. Hence, PMI-IR's excellent performance on this task.

This technique can, indeed, incorporate proximal context, as in the example of the word *tap* in the context of "maple syrup." But most analogical association involves abstract context derived from examples that, if they exist at all in the text corpus, may well be separated by millions of pages from the word under consideration. It is an open question in the field of computational analogy-making as to how this abstract relational structure might be stored and indexed fluidly enough to be accessible for later retrieval in a wide variety of contexts (see Chalmers, French, & Hofstadter, 1992, for a detailed discussion of this issue), but one thing is clear: it is *not* accessible to programs that rely only on local word co-occurrence to produce their semantics.

And, to be fair, this is one way in which humans learn attributionally similar words/concepts. But there is much more to "semantic similarity" than surface similarity.

To reiterate, *relationally (or metaphorically) similar* words require a great deal more than the detection of attributional similarity and physically proximal context. Consider rating a *banana split* as *medicine* (French, 1988, 1990). The number of times that these two items will occur together in any text anywhere is now, and will forever be, infinitesimally small compared to the other associations involving banana splits or medicine. For programs that extract semantics only from text corpora this poses a serious problem, referred to as the problem of data sparseness (Dagan et al., 1994). But the problem is unavoidable. *Of course* the number of Web pages containing the terms "banana split" and "medicine" will be vanishing small because is it not a common association at all, but it remains a perfectly valid, readily understandable one that we can judge without difficulty because we understand it *in relation to our experience with the world*, i.e., to facts like the doctor bringing us a bowl of ice-cream after we have had our tonsils out, with our mother taking us for a sundae to pick up our spirits when our junior high school safety poster was eliminated from the city competition, etc.

In other words, describing one word in terms of another usually involves much more than the above kind of "blunder-mistake-mishap-slip" synonym searching. It involves mentally placing the both words in a variety of *relational* as well as attributional contexts (that can shift fluidly) and converging on a context that fits both words (for detailed discussions of this see: Chalmers, French, & Hofstadter, 1992; Hofstadter, 1995; etc.) If both words fit that context very well, then we give the association a high rating. The more difficult it is to converge on an appropriate context for both words, the lower the rating.

PMI-IR, however, is incapable of extracting these all-important relational and contextual characteristics of situations. Specifically, for questions of the form, "Rate X as a Y," the program is incapable of grasping the relational structure in which each of the words is embedded and then of mapping those two structures onto one another in order to determine the relational similarity of the words.

Conclusions

While we acknowledge the impressive performance on certain lexical tasks of programs that employ co-occurrence analyses on large text corpora, our contention is that these programs lack the capabilities necessary to acquire real (i.e., human) semantics. This paper must not be read as a criticism of these methods per se, but rather as an incentive for researchers to develop new techniques that can incorporate more of the mechanisms by which we humans acquire semantics. These requirements go well beyond the often-cited problems of the lack of syntactic knowledge (Perfetti, 1998) and conceptual disambiguation (Landauer & Dumas, 1997). We have pointed to four problem areas for these programs, areas in which we believe future research should be focused. These areas are i) the ability to cope with the context-dependent deformability of semantic space, ii) the detection of co-occurrences of abstract structures, especially similar, but distal, abstract structures, iii) the means of providing the programs with essential world knowledge, and iv) the elimination of the assumption of words as "atomic" entities. In other words, we maintain that to know a word in a manner even approximately equivalent to how we humans know it, requires far more than merely knowing the "company it keeps."

In short, while the area of text analysis of large corpora is a fascinating and promising one, we believe that in order for real, human semantics to emerge from these techniques, the problems raised in this paper will have to be squarely confronted and overcome.

Acknowledgments

This work was supported in part by grant HPRN-CT-1999-00065 from the European Commission. Christophe Labiouse is supported by a Belgian NFSR Research Fellowship. The authors would also like to thank Jim Friedrich at Willamette University, Oregon,

[French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. NJ:LEA.]

for his help in conducting the informal survey cited in the text.

References

Chalmers, D. J., French, R. M. and Hofstadter, D. R. (1992). High-level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *J. of Experimental and Theoretical and Artificial Intelligence, 4*(3), 185-211.

Church, K.W., and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, pp. 76-83.

Church, K.W., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Uri Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. New Jersey: Lawrence Erlbaum, pp. 115-164.

Dagan, I., Pereira, F. & Lee, L. (1994). Similarity-based estimation of word co-occurrence probabilities. *Proceedings of the 32nd Annual Meeting of the Assoc. for Computation Linguistics,* 272-278.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman (1968).

Fletcher, C. & Linzie, B. (1998). Motive and Opportunity: Some Comments on LSA, HAL, KDC, and Principal Components. *Discourse Processes, 25*(2&3), 355-361.

French, R.M. (1988). Subcognitive Probing: Hard Questions for the Turing Test. *Proceedings of the Tenth Annual Cognitive Science Society Conference*, Hillsdale, NJ: LEA. 361-367.

French, R.M. (1990). Subcognition and the Limits of the Turing Test. *Mind, 99*(393), 53-65.

French, R. M. and Labiouse, C. (2001). Why co-occurrence information alone is not sufficient to answer subcognitive questions. *J. of Experimental and Theoretical Artificial Intelligence.*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170.

Hofstadter, D. R. and the Fluid Analogies Research Group (1995). *Fluid Concepts and Creative Analogies* , New York, NY: Basic Books.

Landauer, T. & Dumais, S. (1997). A solution to Plato's Problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers, 2,* 203-208.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. Discourse Processes, 25, 363-377.

Rips, L. J., Shoben, E. J. & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1-20.

Turney, P.D. (2001a). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001*), in press.

Turney, P. D. (2001b). Answering subcognitive Turing Test questions: A reply to French. *J. of Experimental and Theoretical Artificial Intelligence.*

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 454-46.