Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 381–388
ISBN 978-83-60810-14-9
ISSN 1896-7094

Modeling the Frequency of Phrasal Verbs with Search Engines

Grażyna Chamielec
SuperMemo
Poznan, Poland
ika.chamielec@gmail.com

Dawid Weiss
Institute of Computer Science
Poznan University of Technology
Poznan, Poland
dawid.weiss@cs.put.poznan.pl

***Abstract*—There are well over a thousand phrasal verbs in English. For non-native speakers they are notoriously difficult to remember and use in the right context. We tried to construct a ranking of phrasal verbs according to their estimated occurrence frequency, based on quantitative information available from the public indexable Web. Technically, we used major Web search engines to acquire phrase-occurrence statistics, measured consistency between the rankings implied by their results and confirmed that a rough set of 'classes' of phrasal verbs can be distinguished.**

While this technique relies on inaccurate and possibly biased estimation functions, we show that the overall distribution of ranks seems to be consistent among all the queried search engines operated by different vendors.

I. Introduction

A *PHRASAL verb* is, according to Oxford Advanced Learner's Dictionary [1]:

> [...] a simple verb combined with an adverb or a preposition, or sometimes both, to make a new verb with a meaning that is different from that of the simple verb, e.g., *go in for*, *win over*, *blow up*.

There are a number of phrasal verbs in both spoken and written English ([2] lists over 6000 entries). As the definition states, the meaning of a phrasal verb cannot be easily guessed from individual components—many non-native speakers of English must therefore memorize phrasal verbs in order to be able to understand and use them in the right context. Our motivation for this work was a direct consequence of this observation.

SuperMemo[1] is a company specializing in helping people learn fast, use memory efficiently and aid in self-improvement processes. SuperMemo's line of products include, among others, dictionaries and language courses. While working on a list of English phrasal verbs, we stated the following problem:

- Which phrasal verbs should be memorized first?

There are two other related questions:

- Are there any phrasal verbs that are hardly ever present in a 'live' corpora of written language?
- Are there groups of 'frequent' and 'infrequent' phrasal verbs and is it possible to distinguish these groups?

There is certainly no definite answer to these questions; phrasal verbs and their meaning will vary by region and dialect of English, for example. Our research intuition was telling us though, that by relying on a really large corpora of existing texts rather than book resources or dictionaries, we could come out with a fairly good estimate on which phrasal verbs are common and which are infrequent. In other words, we wanted to measure possibly 'real' average occurrence frequency of each phrasal verb, then sort them in the order of this estimated frequency and distinguish several groups that could provide the basis for the construction of a training course.

[1] http://www.supermemo.com

II. Related work and discussion

There exist a number of dictionaries [2], [3], books and papers concerning phrasal verbs and verb-particle associations at the linguistic layer. There are also on-line resources listing phrasal verbs and providing their meanings. However, we failed to find any resource that would attempt to *quantitatively* measure the frequency of use of phrasal verbs. The paper by Timothy Baldwin and Aline Villavicencio came closest to our expectations [4]. In this work, authors process raw text of the Wall Street Journal corpus using a number of different methods to identify verb-particle occurrences. The best technique reached the f-score of 0.865. The experiment in [4] was performed on an established corpus of press resources. While using a corpus like this (or a balanced language corpus in general) has many advantages, we wanted to stick to the Web because it reflects many different language users, use cases and is a great deal larger than any other corpus available. Although there are various opinions about the coverage of the Web, its information quality and bias (see [5] or [6] for an interesting discussion), we believe that in our case these aspects can be neglected and search engines provide suitable source of knowledge to answer the questions given in the introduction. Obviously, any research based on uncontrolled, proprietary information sources such as search engines should be approached with care. We tried to do our best to cross-validate the results against multiple vendors to make them more confident.

III. Proposed methodology

Every search engine returns an estimation of the number of documents 'matching' a given query (note that this is the number of *documents*, not individual instances of the query). Figure 1 illustrates a query results page with the rough number

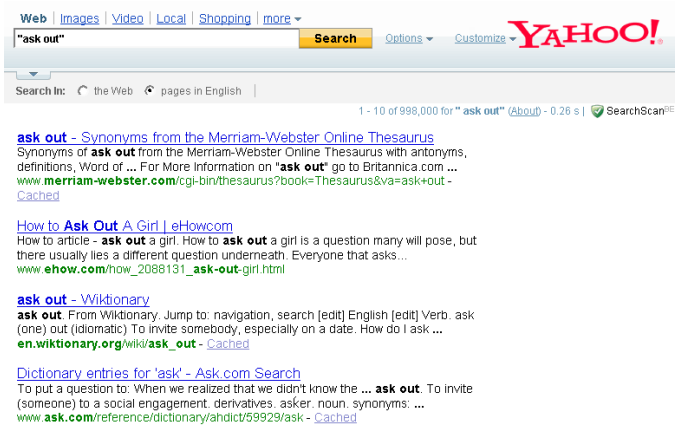

Fig. 1. Query results page from Yahoo search engine. The red rectangle marks the status line displaying the number documents matching the query.

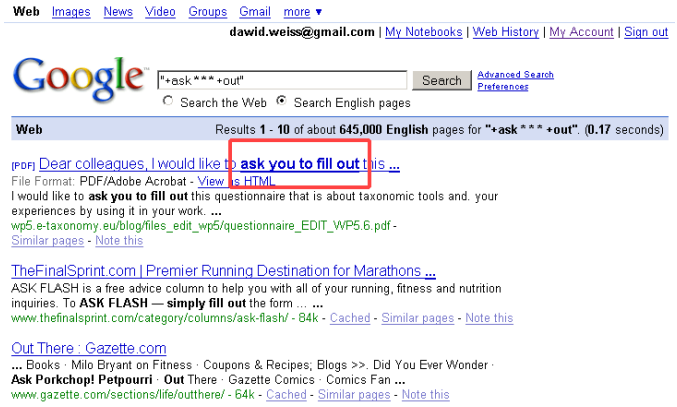

Fig. 2. A wildcard query may result in a false match (see the marked phrase).

of documents matching the exact phrase 'ask out'. While the returned number is merely an estimate and may be inaccurate (we discuss this in Section V), we assumed that the estimation is correct at least to the order of magnitude, thus properly dividing frequent and relatively infrequent phrasal verbs.

IV. Phrasal Verbs Considered

We used a hand-crafted set of phrasal verbs (PV) collected from several on-line resources and books. We made an explicit distinction between separable and inseparable PVs where it was appropriate and placed an asterisk (wildcard) character in places where separation could occur. Then, to every verb we assigned a number of different forms in which it could possibly appear in the text, depending on its tense. Table 3 illustrates an example phrasal verb pattern and all its corresponding variations. The pattern-based representation was used to drive queries to search engines.

V. Potential Ambiguities and Other Problems

There are a few corner cases in counting the number of documents containing a given phrasal verb and they are all a consequence of how text information retrieval methods (implemented in search engines) work. In simplest terms, search engines transform a document into a vector of individual words and their *weights* (relative importance of a given word to the document). This representation of text is called the vector space model [7]. A query to a search engine returns all documents that contain a union of the query's set of words (possibly ordered), but it is rarely possible to specify deeper contextual constraints. Let us explain the possible side-effects of this process on a few examples.

The first problem is that not every word pattern corresponds to an actual phrasal verb. For example, *[to] be in* can appear as *I'm in*, but the sole appearance of this sequence of words without the knowledge of the context may be a false hit (*I'm in Poland right now.*). Unfortunately this will be the case with most verbs that have transitive and intransitive forms. Another issue is caused by multiple meanings of a single phrasal verb, compare *throw up* (vomit) and *throw up* (an idea). Detecting and separating the meaning of these two expressions seems impossible assuming the measurement technique we agreed to use.

The final example concerns separable forms of phrasal verbs. What we intend to do is to query for patterns (sequences of words) that have a few words in between (but not too many). For example, *sign me in* should be counted as an occurrence of *sign in*. However, simply allowing words to appear in between components of a phrasal verb may lead to many mistakes. For instance, as Figure 2 illustrates, the three top-ranked documents for a query *ask out* separated by three other words, are basically wrong. There seems to be no way of filtering out this noise without more complex linguistic analysis (if we had access to whole document content, as in a controlled corpus, we could use POS tags for getting rid of such errors).

Regardless of the above problems, we decided to calculate occurrence statistics and proceed with the experiment. It is our assumption that the number of false matches for less than three wildcards can be neglected compared to the number of true matches (at least for common phrasal verbs). As for phrasal verbs with multiple meanings, all occurrences of these meanings sum up to one figure which reflects the aggregated use of a given sequence of words. Since so, the final ranking

ask/asks/asked/asking * out			
ask out	ask – out	ask – – out	ask – – – out
asks out	asks – out	asks – – out	asks – – – out
asked out	asked – out	asked – – out	asked – – – out
asking out	asking – out	asking – – out	asking – – – out
back/backs/backed/backing off			
back off	backs off	backed off	backing off
crack/cracks/cracking/cracked * up			
crack up	crack – up	crack – – up	crack – – – up
cracks up	cracks – up	cracks – – up	cracks – – – up
cracked up	cracked – up	cracked – – up	cracked – – – up
cracking up	cracking – up	cracking – – up	cracking – – – up

Fig. 3. An example of phrasal verb patterns and matching word sequences. An asterisk (*) symbol represents between zero and three words appearing in its position, we denoted these words using the dash symbol on the right (–).

position is to some extent indicative of the need to learn a given phrasal verb (even if it is ambiguous).

VI. Collecting occurrence statistics

We collected occurrence statistics from several search engines: Google (`www.google.com`), Yahoo (`www.yahoo.com`), AllTheWeb (`www.alltheweb.com`), Gigablast (`www.gigablast.com`) and Microsoft Live (`www.live.com`). With the exception of Gigablast and Microsoft Live, the remaining providers all support the so-called *wildcard queries*, i.e., a query for all documents containing a given phrase separated by one or more unrestricted words inside. With wildcard queries we could estimate the number of occurrences of separable phrasal verbs by querying for the exact phrase, phrase with one, two and three extra words at the point of possible separation. For example, the entry (to simplify, we only show one verb form here):

```
ask * out
```

would result in the following queries to a search engine:

```
ask out
ask * out
ask * * out
ask * * * out
```

An exact format of queries submitted to each search engine varied depending on the service provider's syntax and we omit it here, although we found out that such details are quite crucial because search engines employ various optimizations and query expansion techniques that, in our case, distorted the output. As previously observed in [5], the returned estimation counts have some significant variance within the same query (the same search engine would return a different document count for consecutive executions of an identical query). We took this into account and put together 10 identical query lists, randomized their order and executed all queries at different times and from different machines. Finally, we paid particular attention to restricting the search to documents in the English language and to searching within document content only (exclude links pointing to the page).

The process of querying search engines was partially automated and performed in accordance with each search engine's policies and terms of use specifications (timeouts between queries, use of automated programming interfaces when possible).

VII. Results

Overall, we collected frequency counts for 10 633 various separable and inseparable forms of 991 phrasal verb patterns (some of these were closely related, like *blend in* and *blend into*). For each form, we stored the estimated document count for each of the 10 'samples' made to each single search engine. Even though the querying process was semi-automatic, it lasted over three days (because we had to add the required timeouts between queries) and involved over 30 machines (with different IP addresses). If we had been given access to the search engine's infrastructure, such processing could be made much faster and more accurately by running shallow grammar parsing on the content of each document, splitting the process over multiple machines using the map-reduce paradigm.

We describe the results from several angles in sub-sections below.

A. Differences between search engines

Every search engine is a bit different—these differences usually concern the number of indexed documents, ranking algorithms and technical aspects of estimating the number of matching documents. Our first step was to cross-compare the numbers returned from various search engines to see if they share similar distribution and what shape this distribution is.

We took one sample out of the ten made and for each phrasal verb form we compared document counts between search engines by sorting all forms according to the number of documents returned by Yahoo, placing them (in this order) on the horizontal axis and plotting document counts on the vertical axis. Figures 4–7 demonstrate the results. The overall distribution shape for all search engines is for the most part exponential (vertical axis is on logarithmic scale). Exponential distribution confirms our initial intuition that a small number of phrasal verbs occurs frequently and a great deal of them are relatively infrequent on the Web.

Back to differences between search engines, we can observe notable differences in average document counts between different search engines, but highly correlated distribution shapes. This validates our assumption that search engines are a methodologically sound tool to 'probe' the Web. If (theoretically) we consider the Web to be a global population of documents, then the index of each search engine is basically a random sample taken from this population. If so, the average count of documents between two search engines should be linearly proportional to the degree of a constant multiplier. Another way to put it is that the *ordering* of phrasal verb forms imposed by all search engines should be very similar between search engines. A look at Figures 4–7 and especially at log-log plots in Figure 8 reveals that all search engines returned correlated results. For example, Yahoo and AllTheWeb's results are almost identical (Figure 4) because AllTheWeb's index is powered by Yahoo; minor differences may be a result of different search query routing inside Yahoo's infrastructure. There is also an evident high similarity between Yahoo, Gigablast and Microsoft Live's results (see log-log plots in Figure 8, although Microsoft and Gigablast have an order of magnitude smaller index. The only visibly different engine is Google—not only has it fewer documents compared to Yahoo, but also its count distribution is strikingly different compared to other search engines (although still correlated). Narrowed to only non-wildcard forms, the distribution difference is even more strange because it shows two different 'traces' of frequency distribution in the area of more frequent phrasal verbs (see Figure 8).

We initially thought this difference in Google's case might be caused by the fact that it has the largest infrastructure and queries may be routed to separate index sections, leading to

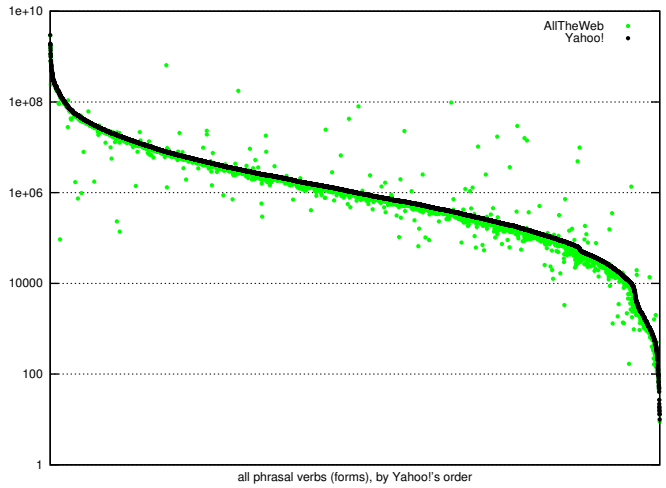

Fig. 4. Document counts for results acquired from AllTheWeb and Yahoo (sorted by Yahoo's results—the black line).

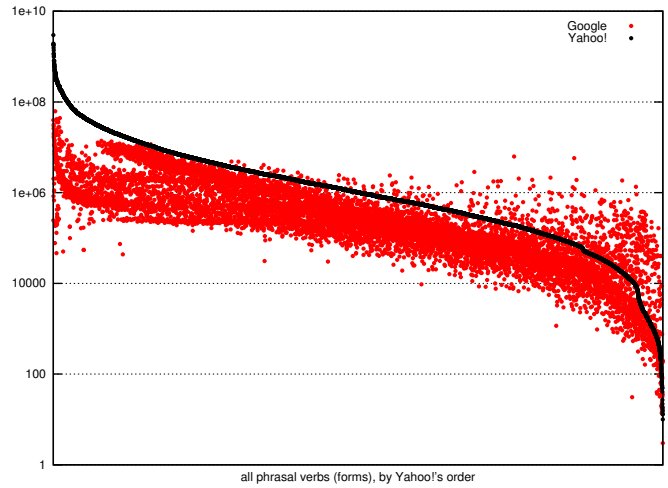

Fig. 5. Document counts for results acquired from Google and Yahoo (sorted by Yahoo's results—the black line).

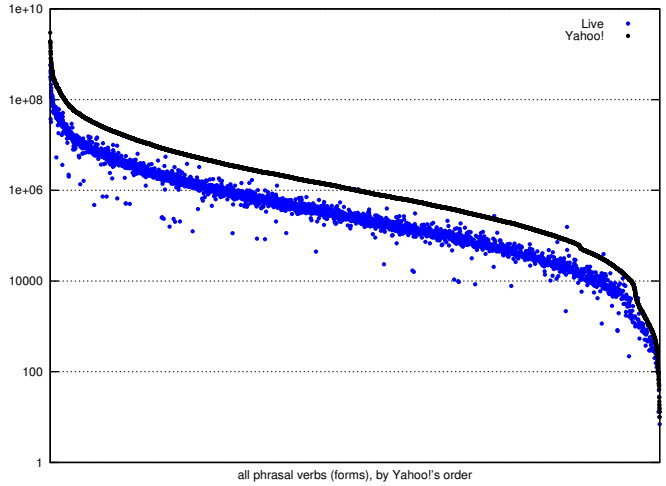

Fig. 6. Document counts for results acquired from Microsoft Live and Yahoo (sorted by Yahoo's results—the black line).

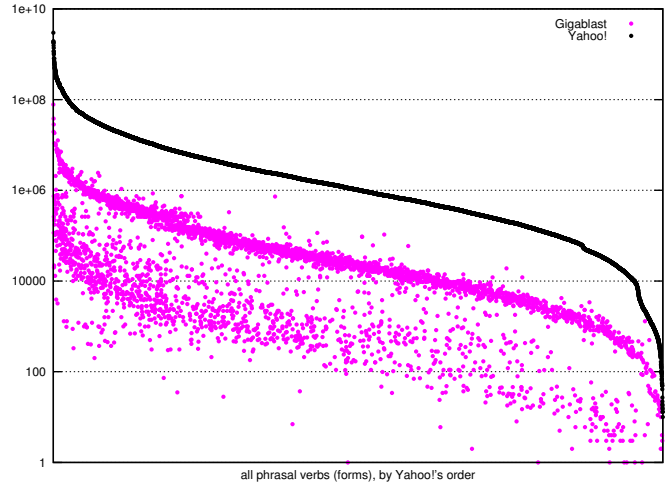

Fig. 7. Document counts for results acquired from Gigablast and Yahoo (sorted by Yahoo's results—the black line).

different estimated count of results. We took a closer look at all ten samples for each query, calculating minimum, maximum, median and a truncated average (average of 6 samples after sorting and removing two minimum and maximum outliers). The outcome of this analysis is that, again, Google has the largest variation between estimated result count for a single query (refer to technical report [8] for a more in-depth analysis). In case of Yahoo the difference between minimum and maximum number of results is relatively small, usually the same. Microsoft Live returns a fairly consistent range of difference—usually in the order of magnitude—with the truncated average usually equal to the maximum. For Google, the difference between min and max is again the order of magnitude, but the average is less predictable and is usually in between min and max (see Figure 9).

B. *Phrasal verb rankings (groups)*

We constructed a *ranking* of phrasal verbs according to their totaled frequency of occurrence on the Web. Note that actual positions in this ranking are a product of multiple heuristics and their values should not be compared directly. The overall ordering should merely help to distinguish subgroups of frequent and infrequent phrasal verbs, as was our initial motivation for this research.

We experimented with many different ways of aggregating information from all samples and forms of each phrasal verb. We produced multiple possible rankings based on the following algorithm steps:

1) for every search engine, aggregate all samples for each phrasal verb form `form_id`, calculate minimum, maximum, median and truncated average (`avg2`) from document counts;
2) consider all variations: forms with $<=$ 0, 1, 2 and 3 wildcards;

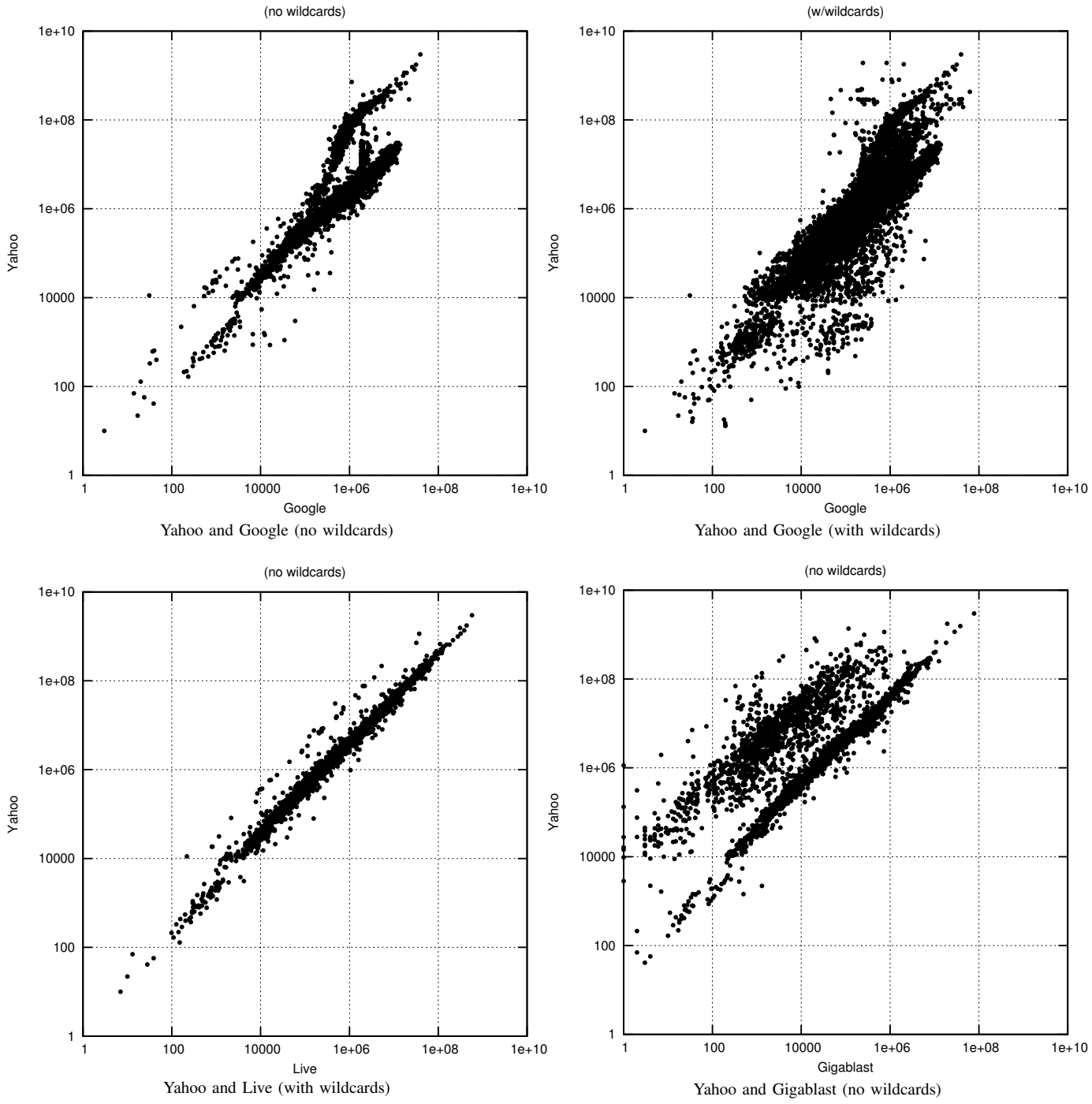

Fig. 8. Log-log plot of document counts between Yahoo and other search engines.

3) sort in descending order all forms according to minimum, maximum, median and avg2 column, assign a rank to each `form_id`;
4) assign a minimum rank of any of its forms to each phrasal verb `pv_id`.

The above procedure has several variables which cause numerous possible variations of output rankings (depending on the engine, number of wildcards and the order column being considered). These rankings, consistently with our previous observations, demonstrate close similarity to each other within a single search engine and between Yahoo, AllTheWeb and Microsoft Live. Only Google is an exception. To give a few examples, the choice of the sorting column did not have much impact on the actual ranking within a single search engine. Cross-engine ranking consistency is shown on plots in Figure 10. The correlation of ranks (measured with correlation coefficient, which in this case equals to Spearman's rank coefficient) between AllTheWeb, Yahoo and Microsoft Live

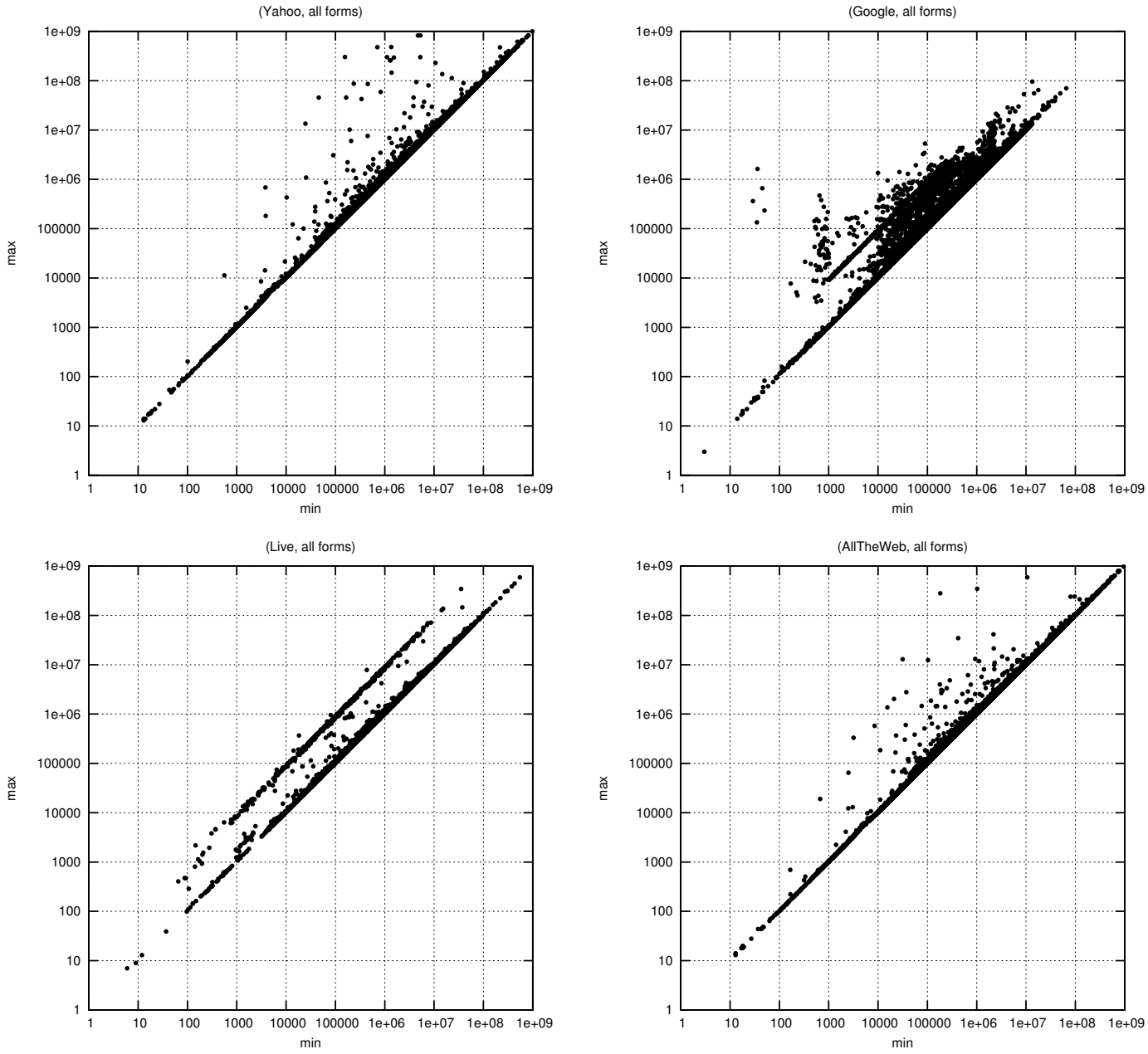

Fig. 9. Relationship between minimum and maximum number of results out of 10 samples for each phrasal verb form (Yahoo, Google, Microsoft Live and AllTheWeb).

was evident and larger than 0.9 for all considered combinations of rank computations. Google is distinctly different from other search engines, but the correlation coefficient is still quite high—between 0.7 and 0.8. We have no clear explanation as to why Google's results turn out to be slightly different than obtained from other search engines.

Even though all rankings were highly correlated, they were still a bit different from each other, so there is no ultimate one answer to our initial question of 'frequent' and 'infrequent' phrasal verbs. Without a doubt the rankings themselves reflect the nature of Web resources (see Table I) by, e.g., boosting phrases common in e-commerce (*sign up*, *check out*). Yet, a tentative and subjective feeling is that the top entries are indeed something that every native user of English should be familiar with and bottom ranking entries are extremely rare, uncommon or denote mistakes in the data set (see Table II).

VIII. Summary and conclusions

We tried to create a ranking of phrasal verbs according to their frequency of actual use on the Web. We designed and performed a computational experiment, measuring estimated document count using several independent search engines. We think the outcomes are interesting from two different viewpoints: the linguistic one and the one concerning (dis)similarities across contemporary search engines, which turn out to be quite intriguing.

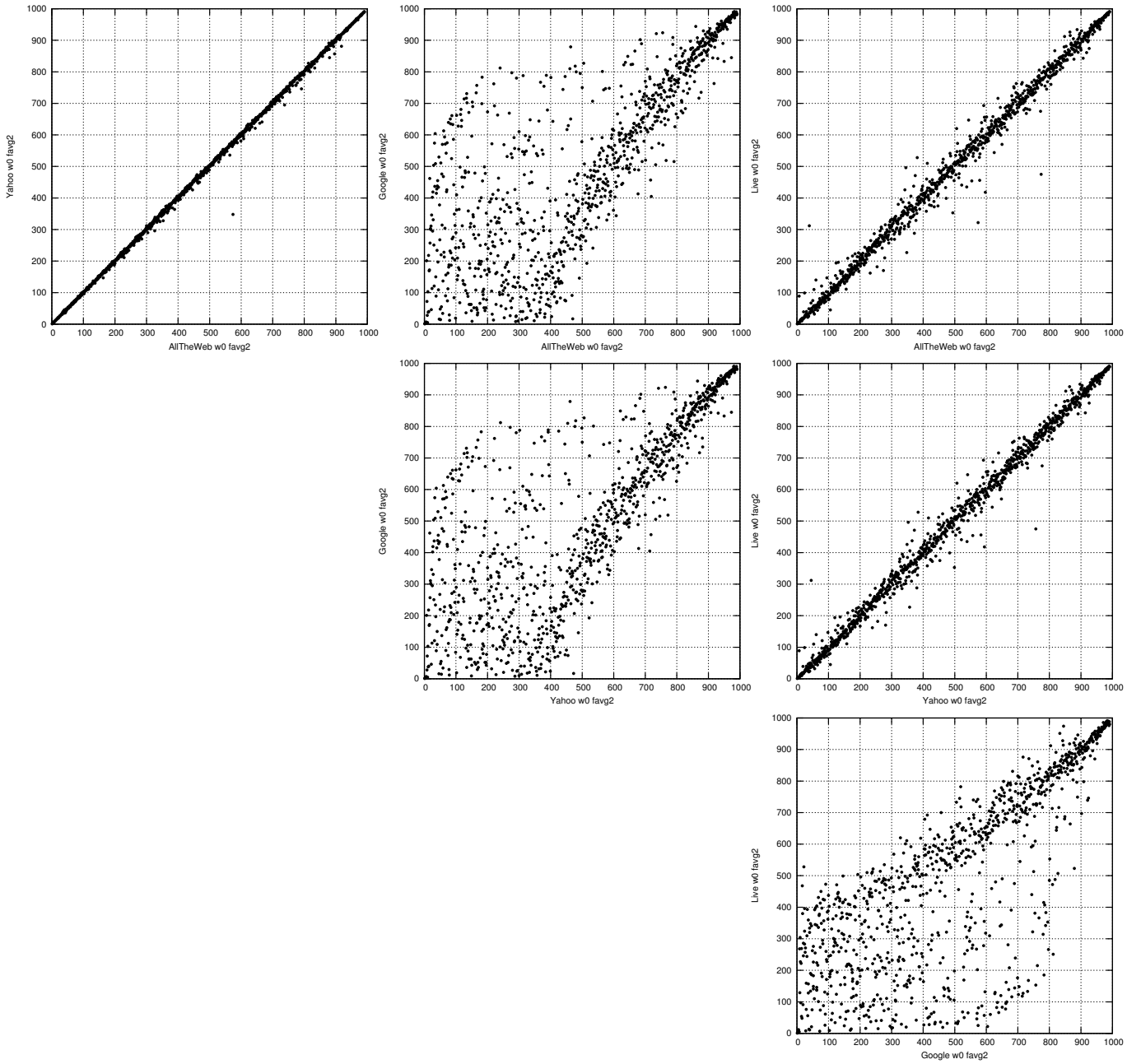

Fig. 10. Relationship between phrasal verb ranks depending on the search engine (search engine on horizontal and vertical axes fixed for rows and columns). Other parameters fixed to: `avg2` column used for sorting, zero wildcards.

As for the linguistic aspect, we are not aware of such search engine based measurement of the frequency of phrasal verbs, although search engines have been used for conducting linguistic experiments before. We think there are clear indications to believe that such an analysis can yield valid results, allowing one to separate frequent and infrequent phrasal verbs. A number of challenging problems remain unsolved:

- Even though the Web is very large, it is also biased; especially phrases that relate to e-commerce are boosted high up the ranking (*sign up*, *check out*). In our case this was not a problem because the rankings (groups) were edited manually for the final application after they were acquired anyway, but in other scenarios this is a problem.
- We currently see no way of disambiguating multi-sense phrasal verbs or no-object phrasal verbs. Given access to the full content of search engine's documents, shallow NLP techniques could be employed here.
- We used wildcard queries and multiple tenses for fetching various potential forms of phrasal verbs. It turned out that this had very little influence over final rankings; is such a step necessary or would it be enough

TABLE I
TOP 10 PHRASAL VERBS ACCORDING TO YAHOO, GOOGLE AND LIVE (0 WILDCARDS, AVG2).

No.	Yahoo	Google	Live
1	sign up	sign up	sign up
2	look for	look for	look for
3	check out	be in	be in
4	be in	check out	check out
5	look at	go back	find out
6	find out	look at	look at
7	arise from	find out	set up
8	come to	be after	come to
9	set up	look in	get to
10	go back	start off	work on

TABLE II
SELECTED 10 PHRASAL VERBS FROM THE BOTTOM OF THE RANKING FOR YAHOO, GOOGLE AND LIVE (0 WILDCARDS, AVG2).

Yahoo	Google	Live
slug out	sob out	fur up
winkle out	slog out	suture up
fur up	swirl down	ravel out
satire up	nestle up	sponge down
skirt round	fur up	push round
sponge down	rein back	hiss off
tail away	skirt round	slog out
be bombed out	sponge down	rap put
slog out	ravel out	scorch along
hiss off	scorch along	stream down upon

to just limit the analysis to present-tense forms?

- The distribution of document counts returned from search engines is exponential, so one could make groups of phrasal verbs each falling into bins related to the frequency's order of magnitude. However, there is no clear dividing line between these bins and there is certainly some room for improvement here.

From the point of view of a researcher interested in search engines, this work provides an interesting insight into differences between major search providers, especially with regard to the estimated matching document set size.

- Yahoo is by far the most *consistent* search engine and its returned estimation does not vary much between the same queries issued at different times,
- Yahoo and Microsoft Live show very correlated counts—nearly identical, in fact. This follows our intuition about 'sample from a large corpus', but is contradicted by results returned by Google. We cannot explain why Google is so much different compared to Yahoo and Live.
- Google and Live return document counts (for the same query) that vary by an order of magnitude.

As for further work on this subject, it would be quite interesting to examine phrasal verb distribution using exact NLP methods (or shallow, but with linguistic context taken into account) on a larger free corpora (such as Wikipedia or a free crawl of the Web) and compare the rankings with those we acquired from search engines. Such effort would allow validating and deriving further conclusions concerning the accuracy of our method. Alternatively, one could try to estimate the estimation error by taking the results returned from a search engine, manually tagging the returned documents as false/ true matches and then establishing true/false hit ratio. This method is used successfully in software engineering to establish the true number of software defects given a number of unreliable referees assessing code quality. Access to input lists of phrasal verbs, crawl results and rankings is given at the following address: http://www.cs.put.poznan.pl/dweiss/research/pv/.

Acknowledgment We are very grateful for anonymous reviewer feedback and his or her constructive suggestions and comments. The work was carried out with the financial support of Polish Ministry of Science and Higher Education.

REFERENCES

[1] *Oxford Advanced Learner's Dictionary*. Oxford University Press, 1995.
[2] *Oxford Phrasal Verbs Dictionary for Learners of English*. Oxford University Press, 2007.
[3] *Cambridge Phrasal Verbs Dictionary*. Cambridge University Press, 2006.
[4] T. Baldwin and A. Villavicencio, "Extracting the unextractable: a case study on verb-particles," in *COLING-02: proceedings of the 6th conference on Natural language learning*. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 98–104.
[5] A. Kilgarriff, "Googleology is bad science," *Computational Linguistics*, vol. 33, no. 1, pp. 147–151, 2007.
[6] N. L. Waters, "Why you can't cite wikipedia in my class," *Communications of the ACM*, vol. 50, no. 9, 2007.
[7] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[8] G. Chamielec and D. Weiss, "Modeling the frequency of phrasal verbs with search engines," Institute of Computing Science, Poznan University of Technology, Poland, Technical Report RA-05/08, 2008.