Evidence Based Library and Information Practice

Article

Measuring the Extent of the Synonym Problem in Full-Text Searching

Jeffrey Beall
Metadata Librarian
University of Colorado Denver
Denver, Colorado, United States of America
E-mail: jeffrey.beall@ucdenver.edu

Karen Kafadar
Rudy Professor of Statistics
College of Arts and Sciences, Indiana University
Bloomington, Indiana, United States of America
E-mail: kkafadar@indiana.edu

Received: 03 September 2008 **Accepted:** 23 October 2008

Abstract

Objective – This article measures the extent of the synonym problem in full-text searching. The synonym problem occurs when a search misses documents because the search was based on a synonym and not on a more familiar term.

Methods – We considered a sample of 90 single word synonym pairs and searched for each word in the pair, both singly and jointly, in the Yahoo! database. We determined the number of web sites that were missed when only one but not the other term was included in the search field.

Results – Depending upon how common the usage is of the synonym, the percentage of missed web sites can vary from almost 0% to almost 100%. When the search uses a very uncommon synonym ("diaconate"), a very high percentage of web pages can be missed (95%), versus the search using the more common term (only 9% are missed when searching web pages for the term "deacons"). If both terms in a word pair were nearly equal in usage ("cooks" and "chefs"), then a search on one term but not the other missed almost half the relevant web pages.

Conclusion – Our results indicate great value for search engines to incorporate automatic synonym searching not only for user-specified terms but also for high usage synonyms. Moreover, the results demonstrate the value of information retrieval systems that use controlled vocabularies and cross references to generate search results.

Introduction and Context of the Study

Full-text searching generates results by matching a word or words in a search query with words in a database. The synonym problem in full-text searching occurs when a searcher looks for information on a topic and enters a search using a single term to represent the topic but does not also enter any synonyms for that topic. For example, a search for information on dentures with only the word "dentures" as a search term could miss documents that refer to this concept by its synonym "false teeth", because the two terms have no words in common. For most full-text searching, “value-added” features such as controlled vocabularies and cross references are not present. These features serve to retrieve and co-locate documents on a given topic in search results regardless of the terms used in the full text of searched documents.

This article seeks to measure the extent of the synonym problem in full-text searching. More precisely, this study looks at single word pairs of synonyms, and for each term measures the proportion of documents that are missed when one term is searched, and the proportion of documents that contain only the synonym. This study is limited to traditional full-text search engines, that is, search engines that match words in a search query with words in full-text documents and return results.

Full-text Search Engines and Synonyms

With the advent of the Internet, full-text searching has proliferated, and with it the desire to retrieve as much information about a topic as possible. A problem that arises with such searches is the potential for the search to return only a subset of the web sites with relevant information because the search concept can be referenced by more than one term. The concept can be described by simple nouns ("false teeth" and "dentures"), or by broader terms, such as "botany" and "plant science", or "aurora borealis" and "northern lights". A search in most search engines on the term "botany" (or "aurora borealis") may well miss web pages that refer to the discipline only as "plant science" (or "northern lights").

A few authors have commented on this effect. For example, in *The Oxford Guide to Library Research* Mann writes:

> When all is said and done, keyword searching necessarily entails the problem of the unpredictability of the many variant ways the same subject can be expressed, within a single language (“capital punishment,” “death penalty”) and across multiple languages (“peine de mort,” “pena capitale”). And no software algorithm will solve this problem when it is confined to dealing with only the actual words that it can retrieve from within the

> given documents (or citations or abstracts) themselves. (102)

Beall refers to this problem as the “synonym problem” and states, “In full-text searching, synonyms hinder effective information retrieval when a searcher enters a term in the search box and the system only returns results that match the term and does not return results that refer to the concept only by one of its synonyms” (“Weaknesses” 439). For example, some use the term “botany” and others use “plant science” to describe the same concept. A search in most search engines on the term “botany” would probably miss web pages that refer only to the discipline as plant science (Beall “Death” 6).

Fugmann uses the term “paraphrase lexicalization” to describe the disconnect between a user's search terms and the terms used in relevant documents. He exemplifies the synonym problem by giving an example of a searcher looking for information on insecticides and missing documents that refer to them as pesticides. He states, “…an inquirer expects all documents to be retrieved in which the concept of the search request is dealt with and in fact independent of how it happens to have been expressed by an author” (223).

Dagan et al. describe the synonym problem from an information science perspective. Their study “investigates conceptually and empirically the novel sense matching task, which requires [one] to recognize whether the senses of two synonymous words match in context” (449). They describe this phenomenon as “lexical substitution”. Their study does not measure the synonym problem but attempts to lay the groundwork for an algorithmic solution to it.

While only a few authors have noted the synonym problem, even fewer have attempted to measure it. The challenges to measuring the extent of the synonym problem include defining an appropriate measure, and designing a study to quantify it. To our knowledge, no previous study has been conducted to measure the synonym problem. This article attempts to fill that void.

On its web page, Google describes a “synonym search”, but it provides very little information about this type of search. On one of its help pages Google states, “If you want to search not only for your search term but also for its synonyms, place the tilde sign ("~") immediately in front of your search term” (“Web Search Help Center”). We suspect that rather few Google users are aware of this feature, and even fewer take advantage of it. Google offers no further explanation of this feature. Slightly more information is provided in the patent application granted to Google in 2002 and issued in 2005 for a process that essentially functions as an algorithmic synonym search, rather than a deterministic synonym search (by matching synonyms from a pre-constructed list). According to the patent's abstract:

> Methods and apparatus determine equivalent descriptions for an information need. In one implementation, if adjacent entries in a query log contain common terms, the uncommon terms are identified as a candidate pair. The candidate pairs are assigned a score based on their frequency of occurrence, and pairs having a score exceeding a defined threshold are determined to be synonyms. (Dean et al. 2005)

We assume that the phrase “equivalent descriptions” here means "synonyms", but it is unclear whether Google has implemented the process described in this patent into its current search algorithms. For proprietary reasons, search engine companies release very little information about the algorithms they employ to generate results. Bade says "… the exact nature of the formulae used remains largely unknown to the public since these are valuable intellectual property for their owners" (831).

At least one library online catalog product offers a synonym search feature. The Innovative Interfaces, Inc. online catalog allows libraries to program in synonyms. Once a synonym pair has been programmed into the system, a keyword search on either of the two words in the pair returns results as if both search terms had been entered. This feature is not used so much for synonyms as it is for variant spellings, such as British and American variants like "labor" and "labour".

Methods

Our original plan was to generate a random sample of synonym groups and then to search them in both Google and Google Book Search. As our source for synonyms, we planned to use printed thesauri from the reference section in the Auraria Library on the campus of the University of Colorado Denver. After collecting the data, we planned to do a statistical analysis to answer our research question.

Difficulties with Synonyms

We soon realized that exact synonyms are rare, and words listed as synonyms in thesauri are close in meaning but frequently are not true synonyms. One example of a false pair of synonyms is the pair "waterfall" and "cascade". While close in meaning, there is a significant semantic difference between these two terms. We sought to study synonym groups that were as semantically identical as possible. We suspected that the use of "non-exact synonyms" such as "waterfall" and "cascade" would result in even more missed web pages, and hence an even more severe problem than what we ultimately observed.

Difficulties with Google

Before we began to collect data we performed numerous test searches, which immediately revealed two significant problems for conducting this research with Google. The first problem was that the Google search software does not allow nested Boolean searching. That is, if a term contains more than one word, Google will not allow a searcher to apply the Boolean operator "not" to the phrase. This was a significant problem for us, because our study objective required us to search for one term but *not* the other. As an example, for the synonym pair "leprosy" and "Hansen's disease", ideally we would perform the following search in Google:

Leprosy -"Hansen's disease"

The minus sign within Google activates the Boolean operator "not" in the search, and the quotation marks indicate a term to be searched as a phrase. Unfortunately, the Google search engine lacks the functionality to correctly perform this type of search. Our test searches showed that when we tried to use nested Boolean terms, the phrases we attempted to exclude often appeared in the pages retrieved by the search. This would prevent us from accurately measuring the number of resources missed due to the synonym problem.

To address the difficulties with synonyms from thesauri, we abandoned printed thesauri as a source for a random selection of synonyms and turned instead to controlled vocabularies. Controlled vocabularies also are frequently called thesauri; they list the preferred term for a concept followed by a list of the variant terms or "cross references". The Library of Congress Subject Headings is an example of a controlled vocabulary, and as one of the most comprehensive we selected this controlled vocabulary as the source for our random selection of synonyms.

At this point we encountered our second major problem with Google: an apparent inconsistency in the search results in the Google database. One of the valuable features of the Google database that benefits information retrieval research is that results of each search include the total number of web sites retrieved. However, as we were conducting our test searches, we found this count to be highly variable. In some cases, for example, the same search performed at two different times retrieved significantly different numbers of "web pages found”. We illustrate this problem with a more detailed description of our study design.

For our study we required data on the number of web pages found from the following searches for each word pair, expressed in Figure 1 as a Venn diagram.

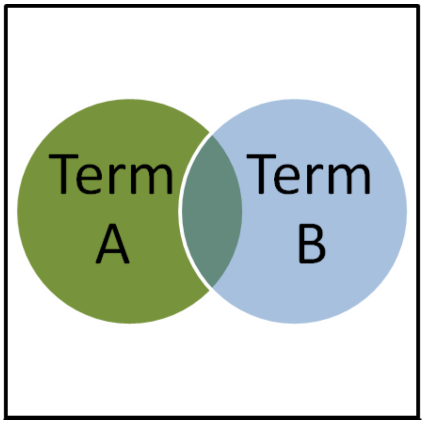

Fig. 1. A Venn diagram that illustrates the data gathered for each word pair.

The Venn diagram in Figure 1 expresses the following Boolean logic:

A not B (Represented by only the area in the left circle that is shaded green)
B not A (Represented by only the area in the right circle that is shaded blue)
A and B (Represented by the blue-green shaded area in the center)
A or B (Represented by the entire shaded area of the diagram)

Our study depends critically on the numbers of web pages found for each of these four searches. One would expect that the sum of the numbers of web pages found from the first three searches should equal the number of web pages found by the fourth search. However, in our test searches we observed discrepancies between these two results, sometimes as large as several million. Indeed, the four individual numbers from the four searches often varied substantially.

We postulated several explanations for the wide discrepancy. First, it could arise if Google actually applies its patented "synonym search" feature described in its help pages. Second, if Google's cited "number of web pages found" is not a deterministic count, but rather is a statistical estimate based on the current version of the search algorithm being used, then one would expect variability in the estimate at different times. A third possible explanation arises from the fact that "every search in Google is part of an experiment" (Pregibon and Lambert), so searches of the same query at different times may result in different algorithms being applied.

Finally, discrepancies could reflect actual changes in the number of web pages available due to new and deleted web pages over time. However, we suspect the numbers of new and deleted web pages on widely diverse topics would not vary much, casting doubt on this fourth possibility as a plausible explanation for the extreme variability we observed in our searches, some of which involved rather obscure terms. Slight discrepancies would have been tolerable, but for our study such huge discrepancies rendered Google searches too variable for our purposes. For this reason, we turned to alternative search engines.

Revised Methods

Because of the inability of Yahoo! (as well as Google) to perform nested Boolean searches, we decided to limit this study to only single word synonym pairs. (We would like to repeat this study on more complicated synonym-phrase pairs when a nested Boolean search feature is implemented in one of the search engines.) We generated a random list of synonym word pairs from the library catalog at the Auraria Library (University of Colorado Denver). Using the search functionality in the "staff" mode of the library's catalog, we created a list of all topical subject authority records that contained at least one cross reference. Because the Auraria Library serves three institutions of higher education, including a comprehensive university, the scope of the headings in the library is unusually broad. Our generated list contained 39,511 records. We then used a program available through the R Project for Statistical Computing <http://www.r-project.org> to generate 100 random numbers distributed uniformly across the range [1, 39511], which identified the indices of the 39,511 records selected for this study.

As indicated above, we limited our study to single word pairs of synonyms, meaning that both the heading and the cross reference had to be single words. We imposed two further conditions on the word pairs for this study to avoid the potential for the 100 pairs to include geographic- or location-specific terms. The two conditions thus relate to the structure and composition of the Library of Congress Subject Headings (LCSH) thesaurus. First, we skipped records whose cross references were also cross references from another record. Second, we insisted on semantically exact word pairs. The LCSH does group semantically related concepts on a single record. For example, the LCSH heading for "mountains" has a "see reference" for the word "hills". While similar, these two concepts are semantically different, even though LCSH groups them together on a single subject authority record for convenience.

To impose the two further conditions, we eliminated a pair and went to the next record in the list of 39,511 if the main heading / cross reference pair (a) involved more than one word for either the main heading or the cross reference; (b) contained a cross reference that was itself a cross reference; or (c) contained terms that were not semantically exact. In ten instances, the random numbers were so close together that no valid single word synonym pair appeared between the previous and the next randomly-selected record. Thus our final sample consisted of ninety pairs of single word synonyms. All searches were conducted by the first author (Beall).

Results

We applied our study plan to search 100 (later revised to 90) synonym, single word pairs in the Yahoo! database. The word pairs and the data are presented in the Appendix. The searches were conducted in March and April, 2007. When we gathered the data, we realized that the searches, like full-text searching, would not be perfect. For example, one of our synonym pairs was *biologicals* / *biologics*. It is likely that one of the terms is the name of a company, or is a word in a foreign language and in many contexts is not a synonym of the other, a situation that would affect our data. But there were far too many search results to examine to determine their context, and the type of searching we are studying, full-text searching, is also burdened by the same problem. We acknowledge this potential contamination by proper company names, but believe it to be quite small.

The pertinent data from this study are the percentages of total references ("A and B") found by searching for "A only" (i.e., number of pages found in search for word A only, divided by the total number of pages found in a search for either ("A or B")) and likewise searching for “B only”. Usually, one of A or B is the more common word, so the percentage for one will be higher, often substantially higher, than the percentage for the other word. Figure 2 displays via boxplots the data for the more common of the words in the pair (Max(%A,%B)), the data for the less common of the words in the pair (Min(%A,%B)), and the difference in the two percentages (Diff(Max–Min)). For convenience in this article, we will designate “A” as the more common word and “B” as the less common word (i.e., a search on “A” returned more web pages than a search on “B”).

Figure 3 displays information similar to the third box in Figure 2, but on an item-by-item basis. For example, the highest percentage among these word pairs occurred for word pair #53: “Mitochondria” but not “Chondriosomes” found 99.992% of the 2,000,189 web pages, while “Chondriosomes” but not “Mitochondria” found only 0.006% of the web pages. The designated line in Figure 3 connects these two proportions: 0.99992 (left side) and 0.00006 (right side). From this display it is clear that if one succeeds in identifying the more common word the search will yield most of the references, but if one asks for the less common word the search will miss almost all the web pages. For about 10-20 of the word pairs the words will each find about half of the available pages. For example, in word pair #72 each of the two searches, on "Preparedness" only and on "Readiness" only, returns about half of the total number of web pages, but also will completely miss the other half.

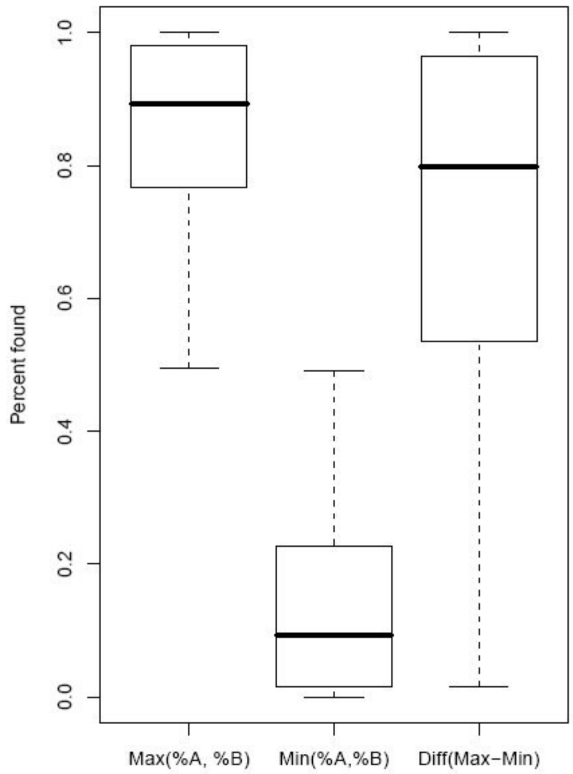

Fig. 2. Boxplot showing results missed from the perspective of the more- and less common word.

If one happens to select the more common of the words in the pair, one is often likely to capture most of the references (on average, about 88% of the references), but in 10 of the 90 pairs a search for even the more common of the words in the pair returned less than 55% of the available web pages. See Table 1 for the list of these 10 word pairs.

In a search, the proportion of missed web pages depends on whether one searched the more common or the less common word in the synonym pair. In these 10 word pairs

Table 1

	A	B	A (not B)	B (not A)	A and B	A or B	Prop. max	Prop min
2	Afrocentrism	Afrocentricity	57900	63800	1040	128000	49.8	45.2
20	Cooks	Chefs	18500000	23600000	2200000	44200000	53.4	41.9
24	Discrimination	Bias	44400000	35900000	2910000	83200000	53.4	43.1
26	Egoism	Egocentricity	1150000	10600	980	3000000	38.3	3.5
27	Electromagnetism	Electromagnetics	953000	750000	46500	1760000	54.1	42.6
50	Marmots	Marmota	325000	299000	8900	645000	50.4	46.4
69	Picornaviruses	Picornaviridae	35400	39200	2360	84200	46.6	42.0
72	Preparedness	Readiness	17200000	16500000	986000	39000000	44.1	42.3
81	Salafiyah	Salafiyya	17500	12100	66	38700	45.2	31.3
93	Tinsmithing	Tinwork	16200	13900	79	41700	38.8	33.3
99	Waka	Tanka	1530000	1580000	10700	3110000	50.8	49.2

even a search on the more common term returned less than 55% of the web pages found if both words were used in the search (i.e., 45% or more web pages were missed when using only one term in the word pair).

How costly can the search be, in terms of missed web pages, if one were to search on the less common of the two words in the pair? Figure 2 shows that, when the more common ("A") of the words is used in the search, often one captures 78% or more of the total web pages (the lower quartile of the percentages of web pages found using "A only" is 78%, as demonstrated by the lower edge on the left-most box in the boxplot). Conversely, if one were so unlucky as to have selected the less common word, one is likely to capture no more than 20% of the web pages (the upper quartile of the percentages of web pages found using "B only" is 20%, as shown by the upper edge of the middle box in the boxplot).

Even when the more common word is searched, the left box in Figure 2 shows that 25% of the searches returned only 50-78% of the available web pages. These results suggest that the "cost" of web searches for information about a topic can be rather high if one unfortunately enters the less common term, which may be frequent depending upon one's native language. (For example, Australians often use the term "jumper" for the American term "sweater", and the British use the term "biscuits" for the American term "cookies".) The third box in Figure 2 shows that the difference in "percentage of web pages found" can be very large -- often as high as 50-95% (lower quartile and upper quartile) -- depending on which word was selected for the search. Figure 3 shows both percentages for each word pair (A, the more common, on the left; B, the less common, on the right), connected by a dashed line. Often one of the words in the word pair is much more common than the other word. But for about one-fourth of the words in our study, both percentages are near 50%: a search for one term or the other fails to capture half of the web pages, regardless of whether one selected the "more" or "less" common word.

Even when the more common word is searched, the left box in Figure 2 shows that 25% of the searches returned only 50-78% of the available web pages. These results suggest that the "cost" of web searches for information about a topic can be rather high if one unfortunately enters the less common term, which may be frequent depending upon one's native language. (For example,

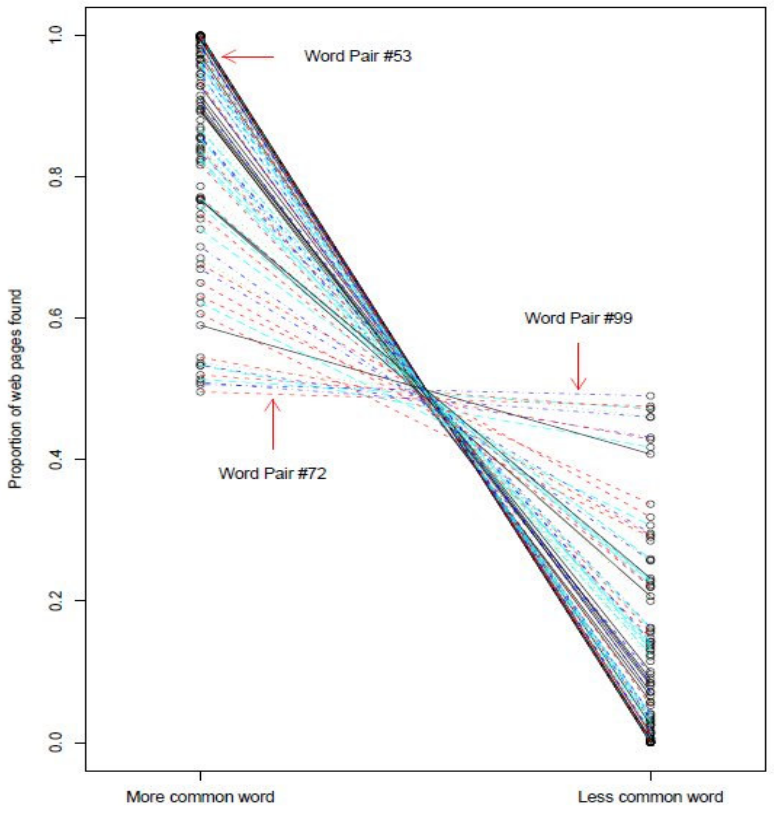

Fig. 3. Display of proportions of web pages found when searching on "More common word" (left side) versus "Less common word" (right side). Segments connect proportions. The greatest discrepancy occurs with word pair #53, "Mitochondria" (0.99992) versus "Chondriosomes" (0.00006). The least discrepancy occurs with word pairs #72 ("Preparedness", 0.496; "Readiness", 0.476) and #99 ("Tanka", 0.507; "Waka", 0.490).

Australians often use the term "jumper" for the American term "sweater", and the British use the term "biscuits" for the American term "cookies".) The third box in Figure 2 shows that the difference in "percentage of web pages found" can be very large -- often as high as 50-95% (lower quartile and upper quartile) -- depending on which word was selected for the search. Figure 3 shows both percentages for each word pair (A, the more common, on the left; B, the less common, on the right), connected by a dashed line. Often one of the words in the word pair is much more common than the other word. But for about one-fourth of the words in our study, both percentages are near 50%: a search for one term or the other fails to capture half of the web pages, regardless of whether one selected the "more" or "less" common word.

Discussion

While small in scope, this study demonstrates the severity of the synonym problem in web searching. Because of cultural or sociological differences in terms, the use of one term instead of its more common counterpart could result in highly incomplete web searches, raising only a fraction of the available web pages on this topic. For example, our study included the word pair “appraisers/assessors”; the former term is more common in some societies (73%), while the latter is more familiar in other contexts (but which captures only 26% of the web pages found by using both terms). For other word pairs, both words are used roughly equally often, but not in the same document, and hence a search on either word, but not the other, misses almost half the web pages found by searching on both (e.g., preparedness 49.6%: readiness 47.6%). Some search engines (such as Google Inc.) appear to offer synonym-searching capability, and based on our study, such a feature would result in more complete searches.

This study involves some important limitations which we need to acknowledge. First, the study is limited in its sample size. We selected only 100 pairs and our study design yielded data on only 90 word pairs, which is not a huge study but is definitely large enough to demonstrate the variability that can arise with the synonym problem. In addition, the uncertainties on the estimates of the reported percentages of missed web pages with each word in the search pair includes the uncertainties in the algorithms used by the search engine. One must keep in mind the extent that the search engine algorithms themselves are based on some sort of sampling strategy that returns estimates on "approximate number of web

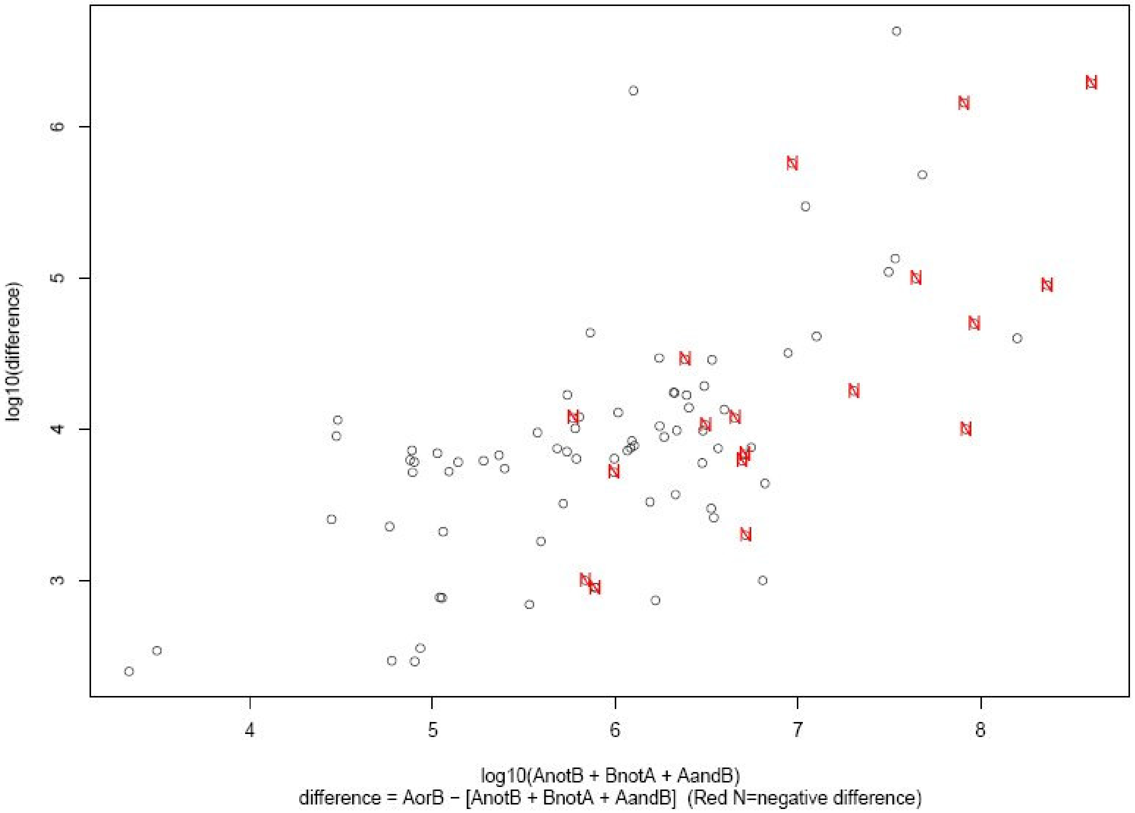

Fig. 4. The fewer the hits in a search, the more precise the estimate of number of web pages found.

pages found", which then affect our reported percentages. Clearly a larger study that involves replication is warranted to yield better estimates of the variability in the percentages reported here.

In a few instances, the data returned were illogical, in that the number of web pages found from a search of (A or B) exceeded the sum of the numbers of web pages found from the three searches combined (A not B) + (B not A) + (A and B). Two factors probably contributed to such events. First, the distributed system architecture can change quickly and repeatedly, resulting in different values at different times. Second, the search engine reports only an estimate, not a precise value, of the number of web pages found. These estimates are less precise for larger numbers of web pages found, as illustrated in Figure 4.

Ideally, the difference between

N1 = #web pages found from "(A or B)" and
N2 = #web pages found from "(A not B) + (B not A) + (A and B)"

should be zero. Figure 4 shows the logarithm (base 10) of the absolute value of this difference, log(|N1-N2|), versus log(N2). As the number of web pages found increases, the discrepancy between N1 and N2 grows, with large N2 (on the order of 100 million web pages) resulting in discrepancies of over 1 million. For the most part, though, this figure shows that the discrepancy is usually less than 1%, but can sometimes be as large as 10%. As search engine algorithms improve, we expect fewer large discrepancies of this type.

This study attempts to address the extent of the synonym problem by comparing the numbers of web pages found by only one of the two words in a synonym word-pair but not the other word. However, a user's main interest may be in capturing not the *total* number of web sites for a given concept but rather the number of *most relevant* web sites. The results from this type of study would indeed be interesting, but we see two immediate problems in attempting to conduct such a study. First, one would have to define what is meant by "most relevant". The easiest definition would be "top 25 web sites", but some of those "top 25" could be duplicates, irrelevant, non-authoritative, or paid by advertisers. Moreover, human subjectivity would be involved in assessing "relevance". At some future time, search engines may offer functionalities that would reduce the human effort in this time-intensive, possibly subjective, laborious process, and we would consider such a study at that point.

Another issue may be whether our study measured "semantic exactness" rather than "extent of the synonym problem". Our criteria for word-pair synonyms in this study included one criterion that was aimed at achieving a high degree of homogeneity in semantic exactness, but this criterion did involve some human judgment. The confounding of these two concepts, "semantic exactness" and "extent of synonymy", may be difficult to resolve with present technology.

Conclusion

The extent of the synonym problem in full-text searching depends on whether one searches the more common of the synonyms. Overall, the measure of what’s missed is as high as 30% in a large (90%) fraction of common word-pairs. Information discovery systems need to take the synonym problem into account and develop solutions for it, both probabilistic and deterministic. This study should be repeated with a wider and more systematic variety of synonym pairs from defined subject areas; searches that include phrases instead of single words in the pairs; replication, to determine the variability in the reported percentages; and more search engines. The methodology here could result in the establishment of a benchmark data set against which various search engines can evaluate their search algorithms in terms of their ability to minimize the synonym problem. Additionally, the data demonstrate the value of vocabulary control and cross references in providing more precise search results.

Acknowledgements: The authors wish to thank the Associate Editor and two anonymous referees for their thoughtful comments on an earlier version of this manuscript. This work was prepared in part with support from Army Research Office, Grant #W911NF0510490, awarded to the University of Colorado Denver (Kafadar).

Works Cited

Bade, David. "Relevance Ranking is Not Relevance Ranking or, When the User is Not the User, the Search Results are Not the Search Results." Online Information Review 31.6 (2007): 831-44.

Beall, Jeffrey. "The Weaknesses of Full-Text Searching." Journal of Academic Librarianship 34.5 (2008): 438-44.

---. "The Death of Full-Text Searching." PNLA Quarterly 70.2 (2006): 5-6.

Dagan, Ido, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. "Direct Word Sense Matching for Lexical Substitution." COLING-ACL 2006: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia, 17-21 Jul. 2006: 449-56. 11 Nov. 2008 <http://eprints.pascal-network.org/archive/00002759/01/P06-1057.pdf>.

Dean, Jeffrey A., Georges Harik, Benedict Gomes, and Noam Shazeer. Methods and Apparatus for Determining Equivalent Descriptions for an Information Need. Google Inc., assignee. Patent 6,941,293. 6 Sep. 2005.

Fugmann, Robert. "The Complementarity of Natural and Controlled Languages in Indexing," Subject Indexing: Principles and Practices in the 90's : Proceedings of the IFLA Satellite Meeting held in Lisbon, Portugal, 17-18 August 1993, and sponsored by the IFLA Section on Classification and Indexing and the Instituto da Biblioteca Nacional e do LIVRO, Lisbon, Portugal. Eds. Robert P. Holley, et al. Munich: Saur, 1995. 215-30.

Mann, Thomas. The Oxford Guide to Library Research. 3rd ed. Oxford: Oxford UP, 2005.

Pregibon, Daryl, and Diane Lambert. "Understanding Online Advertisers." Joint Statistical Meeting, Denver, CO, USA, 5 August 2008.

"Web Search Help Center." Google. 2008. 11 Nov. 2008 <http://www.google.com/help/refinesearch.html>.

Appendix

Table 2
The data collected in this study. "A" is designated as the more common word in the synonym pair, "B" as the less common word.

Number	Terms	max	min	A (not B)	B (not A)	A and B	A or B
1.	A. Adivasis B. Adibasis	0.9979	0.0016	109,000	177	50	110,000
2.	A. Afrocentricity B. Afrocentrism	0.5198	0.4717	63,800	57,900	1,040	128,000
3.	A. Aluminum B. Aluminium	0.6850	0.2853	63,400,000	26,400,000	2,750,000	92,500,000
4.	A. Anomie B. Anomy	0.8556	0.1429	443,000	74,000	776	521,000
5.	A. Appraisers B. Assessors	0.7258	0.2597	6,400,000	2,290,000	128,000	8,850,000
6.	A. Arctiidae B. Lithosiidae	0.9971	0.0016	85,200	138	106	85,800
7.	A. Arthropods B. Arthropoda	0.6061	0.2905	1,400,000	671,000	239,000	2,310,000
8.	A. Berberis B. Barberries	0.9407	0.0565	368,000	22,100	1,090	393,000
9.	A. B.						
10.	A. Biologics B. Biologicals	0.7582	0.2275	2,290,000	687,000	43,200	3,030,000
11.	A. Bleaching B. Blanching	0.9281	0.0715	3,390,000	261,000	1,510	3,660,000
12.	A. Buddhists B. Lamaists	0.9998	0.0001	3,070,000	386	276	3,090,000
13.	A. B.						
14.	A. Bullying B. Bullyism	0.9999	MM	10,100,000	984	160	10,100,000
15.	A. Cachexia B. Cachexy	0.9899	0.0069	245,000	1,720	784	253,000
16.	A. Cannibalism B.Anthropophagy	0.9946	0.0044	2,440,000	10,800	2,330	2,470,000
17.	A. Catalans B. Catalonians	0.9941	0.0056	2,100,000	11,900	643	2,130,000
18.	A. B.						
19.	A. Chimneys B. Smokestacks	0.8340	0.1627	2,820,000	550,000	11,200	3,410,000
20.	A. Chefs B. Cooks	0.5327	0.4176	23,600,000	18,500,000	2,200,000	44,200,000
21.	A. Deacons B. Diaconate	0.9085	0.0578	2,720,000	173,000	101,000	3,000,000
22.	A. Deburring B. Burring	0.8388	0.1557	513,000	95,200	3,410	618,000
23.	A. B.						

24.	A. Discrimination B. Bias	0.5336	0.4314	44,400,000	35,900,000	2,910,000	83,200,000
25.	A. Dreams B. Dreaming	0.8367	0.1246	133,000,000	19,800,000	6,160,000	159,000,000
26.	A. Egoism B. Egocentricity	0.9149	0.0843	1,100,000	93,000	978	190,000
27.	A.Electromagnetism B. Electromagnetics	0.5447	0.4287	953,000	750,000	46,500	1,760,000
28.	A. Embezzlement B. Defalcation	0.9738	0.0230	2,460,000	58,200	7,900	2,540,000
29.	A. Errors B. Mistakes	0.7010	0.2570	162,000,000	59,400,000	9,690,000	231,000,000
30.	A. Eurocentrism B. Eurocentricity	0.9806	0.0178	104,000	1,890	171	113,000
31.	A. Eviction B. Dispossession	0.8953	0.1005	4,590,000	515,000	21,800	5,120,000
32.	A. Extraversion B. Extroversion	0.6307	0.3372	432,000	231,000	22,000	684,000
33.	A. Faience B. Fayence	0.7405	0.2580	861,000	300,000	1,790	1,170,000
34.	A. Fasteners B. Fastenings	0.9559	0.0373	12,100,000	472,000	86,800	12,700,000
35.	A. Fireworks B. Pyrotechnics	0.9453	0.0402	32,200,000	1,370,000	495,000	34,200,000
36.	A. Forearm B. Antebrachium	0.9996	0.0003	4,510,000	1,310	578	4,500,000
37.	A. Formaldehyde B. Formalin	0.7469	0.2196	2,500,000	735,000	112,000	3,350,000
38.	A. Gelatin B. Gelatine	0.7867	0.2003	4,360,000	1,110,000	72,400	5,550,000
39.	A. Greenhouses B. Hothouses	0.9841	0.0152	5,090,000	78,600	3,390	5,170,000
40.	A. Gums B. Gingiva	0.9664	0.0263	4,780,000	130,000	36,200	4,940,000
41.	A. Heme B. Hematin	0.9862	0.0119	975,000	11,800	1,800	995,000
42.	A. Hydrogeology B. Geohydrology	0.9327	0.0525	917,000	51,600	14,600	978,000
43.	A. Intellectuals B. Intelligentsia	0.8409	0.1303	7,810,000	1,210,000	268,000	8,710,000
44.	A. Ischemia B. Ischaemia	0.8551	0.0967	2,060,000	233,000	116,000	2,380,000
45.	A. Kayasthas B. Kayasths	0.8249	0.1613	1,790	350	30	2,420
46.	A. Kimchi B. Kimchee	0.7670	0.2070	930,000	251,000	31,500	1,220,000
47.	A. Lakes B. Lochs	0.9858	0.0130	73,400,000	731,000	90,200	73,100,000
48.	A. Larrea B. Covillea	0.9985	0.0003	568,000	133	82	558,000
49.	A. Libertinage B. Libertinism	0.8570	0.1421	468,000	77,600	460	563,000
50.	A. Marmots B. Marmota	0.5135	0.4724	325,000	299,000	8,900	645,000
51.	A.	0.9999	0.0001	80,700	5	1	80,700

	Mechanoreceptors B. Mechanicoreceptors						
52.	A. Micropipettes B. Micropipets	0.9662	0.0321	72,500	2,410	124	81,300
53.	A. Mitochondria B. Chondriosomes	0.9999	0.0001	2,000,000	144	45	2,000,000
54.	A. Monazite B. Cryptolite	0.9979	0.0018	112,000	207	24	113,000
55.	A. Mutuality B. Mutualism	0.7695	0.2295	798,000	238,000	1,070	1,050,000
56.	A. Natriuresis B. Natruresis	0.9982	0.0012	59,600	71	35	60,000
57.	A. Norsemen B. Northmen	0.7713	0.2206	465,000	133,000	4,860	613,000
58.	A. Ochre B. Ocher	0.8799	0.1151	1,460,000	191,000	8,260	1,660,000
59.	A. Ointments B. Salves	0.6694	0.2958	1,430,000	632,000	74,300	2,140,000
60.	A. Ontogeny B. Ontogenesis	0.8210	0.1428	638,000	111,000	28,100	778,000
61.	A. Organotherapy B. Opotherapy	0.8937	0.066	2,750	203	124	3,420
62.	A. B.						
63.	A. B.						
64.	A. Paramecium B. Paramaecium	0.9282	0.0712	314,000	24,100	207	339,000
65.	A. Parsis B. Parsees	0.7687	0.2228	177,000	51,300	1,960	237,000
66.	A. Pediatrics B. Paediatrics	0.9002	0.0806	18,200,000	1,630,000	388,000	20,200,000
67.	A. Perimenopause B. Premenopause	0.8164	0.1475	631,000	114,000	27,900	772,000
68.	A. Photogravure B. Heliogravure	0.8237	0.1579	395,000	75,700	8,840	487,000
69.	A. Picornaviridae B. Picornaviruses	0.5094	0.4600	39,200	35,400	2,360	84,200
70.	A. Pollination B. Pollinization	0.9988	0.001	2,080,000	2,070	382	2,100,000
71.	A. Porpoises B. Phocoenidae	0.9720	0.0252	709,000	18,400	2,000	773,000
72.	A. Preparedness B. Readiness	0.4959	0.4757	17,200,000	16,500,000	986,000	39,000,000
73.	A. B.						
74.	A. B.						
75.	A. Procellariiformes B. Tubinares	0.9599	0.0293	55,700	1,700	627	60,300
76.	A. Promethium B. Illinium	0.9933	0.0053	137,000	735	185	144,000
77.	A. Radiologists B. Roentgenologists	0.9998	0.0001	1,740,000	221	133	1,770,000

78.	A. Religiosity B. Religiousness	0.8444	0.138	1,040,000	170,000	21,600	1,240,000
79.	A. Rodents B. Rodentia	0.9456	0.0407	6,070,000	261,000	88,000	6,420,000
80.	A. Sago B. Sagu	0.8664	0.1319	1,340,000	204,000	2,690	1,550,000
81.	A. Salafīyah B. Salafiyya	0.5899	0.4079	17,500	12,100	66	38,700
82.	A. Metalloids B. Semimetals	0.6761	0.3187	77,000	36,300	595	116,000
83.	A. B.						
84.	A. Shepherds B. Sheepherders	0.9855	0.0143	6,500,000	94,000	1,610	6,600,000
85.	A. B.						
86.	A. Shrews B. Soricidae	0.8917	0.0895	485,000	48,700	10,200	551,000
87.	A. Skunks B. Mephitidae	0.9983	0.0008	1,270,000	980	1,210	1,280,000
88.	A. Slavists B. Slavicists	0.9725	0.0253	27,100	704	61	30,400
89.	A. Somite B. Metamere	0.9646	0.0317	76,400	2,510	294	85,300
90.	A. Spires B. Steeples	0.8532	0.1362	2,950,000	471,000	36,400	3,460,000
91.	A. Stigmata B. Stigmatization	0.7671	0.2323	1,420,000	430,000	1,080	1,860,000
92.	A. Summer B. Summertime	0.9779	0.0135	396,000,000	5,480,000	3,470,000	403,000,000
93.	A. Tinsmithing B. Tinwork	0.5368	0.4606	16,200	13,900	79	41,700
94.	A. Trilobites B. Trilobita	0.9052	0.0841	339,000	31,500	3,990	384,000
95.	A. Urea B. Carbamide	0.9579	0.0356	3,790,000	141,000	25,500	3,970,000
96.	A. Vietnamese B. Annamese	0.9997	0.0002	45,000,000	10,100	8,800	45,200,000
97.	A. Violin B. Fiddle	0.6499	0.2931	20,400,000	9,200,000	1,790,000	31,500,000
98.	A. Virilization B. Virilism	0.8697	0.1218	67,500	9,450	665	82,800
99.	A. Tanka B. Waka	0.5063	0.4903	1,580,000	1,530,000	10,700	3,110,000
100.	A. Wrasses B. Labridae	0.6217	0.3072	118,000	58,300	13,500	196,000