Survey and evaluation of Web search engine hit counts as research tools in computational linguistics

David Sánchez[a], Laura Martínez-Sanahuja[a], Montserrat Batet[b1]

[a]*Department of Computer Engineering and Mathematics,*
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia (Spain)

[b]*Internet Interdisciplinary Institute (IN3),*
Universitat Oberta de Catalunya,
Avda. Carl Friedrich Gauss, 5, 08860 Castelldefels, Barcelona, Catalonia (Spain)

Abstract

In recent years, many studies on computational linguistics have employed the Web as source for research. Specifically, the distribution of textual data in the Web is used to drive linguistic analyses in tasks such as information extraction, knowledge acquisition or natural language processing. For these purposes, commercial Web search engines are commonly used as the low-entry-cost way to access Web data and, more specifically, to estimate the distribution of the entity(ies) of interest from the *hit count* the search engines provide when querying such entities. Even though several studies have evaluated the effectiveness of Web search engines as information retrieval tools from the perspective of the end users, few authors have assessed the suitability of *hit counts* as research tools in computational linguistics; moreover, studies so far have focused on the most well-known search engines (typically Google, Bing and Yahoo!), and neglected potentially interesting alternatives that have recently surfaced. To fill this gap, in this work, we first compile and survey the general-purpose search engines that are currently available. Then, we evaluate the suitability of the hit counts they provide under several perspectives that are relevant for computational linguistics: flexibility of the query language, linguistic coherence, mathematical coherence and temporal consistency. The results of our survey show that, even though the choice of a particular search engine has been generally ignored by researchers relying on Web data, there are significant quality differences between the hit counts of current search engines, and that the most well-known and widely-used search engines do not provide the best results. In this respect, we also identify the search engines whose hit counts are best suited for research.

Keywords: Web search engines; hit counts; computational linguistics; information distribution; semantic similarity.

[1] Corresponding author: Montserrat Batet; E-mail: montserrat.batet@urv.cat

1. Introduction

The World Wide Web is the largest source of electronic textual data. This has motivated many researchers on computational linguistics to use the Web as source to assist corpus linguistics [1], information extraction [2], classification of textual messages [3] or semantic analyses [4]. Although web text is not fully representative of either spoken or written text, it has been often used as a convenient proxy for them in the belief that its biases and limitations are not large enough to invalidate the results. In this sense, one of the most common uses of Web's data is the assessment the social-scale distribution of information (which is crucial to understand the semantics and dependences underlying textual data) from the probability of (co-)occurrence of linguistic entities in the Web [5]. With this, researchers aim at alleviating the constraints imposed by the static linguistic corpora that has been commonly used in the past that, despite being reliable, are limited in terms of size, coverage and updates [6, 7].

The usual low-cost entry to point to Web data distribution is via commercial Web search engines (WSEs), specifically, from the *hit count* they provide in the result page. *Hit counts* provide estimations of the number of pages indexed by the search engine, after deleting duplicates and near duplicates from some or all matching pages [8]. Despite their limitations, hit counts have been extensively used to estimate the probabilities of the entities of interest. Seminal works using WSEs' hit counts in computational linguistics include the identification of translation for compositional phrases [9], the discovery of synonyms [10] or the assessment of frequencies of bigrams [11]. More recent works include building models of noun compound bracketing [12], large-scale information extraction [13, 14], ontology learning [15-18] or the estimation of the semantic similarity between textual entities [4, 6, 19]. Several application-oriented works also rely on hit counts to assist document annotation [20, 21], to profile users of social networks [22, 23] or to identify privacy risks in textual documents [5, 24]. Hit counts are also essential in other fields of research such as in Webometrics, which uses informetric techniques and hit counts to find, measure and characterize Web-based phenomena [25].

If hit counts are extensively used as proxies of the Web's information distribution in tasks such as the former ones, research outcomes would closely depend on the suitability and accuracy of such hits counts as estimators of the true frequencies of the data. Yet, many researchers ignore the choice of the search engine they use, thus potentially compromising their own research results. In fact, the search engines most commonly used in research are also those that dominate the search engine market, that is, Google, Bing and Yahoo! [4, 6, 24, 26]. This suggests that

researchers usually employ the WSE they are familiar with, which may not necessarily provide the most accurate hit counts.

1.1. Contributions and plan of this paper

The main goal of this work is to offer an up-to-date survey and evaluation of general-purpose WSEs (being commercial or not), which can be used to obtain the hit counts used in research tasks. In contrast to related studies (discussed in the next section), which only focus on the most well-known WSEs (Google, Bing, Yahoo!), our survey covers a much broader spectrum of search engines. With this, we aim to assess whether other less-known or more recent WSEs constitute similar or even better alternatives to the standard ones. For each WSE, we evaluate the suitability of the hit counts they provide under several perspectives that are relevant for computation linguistics. Our analysis is threefold; first, we survey the desirable search features that WSEs should implement in order to provide a flexible calculation of hit counts; second, we evaluate the linguistic coherence of hit counts in one of the core tasks of computational linguistics: the estimation of the semantic similarity between textual entities; then, we test the mathematical coherence of hit counts for queries with multiple terms involving logic operators (AND, OR, NOT), which are extensively used in computational linguistics; finally, we also evaluate the temporal consistency of hit counts over time in order to test their reliability as research tools. As a result of this survey *we identify the Web search engines whose hit counts are best suited for research*.

The rest of the paper is organized as follows. Section 2 depicts related studies on WSEs' hit counts and discusses their limitations, which we tackle in this paper. Section 3 lists the WSEs we gathered and surveys the desirable search features that WSEs should implement. Section 4 evaluates the linguistic coherence of the WSEs' hit counts when used to drive the calculation of the semantic similarity between terms; after that, it evaluates the mathematical coherence of hit counts in front of multiple term queries and their temporal consistency. Section 5 discusses the results and provides advice on choosing the WSEs best suited for research. The final section presents the main conclusions and depicts some lines of future research.

2. Related works

Many researchers have conducted studies on the information retrieval effectiveness of WSEs from the perspective of the end users [27-32]; that is, they assessed the appropriateness the ranked list of results provided by the WSEs with respect to the users' queries. Because these studies aimed at evaluating the end users' experience, they focused on the most popular and

widely-used search engines: Google, Bing (and former Microsoft search engines such as MSN Search and Live Search), Yahoo! Search and AltaVista. Most of them considered informational queries [33], that is, queries in which the user aims at finding documents on a specific topic, whereas only few surveys considered navigational queries, that is, queries in which the user aims at navigating to a known Web page. In general, Google and Yahoo! were considered the best engines with respect to their information retrieval effectiveness, followed by Bing (or its predecessor Live Search).

Few works have analyzed the accuracy and/or consistency of WSEs' hit counts, which is understandable because employing hit counts for research is a collateral use of WSEs. In early surveys on this topic [7, 11], Lapata and Keller investigated the performance of the use of hit counts (gathered from Altavista and Google) in a range of natural language processing tasks. Specifically, they compared hit count-based results with those obtained with a static corpus; they concluded that, despite its limitations, the former constitutes a competitive alternative to the latter, and that the size of the Web compensates the a priori advantages of supervised methods relying on corpora. At the same time, they also observed that the hit counts provided by Google varied substantially over time because of the modifications made to its index and data bases. Finally, they realized that queries involving Boolean queries often returned inconsistent results.

In [8, 34], M. Thelwall compared the hit counts provided by Google, Live Search and Yahoo! from the perspective of Webometrics. Specifically, he measured the divergence between the hit counts and the actual number of URLs resulting for a given query. At the time of the analysis (2008), Google provided the most consistent results. However, the author found that hit counts presented limitations when used as distributional metrics because i) web search engines do not index the whole Web, ii) hit counts are just estimations of the number of pages indexed by the search engine, and iii) these estimations may either consider or not the removal of duplicate or near-duplicate URLs. A similar analysis was conducted by A. Uyar [35]: hit counts for a set of queries with single and multiple terms were compared against the actual number of results. The author also found that all the search engines he considered provided estimates for the number of matching documents, and that estimation patterns greatly differed. The analysis was restricted to queries providing less than 1,000 results, which is the maximum number of results per query indexed by most WSEs. Because of this constraint, the analysis was limited to very specific queries and cannot be generalized to general queries; in any case, Google provided, again, more accurate results than Live Search and Yahoo!.

The more recent studies performed by Yamana et al. [36, 37] focused on analyzing the consistency of the hit counts over time. To do so, the authors analyzed the differences in the hit counts provided by Google, Yahoo! and Bing (formerly Live Search) for a set of queries in successive repetitions during the same day and within a two-month period. They observed significant differences for the surveyed search engines and temporal periods. However, the accuracy of hit counts, as estimators of the frequencies of the entities they refer to, was not evaluated.

In [38, 39], Tian et al. evaluated the logical coherence of hit counts when using positive (AND) and negative (NOT) operators for Google, Yahoo! and Bing. At the time of the study (2010), they found that, in many occasions, hit counts do not monotonically decrease when adding new search words; however, hit counts showed better monotonicity when adding positive words (AND) than with negative ones (NOT). More specifically, Google rarely behaved correctly when negative words were added to the search, whereas Yahoo! and Bing behaved much better.

In [40] we conducted a preliminary survey that identified and characterized a spectrum of WSEs much broader than the related works above. In the current study we update and extend this former survey in several ways. First, we update the list and characterization of WSEs (to September 2016), and provide up-to-date results on the accuracy of hit counts in a core task of computational linguistics: the estimation of the semantic similarity between textual terms. We aim not only to assess the potential performance of WSEs' hit counts (as done in related works), but also to measure their actual performance in a real and widespread research task. Moreover, with respect to [40], we add evaluations of the mathematical coherence of hit counts with multiple term queries and logic operators (AND, OR, NOT), which are of great interest in computational linguistics; we also add evaluations on the temporal consistency of hit counts in order to test their reliability. With these new and more comprehensive results and discussions, we derive more robust conclusions on the WSEs best suited for research.

3. Survey of Web search engines

With the aim of being general and domain-independent, our study focuses on WSEs supporting general-purpose searches and indexing cross-domain Web resources. Therefore, we omit search engines that are constrained to a certain domain of knowledge or that only index a specialized corpus. Some examples of domain-dependent WSEs are PubGene and GoPubMed, which are medical search engines on clinical literature, or BASE and Google Scholar, which index scientific documents.

Moreover, we set additional criteria that are needed for hit count-based analyses and that we use to sift the search engines that we consider in this study. First, the WSE should be active and online, and it should provide a standard search bar for introducing textual queries; in this respect, some search engines provide interactive search interfaces, but not textual inputs. Secondly, the WSE should obviously provide the hit count associated to the performed query, and this should be minimally consistent and representative; accordingly, we discard those search engines that, for a set of general queries (see details in Section 4), either do not provide hit counts or provide very variable results for the same query (see *Exalead*) or nearly zero values (see *Scour, Zuula*). Finally, in order to avoid regional or language biases affecting the calculation of hit counts, we require the search engine to be cross-language.

The search engines considered in this study were compiled in September 2016, by using Wikipedia articles and web surveys on search engines as sources. As a result of this process, 58 individual WSEs were considered and analyzed. However, in some cases, WSEs redirect or use the search engines of other vendors. In such cases, once we checked that the hit counts are equal as those of the main vendor, we limited our analysis to the latter. From the set of 58 WSEs, in Table 1 we list those that did not fulfil some of the criteria above.

Table 1. List of WSEs that did not fulfil some of the functional criteria needed for hit count-based analyses; these WSEs were not considered in our study.

Criteria		*Web search engine*
Hit counts related problems	No hit count for some queries	*Trovator*
	Extremely low hit counts	*Scour; Zuula*
	Very variable hit counts	*Exalead*
	Hit counts not provided	*Ask.com; Dogpile; DuckDuckGo; Excite; Gyffu; info.com; ixquick; Mamma; Qwant; WebCrawler; YaCy; HotBot; Lycos*
Redirection to another WSE	to Bing	*MSN*
	to Google	*iAlgae; Wopa!*
	to Yahoo!	*Alltheweb; Altavista*
Language	only Chinese	*Panguso; Sogou; Sohu; Soso.com; Youdao*
	only Chinese and Japanese	*Baidu*
	only Korean	*Naver*
	only French	*Dazoo FR; LeMoteur; Premsgo*
	only Italian	*Virgilio.it*
	only Swedish	*Swisscows*

Availability	No textual search bar	*Alexa Internet; Blekko; GrayMatter; joongle; Kosmix; Mahalo; Munax; Voila*
	No textual data, just photos	*Specify*
	No online version	*Volunia*
	Down for maintenance	*NowRelevant*
	Inactive	*Yauba; Neuralcoder*

In the following, we list and briefly describe the 11 WSEs that fulfilled all the criteria and were the subject of further analysis:

1. *AOL Search* (https://search.aol.com) was launched in 2005 by AOL Inc. Since 2016, its search is powered by Bing, even though their hit counts slightly differ for some queries.
2. *Bing* (http://www.bing.com) was launched in 2009 by Microsoft. The service has its origins in Microsoft's previous search engines: MSN Search, Windows Live Search and, later, Live Search. By September 2017 it is the second most used search engine with a 7.65% market share [41].
3. *Ecosia* (https://www.ecosia.org) was created in 2009 and it is based in Berlin, Germany. Ecosia's search results are powered by Bing, but the hit counts it provides slightly differ from those of Bing.
4. *Entireweb* (http://www.entireweb.com) is a search engine launched in 2000 by Entireweb Sweden AB.
5. *Gibiru* (http://gibiru.com) was launched in 2009 and provides uncensored and non-personalized anonymous search.
6. *Gigablast* (https://www.gigablast.com) is an independent open source web search engine based in New Mexico that was launched in 2000.
7. *Google Search* (https://www.google.com) is the most-used search engine in the World-Wide Web. As of September 2017 it has a 78.78% market share [41].
8. *Mojeek* (https://www.mojeek.com) was launched in 2009 in the UK, and provides unbiased non-censored search results with no user tracking.
9. *Mozbot* (https://www.mozbot.com) was previously called Reacteur.com and was launched in 2003. It is based in France and provides search results in partnership with Google, even though their hit counts differ significantly.
10. *Yahoo! Search* (https://search.yahoo.com) was created in 1995. Since 2009, its search results are powered by Bing, even though their hit counts differ. As of September 2017 it is the fourth most used search engine with a 4.7% market share [41].
11. *Yandex* (https://www.yandex.com) was launched in 2010 and it is based in Russia.

Because the main focus of WSEs is on the information retrieval process and the hit counts they provide have a merely informative purpose, there may be some limitations when the latter are used in research. In this respect, some authors have questioned the suitability of WSEs as research tools due to a number of issues, regardless the actual accuracy of the hit counts, which we will analyze later. Below, we discuss the main research-oriented problems that have been traditionally attributed to WSEs [42], which we later assess in current search engines.

1) For many years, WSEs have implemented strict keyword-based indexing algorithms in which the query string is matched in a literal way with the indexed contents. However, for computational linguistics, it is usually more desirable to estimate the distribution of concepts rather than of exact words; specifically, *several* lexicalizations are usually associated to the same concept of interest, e.g., *flu* and *influenza* refer to the same concept. In such case, the synonyms, abbreviations, acronyms and different morphological forms of the words referring to the concept of interest should be considered during the retrieval (and hit counting) process because all of them constitute concept occurrences. Notice that the ability of WSEs to lemmatize words and to consider the synonyms of search queries is also an interesting feature for regular Web users because it contributes to improve the information retrieval recall.
2) Another problem that researchers using WSEs usually face is the lack of flexibility of the query language. For regular Web users, a simple query language supporting single or multiple terms is enough. However, researchers usually require operators that allow them, for example, to define Boolean search expressions, to constrain the length of the co-occurrence context for queries with multiple terms (e.g., on document, sentence or n-gram basis) in order to minimize the ambiguity of the search query, or to define character wildcards and build regular expressions to look also for morphological derivations of a certain query.
3) Finally, at an operational level, research efforts relying on WSEs may be limited by the number of consecutive queries the WSE allows, which is an access restriction that is usually implemented to limit abuses or to prevent DoS attacks.

In this first part of our survey, we have evaluated whether the selected WSEs suffer or not from the issues detailed above. In particular, for each WSE listed above, we have analyzed the following aspects:

- *Non-literal searches*: we examined whether the WSEs automatically consider morphological derivations, that is, stemming and lemmatization, and/or synonyms, acronyms and abbreviations of the queries. The terms queried to test this feature were

extracted from the Rubenstein & Goodenough linguistic benchmark [43], which we also use in the qualitative evaluation reported in Section 4.

- *Flexibility of the query language*: we evaluated whether the considered WSEs support Boolean search operators (AND, OR, NOT, ()) for queries with multiple terms and whether they allow defining the length of the co-occurrence context for such multiple term queries by means of proximity search operators (e.g., NEAR). Moreover, we checked if the WSEs support character wildcards for individual terms or characters (e.g., *).
- *Access restrictions*: we checked if the WSEs provide an API to perform searches and retrieve the results and hit counts via programming code. If positive, we report on the maximum number of queries it offers and whether it is free or payment.

Table 2. Features offered by the considered WSEs (by October 2017): non-literal searches, flexibility of the query language and restrictions of access.

Feature/ WSE	*Non-literal search*	*Boolean operators*	*Proximity operators*	*Search wildcards*	*Search API*
AOL Search[2]	Yes	AND, OR, NOT	No	No	No
Bing[3]	Yes	AND, OR, NOT	near:#	No	Payment (1K queries/$3) [4]
Ecosia[5]	Yes	AND, OR, NOT	No	No	No
Entireweb[6]	Yes	AND, OR, NOT	No	No	Free (1K queries/day) [7]
Gibiru[8]	Yes	+, |, -	No	No	No
Gigablast[9]	Yes	AND, OR, NOT, ()	No	No	Payment (1K queries/$0.99) [10]
Google[11]	Yes	AND, OR, -	AROUND(#)	*, $, ..	Free (100 queries/day) [12] Payment (1K queries/$5)
Mojeek[13]	Yes	+, -	No	No	Payment (1K queries/day) [14]
Mozbot[15]	Yes	AND, OR, NOT	No	No	No

[2] https://help.aol.com/articles/aol-search-faqs
[3] https://msdn.microsoft.com/en-us/library/ff795620.aspx
[4] https://azure.microsoft.com/en-us/pricing/details/cognitive-services/search-api/web/
[5] https://ecosia.zendesk.com/hc/en-us/categories/200752332-Search-Engine
[6] http://www.entireweb.com/about/showcase/web/
[7] http://www.entireweb.com/search_api/
[8] http://www.gibiru.com/
[9] https://www.gigablast.com/syntax.html
[10] https://www.gigablast.com/searchfeed.html
[11] https://support.google.com/websearch/answer/2466433?hl=en&ref_topic=3081620
[12] https://developers.google.com/custom-search/json-api/v1/overview
[13] https://www.mojeek.com/advanced.html
[14] https://www.mojeek.com/services/api.html

Yahoo! Search[16]	Yes	AND, OR, NOT	No	No	Free (with ads) [17]
Yandex[18]	Yes	+, |, -	No	*	Free (10K queries/day) [19]

From Table 2, we can see that, nowadays, all search engines implement stemming and lemmatization algorithms in order to detect morphological variations of the queried terms and/or incorporate lists of synonyms for some words. Thanks to these functionalities, the search is performed more at a conceptual level than at a lexical level, which contributes to improve the information retrieval recall and alleviates the limitations of strict keyword-based search engines. The drawback is that most WSEs do not allow disabling the search for synonyms, even for queries with double quotes. Notice also, that WSEs lack the ability to disambiguate polysemous queries and to specify the desired sense of the query, which are also of great interest for linguistic analysis.

Regarding the query language, we can see that WSEs offering a powerful and flexible language are still a minority; only Google provides a proximity search operator (t_1 *AROUND(number of words separating* t_1 *and* t_2*)* t_2) and a variety of search wildcards (any word (*), numbers ($) and number ranges (..)). In general, the most recent search engines (e.g., Ecosia, Gibiru, Mojeek) offer none or a very limited search syntax, which is understandable, because a greater search flexibility requires larger and more sophisticated search indexes and indexing algorithms. Fortunately, most WSEs support the three basic Boolean operators (AND/+, OR/|, NOT/-); only Mojeek does not support OR. Gigablast even supports the definition of complex logical expressions with parenthesis.

Regarding access limitations, the most widely-used WSEs offer either payment service (e.g., Bing) or very limited free queries (e.g., Google). Only Yahoo! offers free unlimited searches, but the results include ads, which may cause some bias. Yandex and Entireweb offer more balanced alternatives, with 10,000 and 1,000 free queries per day and IP, respectively. More recent and minority WSEs (Ecosia, Gibiru and Mozbot) do not provide search API.

[15] https://www.mozbot.com/syntaxe-en.html
[16] https://help.yahoo.com/kb/search-for-desktop/advanced-web-search-sln2194.html
[17] https://developer.yahoo.com/ypa/docs/faq/
[18] https://yandex.com/support/search/how-to-search/search-operators.html
[19] https://tech.yandex.com/xml/doc/dg/concepts/restrictions-docpage/

4. Evaluation of hit counts

Although the former analysis provides some insights on the possibilities of WSEs as tools for research, it gives no evidences on the reliability of WSEs' hit counts as estimators of the Web's information distribution.

Former studies have shown that there may be significant discrepancies among the hit counts provided by different WSEs and, also, inconsistences and variability within the hit counts of a specific WSE. Unfortunately, the hit count estimation algorithms implemented by the WSEs are not public and the causes for such inaccuracies can only be guessed from empirical observations and experiments. Even though ascertaining the causes of hit count inconsistencies is outside the scope of our study, in the following, we summarize the reasons enounced in previous works that may explain hit count inaccuracies:

- Very early surveys on WSEs [44, 45] acknowledged the fluctuations of the results provided by WSEs to the same query. They interpreted that fluctuations could be caused by errors in the indexing or retrieval processes or because documents may be removed of the indexed database (e.g., censorship, removal of duplicates). Another cause is the use of several databases and indexes, which are regularly replicated; then, the same query can be done to different databases, or indexes could be temporally unreachable, thus producing different results. Moreover, some WSEs use performance-dependent algorithms, which means that less accurate results may be provided under high loads.
- M. Thelwall [34] identified the significantly different criteria, filters, and thresholds used by the WSEs to crawl, analyse and index web sites as the main reason behind the discrepancies of the hit counts they provide. Moreover, in [8] he also found that Microsoft Live Search used different algorithms to calculate hit counts according to the number of results of a given query.
- According to A. Uyar [35], because search engines use distributed computers to perform searches, data sets are divided into subgroups that are processed by separated nodes, which are later aggregated in another node [46]. Sometimes (e.g., under heavy loads) a search may not be performed in all subgroups, thus resulting in incomplete and variable results. Also, some search indexes (such as news) may be updated very frequently, and may not be reflected in global indexes. Also, WSEs use multiple levels of caches, some of which may be outdated and produce incomplete results [45]. Finally, queries with multiple terms require combining several search results, which multiply and propagate individual inaccuracies, as it has been also observed by E. Davis [47].
- Satoh and Yamana [36] analyzed the consistency of hit counts through time. They discovered that the distributed and dynamic search indexes used by WSEs cause several

hit count inaccuracies. Specifically, because WSEs crawl and update small parts of their indexes at short periods of time (i.e., seconds) whereas other are updated over days or weeks [48], a high variability of hit counts can be observed during update operations. Differences between cached results and the full search index [49] produce significantly different hit counts depending on the accessed cache. Bar-Ilan et al. [50] also identified that the lack of synchronization between search indexes, especially during index update operations, may cause significant variability in the results of the same query. A. Kilgarriff [42] already hypothesized that similar causes were behind the observed hit count inaccuracies.

In this section, we evaluate the actual accuracy of hit counts by means of a quantitative assessment under the perspectives of *linguistic coherence*, *mathematical coherence* and *temporal consistency*. As far as we know, ours is the only survey that assesses the mathematical coherence of hit counts, and that considers the three former perspectives at once and for such variety of WSEs.

4.1. Linguistic coherence

To assess the *linguistic coherence* of hit counts, that is, how well they estimate the linguistic distribution of data at a social scale [6], we employ an application-oriented evaluation in which hit counts are used as input in a core task of computational linguistics: the estimation of the semantic similarity (or distance, which is the inverse to similarity) between linguistic entities. Semantic similarity/distance quantifies how much the meaning of two entities (i.e., terms or concepts) are alike. This resemblance can be measured by following different paradigms and by exploiting different information or knowledge sources (e.g., raw textual corpora, ontologies, thesauri, etc.) [51, 52]. The notion of semantic similarity/distance is applied in many tasks dealing with textual data, such as the classification and clustering of documents [53], semantic disambiguation [54] or privacy protection [55]. In these tasks, the semantic distance between entities is used as the linguistic equivalence of the arithmetic distance between numerical values; in consequence, the accuracy of the semantic distance calculation is crucial for the performance of the tasks relying on it.

Of the different paradigms employed to measure the semantic similarity/distance between words, in this work we adopt the perspective of *distributional semantics*, which bases the calculation on the distributional characteristics of words in large samples of linguistic data. As mentioned in the introduction, the Web has been widely used as corpus of textual data for distributional measures. In this scenario, WSEs are used as proxies to obtain the distributional

characteristics of the linguistic entities in the Web. The core idea of distributional semantics can be summarized in the so-called *distributional hypothesis* [56]: linguistic entities with similar distributions have similar meanings; that is, words that co-occur in a context tend to be semantically related or similar. This notion can be operationalized in practice by using the WSE's hit count of single term queries to estimate the frequency of appearance of textual terms, and the hit count of the concatenation of pairs of terms (with the AND operator) to assess their frequency of co-occurrence. Then, by dividing both values by the total number of web resources indexed by the WSE, we can calculate both the marginal and joint probabilities of a pair of linguistic entities and, from these, we can compute their semantic resemblance by using a similarity or distance coefficient.

Below, we detail some of the coefficients that have been used to measure the semantic similarity of linguistic terms from the hit counts provided by WSEs. These are the ones we use as means to evaluate the linguistic coherence of hit counts:

- *Pointwise mutual information* (*PMI*): This measures the discrepancy between the probability of co-occurrence of two events ($p(a,b)$) and the probability of occurrence of the individual events ($p(a)$ and $p(b)$) under the assumption of independence between them. In a seminal work by Turney [10] the former probabilities were approximated to the WSEs' hit counts as follows :

$$PMI(a,b) = \log_{10}\frac{p(a,b)}{p(a)\times p(b)} \approx \log_{10}\frac{\left(\frac{hits("a"\ AND\ "b")}{total\_webs}\right)}{\frac{hits("a")}{total\_webs}\times\frac{hits("b")}{total\_webs}}, \quad (1)$$

  where *hits*("*a*"), *hits*("*b*") and *hits*("*a AND b*") are the *hit counts* corresponding to *a*, *b,* and *a* and *b* together, respectively, and *total_webs* is the estimated total number of web pages indexed by the WSE. The PMI can take values in the range (-∞ … *min*(-log $p(a)$,-log $p(b)$)].

- *Normalized PMI* (*NPMI*): This was defined in [57] as a normalized version of the PMI, which is unbounded; it provides values in the [-1,+1] range:

$$NPMI(a,b) = \frac{PMI(a,b)}{-\log_{10}(p(a,b))} \approx \frac{\log_{10}\frac{\left(\frac{hits("a"\ AND\ "b")}{total\_webs}\right)}{\frac{hits("a")}{total\_webs}\times\frac{hits("b")}{total\_webs}}}{-\log_{10}\left(\frac{hits("a"\ AND\ "b")}{total\_webs}\right)} \quad (2)$$

- *Normalized Google Distance* (*NGD*): This is an ad-hoc measure designed to estimate the semantic distance of pairs of linguistic terms by making use of the hit count provided by Google [6]:

$$NGD(a,b) = \frac{\max\left(\log_{10}\left(hits("a")\right), \log_{10}\left(hits("b")\right)\right) - \log_{10}\left(hits("a"\ AND\ "b")\right)}{\log_{10}\left(total\_webs\right) - \min\left(\log_{10}\left(hits("a")\right), \log_{10}\left(hits("b")\right)\right)} \quad (3)$$

Notice the use of the double quotes ("") operator in the queries employed to obtain the hit counts, so that the order and adjacency of words in queries with multiple terms (phrases) are maintained (e.g., "*lung cancer*").

We have chosen these measures because they solely rely on the WSE's hit counts to estimate the semantic similarity or distance. The literature on semantic similarity offers a variety of more sophisticated distributional measures [58] that, by relying on some degree of supervision and/or tuning parameters [59], or by employing more complex queries aimed at minimizing language ambiguity [10, 19] -which are not supported by all the WSEs- are able to provide more accurate results. We left these measures outside our study, so that we avoid a number of variables that may influence the results: tuning parameters, degree of supervision, supported/unsupported query operators, etc. In this way, we ensure that the results *only* depend on the performance of the hit counts that the WSEs provide, which is the aspect that we want to evaluate.

To measure the accuracy of semantic similarity/distance calculations and, thus, the linguistic coherence of the WSE's hit counts in which they rely on, we compare the results provided by the measures (1)-(3) with the similarity ratings provided by human experts on a benchmark of term pairs. Specifically, by measuring the correlation between computerized measures and human ratings we assess how well the former mimic human judgements semantics, which is the ultimate goal behind the design of such measures. We use the Rubenstein & Goodenough [43] benchmark, which is the most well-known and widely used semantic similarity evaluation benchmark [52]. It consists of 65 pairs of general English nouns with associated average similarity ratings (in a 0-4 scale) provided by 51 human subjects.

We use the *Pearson correlation* (*r*) to compare similarity ratings and evaluate the linguistic coherence of hit counts; *r* quantifies the linear statistical dependence between two variables that, in our case, correspond to the sets of computerized and human similarity ratings of the term pairs in the benchmark. The correlation ranges [+1…−1], where 1 means that the ratings are totally dependent; that is, that the computerized measures perfectly mimic human judgements and, thus, that the hit counts optimally captured of the information distribution at a social scale;

0 indicates independence and -1 indicates inverse dependence. In addition to absolute correlation values, we also provide 95% confidence intervals (CI) of *r* so that we can evaluate whether the differences between search engines are statistically significant.

The correlation between successive ratings of the term pairs in the Rubenstein & Goodenough benchmark by the 51 human subjects involved in the construction of the benchmark was r=0.85. Because this value quantifies the discrepancy between human ratings, it defines an upper correlation bound for computerized assessments of semantic similarity.

In our experiments, hit counts have been retrieved by simulating query requests from different Web browsing sessions by means of the Selenium Python library (http://docs.seleniumhq.org/). In this way, we avoid the access limitations imposed by some WSEs and make the hit count retrieval uniform for all WSEs, whether they provide an API or not, and also because some WSEs return different results when using the API and the web interface, due to the latter having more priority and accessing more up-to-date caches [35]. When possible, we set the search engine language to English. Also, because most WSEs do not detail the size of the Web corpus they index, we estimated the *total_webs* constant referred in equations (1)-(3) by following a fixed and general procedure similar to that detailed in [60]. Specifically, we used the known frequency of a commonly-used word in corpora (i.e. 'a' appear in around 67% of the documents [60]), to extrapolate *total_webs* from the hit counts each WSE reports for such word; that is, *total_webs=hits*(*"a"*)/0.67. Even though the estimated *total_webs* may affect the results provided by equations (1)-(3), our evaluation metric (r) considers the dependency among the values, rather than individual absolute numbers. Also, in this experiment, queries involving term pairs (a,b) have been made in the same order as the terms appear in the benchmark (i.e., *hits*("*a*" AND "*b*")). Queries for all search engines were submitted during the second week of September 2016.

Table 3 reports the Pearson correlation (r) and 95% confidence intervals of the Pearson correlation for the search engines and measures we consider. Notice that, because the NGD quantifies distances rather than similarities, which results in negative correlations, we report -r instead of r.

The results show that the three measures provide similar correlation values and, therefore, define similar relative rankings with respect to the WSEs' hit counts linguistic coherence. In all cases, Mojeek and Google achieve the highest accuracies; these are followed by Mozbot, AOL Search, Ecosia, Gibiru and Yahoo!, with small differences among them and slight variations in

the WSEs' rankings across the measures. The remaining search engines provide too low correlation values to be reliably used in research. In this respect, it is especially surprising to observe the nearly random results provided by Bing (formerly Live Search) and the negative correlations of Yandex, which suggest that hit count estimation in these WSEs has been poorly implemented.

Table 3. Pearson correlation (*r*) and 95% CI intervals between the hit count-based PMI, NPMI and NGD measures and the similarity ratings of the human experts in the Rubenstein & Goodenough benchmark for each WSE. Queries were submitted during the second week of September 2016.

WSE	total_webs	*PMI*		*NPMI*		*NGD*	
	(millions)	*r*	*95% CI*	*r*	*95% CI*	*-r*	*95% CI*
AOL Search	26,700	0.384	0.155; 0.574	0.378	0.148; 0.569	0.374	0.144; 0.566
Bing	26,700	0.079	-0.168; 0.316	0.096	-0.151; 0.332	0.064	-0.182; 0.303
Ecosia	26,700	0.383	0.154; 0.573	0.377	0.147; 0.568	0.374	0.144; 0.566
Entireweb	27,000	0.166	-0.081; 0.393	0.293	0.053; 0.501	0.172	-0.075; 0.399
Gibiru	4,515	0.366	0.135; 0.559	0.446	0.227; 0.622	0.367	0.136; 0.56
Gigablast	1,261	0.158	-0.089; 0.386	0.163	-0.084; 0.391	0.169	-0.078; 0.396
Google	37,905	0.466	0.251; 0.637	0.495	0.286; 0.659	0.467	0.252; 0.638
Mojeek	1.7	0.538	0.339; 0.691	0.610	0.431; 0.743	0.540	0.342; 0.692
Mozbot	3,435	0.387	0.159; 0.576	0.397	0.17; 0.584	0.383	0.154; 0.573
Yahoo! Search	26,560	0.382	0.153; 0.572	0.338	0.103; 0.537	0.383	0.154; 0.573
Yandex	843	-0.202	-0.424; 0.044	-0.180	-0.406; 0.066	-0.208	-0.43; 0.037

4.2 Mathematical coherence

The use of hit counts as estimators of the individual and joint distributions of terms may be hampered if the hit counts lack *mathematical* coherence with respect to the syntax of the queries. Specifically, hit counts of queries with multiple terms involving logical operators (i.e., AND, OR, NOT) should be:

(i) Independent to the order of the terms; that is, AND and OR should be symmetric with respect to hit counts, e.g., *hits*("*a*" AND "*b*")=*hits*("*b*" AND "*a*").

(ii) Coherent with the logic operators used to link them, for example, *hits*("*a*" OR "*b*")=*hits*("*a*")+*hits*("*b*")-*hits*("*a*" AND "*b*").

In practice, these strict equalities rarely hold due to the technical issues we discussed at the beginning of the section; specifically, the fact that WSEs may access to different caches, even

for consecutive queries [45], because some WSEs may use different hit count estimation algorithms based on the number of results [8], or because queries with multiple terms require aggregating several search results, which cause an accumulation of estimation errors [35]. In this section, we evaluate whether these issues affect the *mathematical coherence* of hit counts.

In the first experiment, we tested the symmetric property of the AND operator with respect to the hit counts by querying the term pairs in the Rubenstein & Goodenough benchmark with different orders. Specifically, because the order of the term pairs should not influence the similarity results (i.e., similarity measures are *symmetric*, that is, $sim(a,b)=sim(b,a)$), we compared the correlation of the different measures and WSEs with respect to the human ratings in the benchmark when varying that order of terms of each pair. Results are reported in Table 4; notice that the values for the (*a* AND *b*) order are those reported in Table 3, which correspond to the original order of the term pairs in the benchmark.

Table 4. Pearson correlation (*r*) for the Rubenstein & Goodenough benchmark when varying the order of the terms of each pair. Queries were submitted during the second week of September 2016.

WSE		*PMI*		*NPMI*		*NGD*	
		r	*Difference*	*r*	*Difference*	*-r*	*Difference*
AOL Search	*a AND b*	0.384	0.029	0.378	0.021	0.374	0.027
	b AND a	0.355		0.357		0.347	
Bing	*a AND b*	0.079	0.113	0.096	0.119	0.064	0.108
	b AND a	0.192		0.215		0.172	
Ecosia	*a AND b*	0.383	0.031	0.377	0.022	0.374	0.03
	b AND a	0.352		0.355		0.344	
Entireweb	*a AND b*	0.166	0.131	0.293	0.038	0.172	0.124
	b AND a	0.297		0.331		0.296	
Gibiru	*a AND b*	0.366	0.018	0.446	0.021	0.367	0.019
	b AND a	0.348		0.425		0.348	
Gigablast	*a AND b*	0.158	0.124	0.163	0.126	0.169	0.129
	b AND a	0.034		0.037		0.040	
Google	*a AND b*	0.466	0.1	0.495	0.102	0.467	0.121
	b AND a	0.366		0.393		0.346	
Mojeek	*a AND b*	0.538	0	0.610	0	0.540	0
	b AND a	0.538		0.610		0.540	
Mozbot	*a AND b*	0.387	0.024	0.397	0.004	0.383	0.017
	b AND a	0.363		0.393		0.366	
Yahoo! Search	*a AND b*	0.382	0.087	0.338	0.054	0.383	0.099
	b AND a	0.295		0.284		0.284	

Yandex	*a AND b*	-0.202	0.003	-0.180	0.004	-0.208	0.004
	b AND a	-0.199		-0.184		-0.204	

Bing, Entireweb and Gigablast perform particularly poorly, with divergences between the correlations for the different orders that are near 100% for some measures. These divergences are extremely high, and suggest a lack of coherence in the assessment of hit counts by the WSEs. As a matter of fact, these WSEs also provide very low *r*, which indicates that their hit counts are too random to be usable for research beyond the anecdotal. Google and Yahoo! provide better but still significantly divergent results, with correlation differences in the 25% range, whereas AOL Search, Ecosia, Gibiru, Mozbot and Yandex achieve reasonably similar correlations, which diverge within the 1-10% range. Finally, Mojeek is the only WSE that provides *exact* results regardless of the ordering of multiple term queries; in fact, hit counts are *identical* for the different orderings for all the term pairs. This is of great interest for algorithms requiring perfectly coherent frequencies, such as clustering or classification of textual resources.

In a second experiment, we also tested the mathematical coherence of the hit counts with respect to the logic operators (AND, OR and NOT) supported by the WSEs for multiple term queries. Specifically, we tested whether the following logical equalities hold for the hit counts of each WSE:

$$hits\left("a"\,OR\,"b"\right) = hits("a") + hits("b") - hits\left("a"\,AND\,"b"\right) \quad (4)$$

$$hits\left("b"\,OR\,"a"\right) = hits("b") + hits("a") - hits\left("b"\,AND\,"a"\right) \quad (5)$$

$$hits\left("a"\,NOT\,"b"\right) = hits("a") - hits\left("a"\,AND\,"b"\right) \quad (6)$$

$$hits\left("b"\,NOT\,"a"\right) = hits("b") - hits\left("b"\,AND\,"a"\right) \quad (7)$$

For each term pair $t=(a,b)$ in the benchmark, we measured the divergence observed between the expressions ($expr_i$) at each side of the equalities in equations (4)-(7) (e.g.,$expr_1$= *hits*("*a*" OR "*b*"), $expr_2$=*hits*("*a*")+*hits*("*b*")-*hits*("*a*" AND "*b*")) according to the *absolute error*:

$$error\left(\mathrm{expr}_1(t), \mathrm{expr}_2(t)\right) = \left|\mathrm{expr}_1(t) - \mathrm{expr}_2(t)\right| \quad (8)$$

In order to make the error between term pairs comparable regardless their absolute hit counting, we normalized the absolute error by the largest magnitude. The resulting *relative error* is expressed as a percentage of divergence between the two equivalent expressions, $expr_1$ and $expr_2$, for each term pair *t*, as follows:

$$rel\_error\left(\text{expr}_1(t), \text{expr}_2(t)\right) = \frac{error\left(\text{expr}_1(t), \text{expr}_2(t)\right)}{\max\left(\text{expr}_1(t), \text{expr}_2(t)\right)} \times 100 \quad (9)$$

Finally, to obtain a normalized and aggregated value of divergence for each WSE and expression, we measured the arithmetic *mean of the relative error* for all the pairs in the benchmark as follows:

$$\overline{rel\_error\left(\text{expr}_1, \text{expr}_2\right)} = \frac{1}{|bench|} \sum_{\forall t \in bench} rel\_error\left(\text{expr}_1(t), \text{expr}_2(t)\right), \quad (10)$$

where $|bench|$ is the number of term pairs in the benchmark. Average errors close to 100% suggest very large divergences, whereas values close to 0% indicate very accurate results.

The mean relative errors (and standard deviations) between the expressions in the logical equalities depicted above (eq. (4)-(7)) are reported in Table 5 for all WSEs.

Table 5. Mean relative error (and standard deviation σ) between the hit count-based expressions in the logical equalities (4)-(7) for each WSE. Queries were submitted during the second week of September 2016.

WSE	*expr₁=hits*(“*a” OR “b”*) *expr₂=hits*(“*a”*) + *hits*(“*b”*) –*hits*(“*a” AND “b”*)	*expr₁=hits*(*“b” OR “a”*) *expr₂=hits*(*“b”*) +*hits* (“a*”*) –*hits*(*“b” AND “a”*)	*expr₁=hits*(*“a” NOT “b”*) *expr₂=hits*(*“a”*) – *hits*(*“a” AND “b”*)	*expr₁=hits*(*“b” NOT “a”*) *expr₂=hits*(*“b”*) – *hits*(*“b” AND “a”*)
AOL Search	36.76% (σ=20.52%)	39.25% (σ=23.97%)	37.62% (σ=24.38%)	43.58% (σ=29.55%)
Bing	36.40% (σ=21.79%)	36.32% (σ=22.20%)	35.49% (σ=29.50%)	43.10% (σ=28.36%)
Ecosia	36.99% (σ=20.68%)	38.82% (σ=24.12%)	37.65% (σ=23.99%)	43.16% (σ=30.85%)
Entireweb	35.03% (σ=19.53%)	44.47% (σ=26.89%)	41.12% (σ=27.87%)	45.80% (σ=33.31%)
Gibiru	20.19% (σ=14.62%)	20.18% (σ=14.54%)	22.25% (σ=18.04%)	30.71% (σ=22.58%)
Gigablast	83.35% (σ=13.51%)	83.45% (σ=13.51%)	17.49% (σ=18.13%)	23.00% (σ=21.26%)
Google	17.25% (σ=20.09%)	14.16% (σ=12.47%)	19.17% (σ=22.67%)	25.32% (σ=21.32%)
Mojeek	*OR not supported*	*OR not supported*	5.00% (σ=8.89%)	4.63% (σ=5.70%)
Mozbot	22.58% (σ=17.13%)	23.32% (σ=18.43%)	18.57% (σ=22.02%)	28.42% (σ=25.47%)
Yahoo! Search	36.61% (σ=19.83%)	46.88% (σ=26.21%)	38.41% (σ=24.07%)	38.22% (σ=30.10%)
Yandex	9.30% (σ=9.03%)	10.27% (σ=9.19%)	55.95% (σ=21.50%)	60.21% (σ=20.03%)

The figures show that, in general, the WSEs that provided the most accurate results in the former experiments are also the ones that exhibit the lowest errors. However, due the relatively large standard deviations of the errors among the word pairs in the benchmark, we should interpret the results with caution. Interestingly, the WSEs that produced the largest errors (i.e.,

Gigablast for the AND/OR operators and Yandex for the NOT operator) tended to exhibit the lowest relative variances, which indicate that their results are "consistently poor". The differences and variances of the errors when varying the order to the queries (i.e., expression (4) vs (5) and expression (6) vs. (7)) are also consistent with those observed in the symmetric property test, with Yandex and Mojeek providing the smallest differences.

Regarding the coherence of the OR/AND operators (eq. (4)-(5), second and third columns in Table 5), the results show that only Google and Yandex are able to maintain the mean relative error below 20%, with Mozbot and Gibiru being slightly above 20%. Standard deviations are similar for the three search engines, with a coefficient of variation (i.e., standard deviation/mean) near 1, in most cases. Unfortunately, Mojeek does not support the OR operator; therefore, the only option to quantify the *union* of hits between pairs of terms is to use the expression *hits*(*a*)+*hits*(*b*)-*hits*(*a AND b*). The remaining WSEs produce too large mean errors (above 35%), with Gigablast providing nearly random results (mean error above 80%).

The results involving the NOT operator (eq. (6)-(7), fourth and fifth columns in Table 5) are also reasonably accurate for Google, Gibiru and Mozbot, even though they are now surpassed by Gigablast and Mojeek. The fact that Gigablast is significantly more accurate for the NOT operators than for the OR/AND operators discussed above suggests a poor implementation of the OR operator for this search engine. The opposite can be said for Yandex, which performs significantly poorer for the NOT operator than for the OR/AND operators. Mojeek is especially interesting, since it achieves the lowest mean errors and deviations (around 5-8%), which are also quite similar when varying the order of the terms (recall that Mojeek was the only WSE providing exact results in the symmetric property test). The remaining WSEs exhibit a similar (poor) behavior as with the AND/OR operator.

4.3 Temporal consistency

Another aspect that is of great interest when using hit counts for research is their consistency over time, that is, their *temporal consistency*. It is normal that hit counts increase over time as more web pages are created and get indexed by the search engine. Ideally, for "static" general concepts, the variation of hit counts should be even and proportional to the amount of indexed content. However, as discussed above, the differences we observe when querying the same term at different times are often a consequence of the update operations carried out by WSEs on their distributed search indexes [36, 37], and by the lack of synchronization among indexes [50]. This hampers the reproducibility of hit count-based research results.

In this section, we aim at evaluating whether these technical issues significantly hamper the temporal consistency (and, thus, the reliability) of hit counts for the different WSEs. To do so, we compare the results reported in Table 3 (obtained in the second week of September 2016) with those obtained in our former study for the same WSEs [40] (obtained in the fourth week of May 2016). Because the concepts in the Rubenstein and Goodenough benchmark are general, their semantic similarity and, thus, the correlation results should be maintained; thus, ideally, there should not be significant differences between their relative distributions in such a short period of time. In fact, because the absolute number of hits tends to increase over time, comparing the relative distributions of terms (as we do) is preferable to comparing absolute hit counts, as done by related studies [36, 37]; in the latter case, the observed differences would be less conclusive because one would not be able to discern if they are the result of a technical issue or the consequence of the natural increase of indexation coverage. Table 6 reports the correlation values and differences of the two repetitions of the experiment.

Table 6. Pearson correlation (*r*) for two repetitions of the experiment reported in Table 3.

WSE		*PMI*		*NPMI*		*NGD*	
		r	*Difference*	*r*	*Difference*	*-r*	*Difference*
AOL Search	*Sept. 2016*	0.384	0.04	0.378	0.006	0.374	0.029
	May 2016	0.344		0.372		0.345	
Bing	*Sept. 2016*	0.079	0.015	0.096	0.001	0.064	0.003
	May 2016	0.064		0.097		0.067	
Ecosia	*Sept. 2016*	0.383	0.281	0.377	0.261	0.374	0.289
	May 2016	0.103		0.116		0.085	
Entireweb	*Sept. 2016*	0.166	0.085	0.293	0.025	0.172	0.08
	May 2016	0.251		0.318		0.252	
Gibiru	*Sept. 2016*	0.366	0.187	0.446	0.159	0.367	0.188
	May 2016	0.553		0.605		0.555	
Gigablast	*Sept. 2016*	0.158	0.06	0.163	0.091	0.169	0.077
	May 2016	0.098		0.072		0.092	
Google	*Sept. 2016*	0.466	0.022	0.495	0.028	0.467	0.011
	May 2016	0.444		0.467		0.456	
Mojeek	*Sept. 2016*	0.538	0.053	0.610	0.05	0.540	0.072
	May 2016	0.591		0.660		0.612	
Mozbot	*Sept. 2016*	0.387	0.157	0.397	0.128	0.383	0.119
	May 2016	0.554		0.525		0.502	
Yahoo! Search	*Sept. 2016*	0.382	0.193	0.338	0.089	0.383	0.196
	May 2016	0.189		0.249		0.187	

Yandex	*Sept. 2016*	-0.202	0.066	-0.180	0.068	-0.208	0.066
	May 2016	-0.268		-0.248		-0.274	

We see that AOL Search, Google and Mojeek provide correlations that are very stable (10% or lower divergence) over the three-and-a-half-month period. Strictly speaking, Bing also maintains the stable correlation, but the results were and are so close to 0, that they should be considered random. The correlation values provided by Ecosia, Entireweb Gigablast, Yahoo! and Yandex are too low and variable in either or both tests to draw reliable conclusions; whereas Gibiru, Mojeek and Mozbot show too much divergence (e.g. more than 50% for Gibiru and the PMI) to consider their results consistent over time.

5. Discussion

Before going into further discussions on the results reported above, we should acknowledge the limitations of our study:

- Because search engine algorithms change frequently, we cannot know how long the conclusions of our study would be valid [34]. However, our evaluation framework can be executed periodically to reassess and update the findings.
- Our analysis is restricted to English web sites and English queries. Because some WSEs implement language-dependent analysis (e.g., to consider synonyms and lexicalizations), their behavior and capabilities may be different for other more minority languages.
- We evaluated hit counts in semantic similarity, which is a core task of computational linguistic of which many other tasks, such as semantic clustering or classification, depend on. However, this is not fully representative of the interest of WSEs in other linguistic research tasks. In any case, as we argue later, the choice of a WSE may vary according to the particularities of the task to be conducted (e.g., requirements on the query syntax, volume of queries to be submitted, etc.), regardless the accuracy in the task we considered.
- Our empirical study did not use search wildcards and POS searches because, as depicted in Table 2, only a small subset of WSEs support them, which would make our analysis very narrow.

Several findings can be extracted from the tests conducted in Sections 3 and 4. First, from the evaluation of the linguistic coherence of hit counts (Table 3), it is surprising to see the poor results achieved by Bing, which provided almost zero correlation for most measures. This fact contrasts with the effectiveness that this search engine achieved in former studies focusing on

its information retrieval abilities [29, 31]. As noticed in [8], Bing (formerly Live Search) was used to implement different hit count estimation algorithms according to the number of results of the query. Moreover, given other WSEs powered by Bing's results (AOL Search, Ecosia and Yahoo) obtained similar and significantly higher correlations (0.38 vs. 0.08), we hypothesize that either Bing was subjected to updates during our tests (whereas the other WSEs used older and more stable caches), or Bing implements performance-based algorithms that significant degrade the hit counts when a batch of queries are executed from the same origin.

Yahoo! Search, AOL Search and Ecosia, all of them powered by Bing, provided similar results in most tests; even though, they showed significantly different temporal variabilities, which would likely be caused by the access to different caches. Their mathematical consistency was remarkably poor (as it was for Bing), which suggests a poor aggregation of hit counts for queries with multiple terms.

Among the three most used search engines (Google, Bing and Yahoo!), only Google obtained reasonably high and stable correlations over time (Table 6). Moreover, it was one of the two WSEs below 20% divergence for the mathematical coherence test with the AND/OR operators (eqs. (4) and (5)). In this case, the figures we obtained are consistent with the retrieval effectiveness assessed in former studies [27, 29, 30], which concluded that Google provided the best results. However, Google produced divergent results during the evaluation of the symmetric property of the AND operator (Table 4). From a qualitative point of view, Google also offers the most flexible query language (Table 2), supporting proximity operators and a variety of search wildcards. However, it is also one of the most restrictive WSEs in terms of search API (only 100 free queries per day and IP).

Among the less widespread search engines considered in our survey, Mojeek constitutes the most interesting alternative to the mainstream WSEs. First, it obtained the highest correlations for all the measures, which were also higher than those of Google (e.g., Mojeek's 0.610 vs. Google's 0.495 for the NPMI, Table 3). In fact, considering the human correlation upper-bound for the benchmark we use (0.85 intra-subject correlation) and the bare bone nature of the measures we employ, Mojeek's hit counts can be considered reasonably good estimators of the true frequencies of terms. Moreover, Mojeek was also the only WSE whose hit counts *exactly* fulfilled the symmetric property for the AND operator (Yandex was very close, but its linguistic coherence was very poor). This property will be of great interest for researchers employing co-occurrence frequencies of multiple terms, because it ensures that the results are independent of the order of the terms. Finally, Mojeek was also the WSE that provided the smallest errors in the

mathematical coherence test with the AND/NOT operators. The good results achieved by Mojeek could be explained by the fact that it is not driven by commercial interests and, on the contrary to most WSEs, it provides uncensored and unbiased searches. This lack of bias is of great importance when analyzing the Web's information distribution, and would contribute to make the hit count more reliable and representative of the true distribution of information in society. Also, due to the narrow indexing of Mojeek (less than 2 million webs, see Table 3), it is less affected by the hit count inconsistencies caused by the highly distributed implementation of large WSEs (see Section 4). The main drawbacks of Mojeek are i) the fact that its query language is restricted to the very basic Boolean operators (in fact the OR operator is not supported), which may be of limited usefulness in complex linguistic analyses and ii) its payment search API.

In a second group, Mozbot and Gibiru also provided usable results (Table 3) and reasonably fulfilled the symmetric property (Table 4). They also showed a similarly reasonable mathematical coherence and temporal variability, albeit the results were worse in the more recent test. Like Mojeek, Gibiru offers unbiased and uncensored searches, which will be of special interest for queries that might be object of such biases. The remaining WSEs (Entireweb Gigablast and Yandex) provided too low linguistic and mathematical coherence to be used in research beyond the anecdotal.

6. Conclusions and future work

Many researchers have used the Web as linguistic data source and, more specifically, the hit counts provided by WSEs as proxies of the distribution of information at a social scale. Even though researchers have usually ignored the choice of the particular WSE, in this study, we have shown that there are very significant differences among WSEs, and that the most well-known and widely-used ones are not always the best-suited for research.

In comparison with related studies [35-37, 40], we provide an up-to-date and much broader survey of WSEs, and a comprehensive and integral threefold evaluation of the "quality" of hit counts under the perspectives that are relevant in computational linguistics: linguistic coherence (in a core task of computational linguistics), mathematical coherence (for multiple term queries) and temporal consistency.

Whereas Bing provided particularly poor results, Google was the most balanced option, with reasonably high accuracy and the most powerful query language. Moreover, we also identified

other less known –and, so far, unstudied– WSEs (i.e., Mojeek in a first tier and Mozbot and Gibiru in a second tier) that achieved results even better than Google in some aspects, specifically, on the mathematical and linguistic coherence of hit counts. Researchers may carefully consider which of these aspects are the most relevant for their application and, thus, take an informed decision on the WSE to use according to the results we report.

As future work, we plan to extend our analysis to concrete domains, in order to test whether domain-specific WSEs (e.g., medical ones) are able to improve general-purpose ones when dealing with domain-specific terms. To do so, we will base the evaluation on domain-specific benchmarks, such as the Pedersen's benchmark of biomedical terms [61]. Another interesting experiment would be comparing the results provided by general-purpose WSEs with those obtained from large corpora created for research on computational linguistics [42], and checking whether the latter are a preferable alternative (with respect to accuracy, coverage and updates) to the former. In this sense, it is important to note that even specialized corpora are not free of flaws [62].

Disclaimer and acknowledgements

This work was partly supported by the European Commission (projects H2020-644024 "CLARUS" and H2020-700540 "CANVAS"), by the Spanish Government (projects TIN2014-57364-C2-R "SmartGlacis", TIN2015-70054-REDC "Red de excelencia Consolider ARES" and TIN2016-80250-R "Sec-MCloud"). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of UNESCO.

References

[1] P. Resnik, N. Smith, The Web as a parallel corpus, Computational Linguistics 29 (2003) 349-380.
[2] T. Özacar, A tool for producing structured interoperable data from product features on the web, Information Systems 56 (2016) 36–54.
[3] S.R. Yerva, Z. Miklós, K. Aberer, Quality-aware similarity assessment for entity matching in Web data, Information Systems 37 (2012) 336–351.
[4] D. Bollegala, Y. Matsuo, M. Ishizuka, Measuring Semantic Similarity between Words Using Web Search Engines, in: Proc. of 16th international conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007, pp. 757-766.

[5] R. Chow, P. Golle, J. Staddon, Detecting Privacy Leacks Using Corpus-based Association Rules, in: Proc. of 14th Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, 2008, pp. 893-901.
[6] R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering 19 (2006) 370-383.
[7] M. Lapata, F. Keller, Web-based models for natural language processing, ACM Transactions on Speech and Language Processing 2 (2005) 1-31.
[8] M. Thelwall, Extracting accurate and complete results from search engines: Case study windows live, Journal of the American Society for Information Science and Technology 59 (2007) 38-50.
[9] G. Grefenstette, The WWW as a resource for example-based MT tasks, in: Proc. of ASLIB Conference on Translating and the Computer 21, London, U.K., 1999, pp.
[10] P.D. Turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, in: Proc. of 12th European Conference on Machine Learning, ECML 2001, Freiburg, Germany, 2001, pp. 491-502.
[11] F. Keller, M. Lapata, Using the web to obtain frequencies for unseen bigrams, Computational Linguistics 29 (2003) 459-484.
[12] P. Nakov, M. Hearst, Search engine statistics beyond the n-gram: Application to noun compound bracketing, in: Proc. of Ninth Conference on Computational Natural Language Learning, Ann Arbor, Michigan, US, 2005, pp. 17-24.
[13] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, A. Yates, Unsupervised named-entity extraction form the Web: An experimental study, Artificial Intelligence 165 (2005) 91-134.
[14] D. Sánchez, D. Isern, Automatic extraction of acronym definitions from the Web, Applied Intelligence 34 (2011) 311-327.
[15] D. Sánchez, A. Moreno, Learning non-taxonomic relationships from web documents for domain ontology construction, Data & Knowledge Engineering 63 (2008) 600-623.
[16] D. Sánchez, A. Moreno, Pattern-based automatic taxonomy learning from the Web, AI Communications 21 (2008) 27-48.
[17] D. Sánchez, A methodology to learn ontological attributes from the Web, Data & Knowledge Engineering 69 (2010) 573-597.
[18] D. Sánchez, A. Moreno, L.D. Vasto-Terrientes, Learning relation axioms from text: An automatic Web-based approach, Expert Systems with Applications 39 (2012) 5792-5805.
[19] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems 35 (2010) 383-413.

[20] P. Cimiano, S. Handschuh, S. Staab, Towards the self-annotating web, in: Proc. of 13th international conference on World Wide Web, WWW 2004, New York, USA, 2004, pp. 462 - 471.
[21] D. Sánchez, D. Isern, M. Millán, Content annotation for the Semantic Web: an automatic Web-based approach, Knowledge and Information Systems 27 (2011) 393-418.
[22] A. Viejo, D. Sánchez, J. Castellà-Roca, Preventing automatic user profiling in Web 2.0 applications, Knowledge-Based Systems 36 (2012) 191-205.
[23] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, M. Ishizuka, Polyphonet: An advanced social network extraction system from the Web, Web Semantics: Science, Services and Agents on the World Wide Web 5 (2007) 262-278.
[24] D. Sánchez, M. Batet, C-sanitized: A privacy model for document redaction and sanitization, Journal of the Association for Information Science and Technology 67 (2016) 148-163.
[25] L. Björneborn, P. Ingwersen, Toward a basic framework for Webometrics, Journal of the American Society for Information Science and Technology 55 (2004) 1216-1227.
[26] O. Chapelle, Y. Chang, Yahoo! Learning to Rank Challange Overview, in: Proc. of Yahoo! Learning to Rank Challenge at ICML 2010, Haifa, Israel, 2011, pp. 1-24.
[27] D. Lewandowski, Evaluating the retrieval effectiveness of Web search engines using a representative query sample, Journal of the Association for Information Science and Technology 66 (2015) 1763-1775.
[28] A. Macfarlane, Evaluation of web search for the information practitioner, Aslib Proceedings: New Information Perspectives 59 (2007) 352-366.
[29] S.K. Deka, N. Lahkar, Performance evaluation and comparison of the five most used search engines in retrieving web resources, Online Information Review 34 (2010) 757-771.
[30] D. Bilal, Ranking, relevance judgment, and precision of information retrieval on children's queries: Evaluation of Google, Yahoo!, Bing, Yahoo! Kids, and ask Kids, Journal of the American Society for Information Science and Technology 63 (2012) 1879-1896.
[31] J. Zhang, W. Fei, Search engines? responses to several search feature selections, The International Information & Library Review 42 (2010) 212-225.
[32] J. Bar-Ilan, M. Levene, A Method to Assess Search Engine Results, Online Information Review 35 (2011) 854-868.
[33] A. Broder, A taxonomy of web search., ACM Sigir forum 36 (2002) 3-10.
[34] M. Thelwall, Quantitative comparisons of search engine results, Journal of the American Society for Information Science and Technology 59 (2008) 1702-1710.
[35] A. Uyar, Investigation of the accuracy of search engine hit counts, Journal of Information Science 35 (2009) 469-480.

[36] K. Satoh, H. Yamana, Hit Count Reliability: How Much Can We Trust Hit Counts?, in: Proc. of 14th Asia-Pacific international conference on Web Technologies and Applications, 2012, pp. 751-758.
[37] T. Funahashi, H. Yamana, Reliability Verification of Search Engines' Hit Counts: How to Select a Reliable Hit Count for a Query, in: Current Trends in Web Engineering, Springer, 2010, pp. 114-125.
[38] T. Tian, J. Geller, S.A. Chun, Predicting Web Search Hit Counts, in: Proc. of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, ON, Canada, 2010, pp. 162-166.
[39] T. Tian, S.A. Chun, J. Geller, A prediction model for web search hit counts using word frequencies, Journal of Information Science 37 (2011) 462-475.
[40] L. Martínez-Sanahuja, D. Sánchez, Evaluating the Suitability of Web Search Engines as Proxies for Knowledge Discovery from the Web, in: Proc. of 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, York, UK, 2016, pp. 169-178.
[41] Netmarketshare. Desktop Search Engine Market Share. March 2017.Available at https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0.
[42] A. Kilgarriff, Googleology is Bad Science, Computational Linguistics 33 (2007) 147-151.
[43] H. Rubenstein, J. Goodenough, Contextual correlates of synonymy, Communications of the ACM 8 (1965) 627-633.
[44] H. Snyder, H. Rosenbaum, Can search engines be used as tools for web-link analysis? A critical view, Journal of Documentation 55 (1999) 375-384.
[45] W. Mettrop, P. Nieuwenhuysen, Internet search engines - Fluctuations in document accessibility, Journal of Documentation 57 (2001) 623-651.
[46] J. Zhang, X. Long, T. Suel, Performance of compressed inverted list caching in search engines, in: Proc. of 17th international conference on World Wide Web Beijing, China, 2008, pp. 387-396.
[47] E. Davis. A difference of a factor of 70,000 between hit counts and results returned in Google. In: Unpublished technical note; 2015.
[48] J. Dean. Challenges in Building Large-Scale Information Retrieval Systems. In: 2009 Conference on Web Search and Data Mining; 2009.
[49] G. Skobeltsyn, F. Junqueira, V. Plachouras, R. Baeza-Yates, ResIn: a combination of results caching and index pruning for high-performance web search engines, in: Proc. of 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore, 2008, pp. 131-138.

[50] J. Bar-Ilan, M. Levene, M. Mat-Hassan, Methods for evaluating dynamic changes in search engine rankings: a case study, Journal of Documentation 62 (2006) 708-729.
[51] M. Batet, D. Sánchez, Review on Semantic Similarity, in: Encyclopedia of Information Science and Technology (3rd edition), IGI Global, 2014, pp. 7575-7583.
[52] Juan J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, F. Chirigati, HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset, Information Systems 66 (2017) 97–118.
[53] M. Batet, Ontology based semantic clustering, AI Communications 24 (2011) 291-292.
[54] B.T. McInnes, T. Pedersen, Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text, Journal of Biomedical Informatics 46 (2013) 1116-1124.
[55] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of Biomedical Informatics 46 (2013) 294-303.
[56] M. Sahlgren, The Distributional Hypothesis, Rivista di Linguistica 20 (2008) 33-53.
[57] G. Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction, in: Proc. of Biennial GSCL Conference 2009, Tübingen, Germany, 2009, pp. 31–40.
[58] S. Mohammad, G. Hirst. Distributional Measures of Semantic Distance: A Survey. In: http://arxiv.org/abs/1203.1858; 2006.
[59] D. Bollegala, Y. Matsuo, M. Ishizuka, A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web, in: Proc. of Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, Singapore, Republic of Singapore, 2009, pp. 803–812.
[60] A. van den Bosch, B. T., d.K. M., Estimating search engine index size variability: a 9-year longitudinal study, Scientometrics 107 (2016) 839-856.
[61] T. Pedersen, S. Pakhomov, S. Patwardhan, C. Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics 40 (2007) 288-299.
[62] S. Zhang. The pitfalls of using Google Ngram to study language. In: Wired; 2015.