Analyzing the Sense Distribution of Concordances Obtained by Web As Corpus Approach

Abstract. In corpus-based lexicography and natural language processing fields some authors have proposed using the Internet as a source of corpora for obtaining concordances of words. Most techniques implemented with this method are based on information retrieval-oriented web searchers. However, rankings of concordances obtained by these search engines are not built according to linguistic criteria but to topic similarity or navigational oriented criteria, such as page-rank. It follows that examples or concordances could not be linguistically representative, and so, linguistic knowledge mined by these methods might not be very useful. This work analyzes the linguistic representativeness of concordances obtained by different relevance criteria based web search engines (web, blog and news search engines). The analysis consists of comparing web concordances and *SemCor* (the reference) with regard to the distribution of word senses. Results showed that sense distributions in concordances obtained by web search engines are, in general, quite different from those obtained from the reference corpus. Among the search engines, those that were found to be the most similar to the reference were the informational oriented engines (news and blog search engines).

1 Introduction

Most statistical approaches to solving tasks related to Natural Language Processing (NLP) as well as lexicographic works use corpora as a resource of evidences. However, one of the biggest problems encountered by these approaches is to obtain an amount of data that could be large enough for statistical and linguistic analysis. Taking into account the rapid growth of the Internet and the quantity of texts included in it, some researchers have proposed using the Web as a source for building corpora [7]. Two strategies have been proposed for exploiting the web with that objective in mind:

- *Web As Corpus*: The web is accessed directly as a corpus. This access is usually performed by means of commercial web search engines (*Bing*, *Yahoo*, *Google*...), which are used to retrieve concordance lines showing the context in which the user's search term occurs. *WebCorp* [11] is a representative linguistic tool based on this strategy.
- *Web For Corpus*: This strategy consists of compiling a corpus from the web to be accessed later. This compilation process can be performed by crawling the web and also by using commercial web search engines. This latter approach consists of sending a set of queries including seed terms corresponding to a certain topic or domain, and then retrieving the pages returned by the engine. *BootCat* is a tool based on this approach [3].

adfa, p. 1, 2011.

The two strategies usually rely on web search engines (SEs) in order to retrieve pages with text and in this way build word concordances or text corpora. Using APIs provided by SEs offers several advantages. It makes the treatment of spam and other low-quality, undesired contents easier. Besides, these APIs provide a high coverage of the web.

The Web as Corpus approach is more suitable than Web for Corpus for those tasks requiring an acceptable quantity of examples of concordances for any word (e.g. Distributional Similarity, Information Extraction, Word Sense Disambiguation, etc...). However, some problems can arise from using SEs for concordance compilation. For example, Aguirre et al. [1] found that the great number of erotic web pages strongly influenced their experiments on WSD. The set of pages retrieved by the web SE is dependent on ranking criteria1, which are not specified according to linguistic features such as frequency of use of each sense. The users of commercial SEs have other needs than those focused on obtaining specific pieces of text information. Broder [5] states that the target of search queries is often non-informational (more than 50%), since it might be navigational when the queries are looking for websites, or transactional when the objective is to shop, download a file, or find a map. Thus, criteria related to these needs are also reflected in the above mentioned ranking factors. The main page-ranking factors mainly rely on popularity, anchor text analysis and trusted domains, but not on content.

Our work is based on two main assumptions:

- *Assumption 1*. *SemCor* is a sense-tagged corpus [10], which can be regarded as a gold-standard reference of the "real" distribution of word senses in an open domain. According to Agirre and Martinez [2], corpora with similar distribution to that of *SemCor* get the best results in the task of WSD for an open domain. Word Sense Disambiguation (WSD) systems with the best performance in Senseval-2 were trained with it.
- *Assumption 2*. The Web is, in terms of variety of genres, topics and lexicons, close to the most traditional open domain "balanced" corpora, such as the Brown or BNC [3].

On the basis of these two assumptions, we aim to validate the following hypothesis: the ranked results obtained by SEs are not a representative sample in terms of sense distribution, since they follow ranking criteria focused on non-linguistic relevance and only first results are usually compiled. In other words, the linguistic examples or concordances extracted by web SEs are biased by non-linguistic criteria. A high bias would indicate that web-based concordances compilation is not useful at all in some NLP and lexicography tasks. For example, linguistic information obtained by an SE used for knowledge extraction such as cross-lingual distributional similarity [12], semantic-class learning [8], or consultation (e.g. lexicographic, translators, writers or language learners) might not be reliable. In the remaining sections, we will attempt to confirm or reject such a hypothesis.

---

[1] An example of factors used for search engine rankings are listed here: http://www.seomoz.org/article/search-ranking-factors.

2 Related Work

There are few works which deal with linguistic adequacy of concordances obtained by SEs. Chen et al. [6] describe an experiment to manually evaluate concordances included in web documents retrieved from an SE for 10 test words. They annotated by hand about 1,000 to 3,000 instances for each test word. In particular, the authors evaluate two pieces of information: the quality of the web documents returned by the SE, and sense distributions. Concerning sense distribution, they concluded that, on the one hand, the most frequent senses from web-based corpora are similar to *SemCor* and, on the other, web-based corpora may provide more diverse senses than *SemCor*. However they do not perform any correlation analysis to draw that conclusion. As we will show later in Section 5, a correlation analysis performed over their results shows a low correlation in terms of sense distribution between web-based corpus and *SemCor*.

Other works analyze some aspects related to linguistic adequacy but for different purposes. Diversity in web-rankings is a topic closely related to word-sense distribution analysis. [13] propose a method to promote diversity of web search results for small (one-word) queries. Their experiments showed that, on average, 63% of the pages in search results belong to the most frequent sense of the query word. This suggests that "diversity does not play a major role in the current *Google* ranking algorithm". As we will show in Section 5, the degree of diversity of the concordances we have retrieved is still lower: in our experiments 72% of the concordances belong to the most frequent sense.

3 Methodology for analyzing the adequacy of *Web As Corpus*

Our objective is to verify whether the distribution of senses in the rankings obtained from SEs are representative in linguistic terms for an open domain. Our analysis relies on *SemCor* as evidence, since it is a well-known, manually annotated open domain reference corpus for word senses according to *Wordnet*. In addition, we also measure two further properties of the concordances retrieved by SEs, namely sense diversity and linguistic coherence (i.e., typos, spelling errors, etc...).

3.1 How to obtain concordances of a word from the web?

Several SEs have been used in the literature in order to collect examples of concordances from the web. Most authors use SEs directly by collecting the retrieved snippets. Web As Corpus tools such as *WebCorp* are more linguistically-motivated tools. In that sense they offer parameters to post-process SE rankings (case sensitive searches to avoid named entities, no searches over links to avoid a somehow navigational bias...). Anyway they still depend on SE rankings. So they raise the same problems mentioned in Section 1. Other emergent SEs are those focused on specific domains and genres such as news or blogs. These SEs are interesting for linguistic purposes because bias produced by factors related to navigational and transactional is avoided. In addition, their text sources are not domain restricted. In fact, newswire-based cor-

pora are often built for open domain knowledge extraction purposes. However, some authors [3] point out that, in terms of variety of genres and topics, Web is closer to traditional "balanced" corpora such as the BNC. Blog is a new genre not present in traditional written sources but similar to them in terms of distribution of topics. In order to analyze and compare the influence of these characteristics, the following engines have been evaluated: *WebCorp*, *Google News Archive* (*GNews*), and *Google Blog Search* (*GBlog*). See Table 1 for more details.

Table 1. Characteristics of SEs

SE	Domain	Genre	Query	Ranking
WebCorp	Open	Open	inform.	topic
			navig.	popularity
			transac.	… (See first note)
GBlog	Open	Blogs	inform.	topic
GNews	Open	News	inform.	topic

In order to guarantee a minimum linguistic cohesion of the concordances, the following parametrizations were used for each SE. English is selected as the target language in all of them. In *WebCorp*2, Bing has been selected as the API because provides the best coverage. Case-sensitive searches were performed. The search over links option was disabled in order to mitigate navigational bias. Span of ± 5 words for concordances was established. *GNews* and *GBlog* do not offer choice for case-sensitive searches. So, case-sensitive treatment was done after retrieving the snippets. In the cases of *GNews* and *GBlog*, searches were performed only on the body of documents and not on the titles (*allintext* operator was used).

3.2 Selecting test words

10 test words (see Table 2.) are randomly selected from the *SemCor* 1.6 corpus, a corpus where all words are tagged with their corresponding sense according to *WordNet* 1.6. Due to the small size of the sample several conditions were established in order to guarantee the representativeness of the test words and the corresponding contexts:

- Nouns are selected because they are the largest grammatical group.
- More than 1 sense in *SemCor* because we want to focus on ambiguous words.
- Minimum frequency of 50 on *SemCor* corpus. As McCarthy [9] pointed out, *SemCor* comprises a relatively small sample of words. Consequently, there are words where the first sense in *WordNet* is counter-intuitive. For example, the first sense of "*tiger*" according to *SemCor* is an audacious person, whereas one might expect carnivorous animal to be a more common usage.

[2] *WebCorp* has included recently *Gnews* and *Gblog* APIs.

Table 2. Selected test words from *SemCor* and their sense distribution

Word	Sense distribution
church	1=0.47,2=0.45,3=0.08
particle	1=0.63,2=0.35,3=0.02
procedure	1=0.73,2=0.27
relationship	1=0.60,2=0.21,3=0.19
element	1=0.71,2=0.21,3=0.06,4=0.02
function	1=0.58,2=0.32,3=0.09
trial	1=0.45,2=0.03,4=0.52
production	1=0.64,2=0.21,3=0.11,4=0.04
newspaper	1=0.66,3=0.02,2=0.29,2;1=0.02
energy	1=0.74,2=0.10,3=0.12,4=0.05

3.3 Annotation of web based concordances

Each test word is submitted to the three different SEs. The number of retrieved snippets, i.e., word concordances, may change depending on both the query word and the SE. So, in order to obtain more comparable samples, the first 250 concordances are retrieved for each case. As the number of test examples is still too much to analyze by hand, to save work without missing the rank information, only an interpolated sample of 50 concordances was analyzed. The hand analysis involves manually tagging the sense of the test words according to *WordNet* 1.6.

3.4 Measuring the adequacy of concordances

The main objective here is to measure differences in terms of sense distribution between *SemCor* and the concordances retrieved by the SE. However, besides sense distribution, our aim is also to measure both sense diversity and linguistic coherence. Let us describe first how we measure sense diversity, then linguistic coherence, and finally sense distribution.

3.4.1 Sense diversity

We associate the term "*sense diversity*" with text corpora whose word occurrences cover a great variety of senses. It is possible to know to a certain extent the degree of diversity of a corpus by observing the senses of a sample of words. In particular, diversity can be measured by comparing the number of possible senses of the test words (e.g., their *WordNet* senses) with the number of different senses that are actually found in the corpus, i.e., in our collections of concordances. The higher the number of senses covered by the concordances, the greater their degree of diversity. Concordances with much sense diversity tend to be open to many domains.

3.4.2 Linguistic coherence

The quality and linguistic coherence of the retrieved concordances can vary from totally nonsensical expressions to high quality texts. So, coherence, or more precisely "level of coherence", is also taken into account in our evaluation protocol. To do this, the annotators can assign four possible coherence values to each retrieved concordance:

- Score 0. The concordance has no problems.
- Score 1. The concordance has serious typographical errors or morphosyntactic problems, but it can be understood.
- Score 2. The query word is part of a Named Entity, e.g., "*town*" in "*Luton Town FC Club*".
- Score 3. The concordance is totally nonsensical and cannot be understood at all.

The range of values is from 0 (coherent) to 3 (totally incoherent or nonsensical). It should be borne in mind that values 1 and 2 could be unified since named entities written in lower-case seem to be typographical errors. However, we preferred to keep the two coherence levels because value 1 still allows us to assign a *WordNet* sense to the key word, but it is not the case when the coherence level is 2. On the basis of the notion of level of coherence, we define "degree of incoherence", which is associated with a concordance collection. The degree of incoherence of a concordance collection, noted ϕ, is computed as follows:

$$\phi = \left(\sum_{i}^{n} L(c_i) \right) / 3\mathrm{n} \qquad (1)$$

where $L(c_i)$ stands for the level of coherence of concordance c_i, and n is the number of concordances in the collection. Let us suppose that we have a collection of 4 concordances, with the following levels of coherence for each concordance: 0, 1, 0, 3. The degree of incoherence of the total collection is then 4/12 = 0.33. The values of this function are ranged from 0 (fully coherent) to 1 (totally incoherent).

3.4.3 Sense distribution

The distributions of senses found in the three SE concordance collections are compared with those extracted from *SemCor* by analyzing the Pearson correlation between them. The correlation between two sets of concordances is computed by considering the relative frequencies of those senses of the word that have been found in, at least, one of the two sets.

Besides the strict correlation, we are also interested in verifying properties concerning sense dominance. For instance, two collections may share (or not share) the same dominant sense. When their dominant senses are not the same, and one of them is domain-specific (scientific, technical, etc.), then the two collections should be considered very different in terms of sense distribution, regardless of their specific Pearson correlation. On the other hand, when they share the same dominant sense, it is

also important to observe whether there are differences in terms of degree of dominance. If the main sense is very dominant in one collection and not so dominant in the other one, we may infer that there are significant differences in sense distribution. This is true even if the Pearson correlation is actually very high. Let us see an example. Word "production" have 4 senses with the following two sense distributions, in *SemCor* and *GNews*, respectively:

- *SemCor*: 0.64 0.21 0.11 0.04
- *GNews*: 0.98 0.02 0.0 0.0

The Pearson correlation between *SemCor* and *GoogleNews* is very high: > 0.97. However, from a linguistic perspective, the distributions are very different. While the sense distribution in *SemCor* may be considered as an evidence for content heterogeneity (there are three senses with more than 10% occurrences), sense distribution in *GoogleBlogs* shows that concordances are content homogeneous. As in *GoogleBlogs* only one sense covers more than 98% of the word occurrences, it means that concordances are retrieved from a domain-specific source. By contrast, concordances of *SemCor* seem to represent more open and balanced text domains.

The rank of retrieved concordances for each SE is also analyzed. We are interested in observing the order of appearance of the senses among the web ranking. An adequate order will be that one in which concordances are ordered according to the probability of senses of the search word. Thus, concordances including the most probable ones should be on the top of the rank. For linguistic consultations, for instance, it is better to show concordances including the most common senses of the search word at the top of the ranking. In addition, those strategies that only retrieve the first concordances of the SE could also perform better if top concordances corresponded to the most common senses. Once again, we use *SemCor* to prepare a reference rank according to sense probabilities calculated from *SemCor*. In order to measure the adequacy of the rank of web-concordances, we compute the Spearman correlation between the web concordances rank and the *SemCor* based reference rank.

4 Results

The results concerning sense diversity, linguistic coherence, and sense distribution are shown and analyzed as follows:

4.1 Sense Diversity

In total, the 10 test words have 49 different *WordNet* senses. Among these 49 senses, the collection of concordances from *WebCorp* contains instances of 34 senses, two more (32) than the senses found in *SemCor* for the same 10 words. The concordances of *GBlog* contain a different 31 senses while those of *GNews* only 27. In Table 3, we show the percentage of different senses we found in each corpus with regard to the total number of *Wordnet* senses attributed to the 10 test words (first column), as well as to the number of senses these words have in *SemCor* (second column). In the last

column, we show the number of senses appearing in each collection of concordances that do not appear in *SemCor*. It follows that the *WebCorp* corpus may provide more diverse senses and, therefore, more domain diversity, than the two corpora built from the Google engines. In addition, we may also infer that the journalism articles seem to be more restricted in terms of domain diversity than the posts of blogs. However, we have to take into account that high diversity does not imply balanced sense distribution.

Table 3. Sense diversity

	% senses of test words	% senses in *SemCor*	#new-senses
WebCorp	69%	81%	8
GBlog	63%	78%	6
GNews	55%	72%	5

4.2 Linguistic coherence

Table 4 shows information on levels of incoherence associated with the three web-based concordance collections. The three first columns show the total values of 3 levels of coherence (level 0 is not shown but can be inferred). The forth column measures the degree of incoherence for each collection, according to formula 1 (see above). In the last column, we show the percentage of concordances having some positive incoherence value (i.e., 1, 2, or 3) for each collection.

We can observe that *WebCorp* is the SE that provides more incoherent concordances at the 3 levels. In addition, in *WebCorp* almost 1 context out of 4 has some problems of coherence. This is probably due to the fact that *WebCorp* covers the whole Web, containing many not very confident text sources. The degree of incoherence in *GBlog* is also relatively high (0.12), against only 0.04 of *GNews*, which is then the most reliable source of textual data in our experiments So, the linguistic quality of the corpora built with Google engines is clearly better than that of *WebCorp*.

Table 4. Sense diversity

	level 1	level 2	level 3	ϕ	incoherence (%)
WebCorp	68	6	48	0.15	24%
GBlog	34	1	25	0.07	12%
GNews	32	0	9	0.04	8%

4.3 Sense distribution

4.3.1 Pearson Correlation

The senses found in the web-based concordances are compared with those extracted from *SemCor* by analyzing the Pearson correlation between them (see Table 5). This table is organized as follows. The test word is in the first column. The following columns show the Pearson correlation between the *SemCor* and sense distributions corresponding to the different web-based concordances (*WebCorp*, *GNews*, or *GBlog*).

As far as the Pearson coefficient is concerned, the average correlation of WebCorp and *SemCor* is 0.51, which is the lowest correlation. The average correlation between *GBlog* and *SemCor* is 0.56, and the one between *GNews* and *SemCor* is the highest: 0.66. As the correlation values between 0.51 and 0.79 are interpreted as being "low", we may consider that there is always a low correlation between *SemCor* and our three web-based concordance collections. If we conduct a more detailed analysis word by word and compute the correlations of each test word, we can observe that there are three words ("*newspaper*", "*production*", and "*procedure*") with moderate correlations (between 0.80 and 0.86), four words ("*energy*", "*church*", "*trial*", and "*element*") with low correlations (between 0.51 and 0.79), and three ("*particle*", "*function*", and "*relationship*") without any correlation at all, since their values are lower than the significance level (0.35, $p<.01$) established for tests with 50 pairs.

Table 5. Pearson correlation of sense distribution regarding to *SemCor*

	WebCorp	*GBlog*	*GNews*
church	0.78	0.85	0.70
particle	0.10	0.53	0.20
procedure	0.62	0.88	0.89
relationship	-0.28	-0.41	0.50
element	0.28	0.29	0.95
function	0.50	-0.03	0.03
trial	0.60	0.63	0.7
production	0.61	0.89	0.97
newspaper	0.93	0.99	0.66
energy	0.97	0.97	0.98
average	0.51	0.56	0.66

It should be noticed that these results seem to be not in accordance with those experiments reported in [6], where the web-based corpus is "quite similar" to *SemCor* in terms of sense distribution. In that work, the notion of "quite similar" must be considered as a non-technical and naive intuition, since no correlation measure was computed. Considering the sense frequencies reported in that work, we computed the Pearson correlation between *SemCor* and their concordances for the 9 ambiguous test words

used in their experiments. Table 6 shows that the average correlation is low (0.59), very close to our results (total average of the three collections: 0.58).

Table 6. Pearson correlation between concordances reported by Chen et al. [6] and *SemCor*

	Chen et al.[6] concordances
author	1
back	0.94
cart	-1
case	0.85
center	0.15
core	0.35
mind	1
sequence	1
toast	0.99
average	0.59

4.3.2 Analysis on Dominant Senses

Besides the statistical test, it is also important to verify further qualitative aspects, in particular whether concordances are comparable in terms of sense dominance. More precisely, given two collections of concordances, we checked both whether they share the same dominant senses and whether the same dominant senses have a similar degree of dominance.

We observed that *SemCor* and each web-based concordances do not share the same dominant sense in several cases. In addition, for most of these words, their dominant senses in the web-based concordances are domain-specific senses, e.g. physics for "*particle*", computer science for "*function*", show business for "*production*", or gossip news for "*relationship*". It should be noticed that these specific senses are not dominant in *SemCor* and they do not correspond to the first sense in *WordNet*. The high relevance given by non-linguistic ranking criteria to webs dealing with scientific, business or gossip topics could explain the large number of domain-specific senses in the concordances.

On the other hand, for many cases where the test word shares the same dominant sense in both *SemCor* and the web-based concordances, we observe that there are significant differences in terms of degree of dominance. In general, the sense distribution of *SemCor* seems to be more balanced than that of web-based concordances. Six words had same dominant sense in both *SemCor* and *GNews*, but in all cases the shared sense is clearly more dominant in *GNews*. The average degree of dominance in *SemCor* is 62% against 72% in the web-based concordances. As we showed above in 4.4.1, these differences may not be very significant for the Pearson correlation, but from a linguistic point of view, they are very significant since they denote that the web-based concordances are more homogeneous in terms of linguistic content. Once

again, the ranking criteria of the SE could give more relevance to a very restricted subset of topics among all of those we can find in the open domain web.

In addition, we can find further qualitative differences between *SemCor* and web-based concordances. On the one hand, it should be noted that web-based concordances introduce new technical or domain-specific senses that are not in *SemCor*. On the other hand, we can find cases where the transactional function of some webs may influence sense distribution. For instance, the second sense of "*trial*" (very marginal in *SemCor*) is very important in *WebCorp* and *GBlog* because of the high number of commercial pages with "free trial" software.

4.3.3 Spearman Correlation

Finally, the order of the senses appearing in the web concordances ranking is also analyzed. We check whether web-concordances of test words are sorted by the frequency of use of the included sense. For this purpose, the reference ranking for each test word and SE is prepared by sorting all the collected concordances according to sense probabilities mined from *SemCor*. For example, suppose we collect the following concordances ranking from the web for a test word including 4 senses: {($context_1$,sen_1), ($context_2$,sen_2), ($context_3$,sen_1), ($context_4$,sen_1)}. The *SemCor*-based ranking (the reference) is built by sorting all concordances according to sense probability estimated from *SemCor* (sen_2=0.8,sen_1=0.2): {($context_2$,sen_2) ,($context_1$,sen_1), ($context_3$,sen_1), ($context_4$,sen_1)}. Contexts with the same sense keep the original order. Then, the Spearman correlation between original concordances ranking and *SemCor*-based reference ranking is calculated (see Table 7).

Table 7. Spearman correlation of concordance ranking with respect to SemCor

	WebCorp		*GBlog*		*GNews*	
	All	top50	All	top50	All	top50
church	0.41	0.98	0.58	0.68	0.43	-0.25
particle	0.63	1	0.57	1	0.58	0.65
procedure	-0.59	-0.76	0.46	1	0.11	1
relationship	0.61	0.53	0.49	-0.25	0.70	0.37
element	0.53	1	0.53	0.99	0.38	0.95
function	-0.42	0.94	-0.08	-0.59	-0.75	-0.75
trial	0.13	0.58	0.80	0.99	0.55	0.82
production	-0.11	-0.01	1	1	0.19	0.82
newspaper	0.80	0.79	0.30	0.2	0.17	0.31
energy	0.42	0.66	0.68	1	0.51	0.4
average	0.24	0.57	0.53	0.6	0.30	0.43

Notice that if all concordances of the ranking are analyzed, we observe that only *GBlog* concordances are correlated. Other SEs provide some correlation only if the first 50 concordances are selected. So, it seems that top of rankings are more adequate in terms of sense probability.

5 Conclusions

We have proposed an experimental method to verify whether the distribution of senses in the rankings obtained from SEs are balanced, representative, and coherent in linguistic terms. Taking *SemCor* as a balanced reference, we observed that the concordances retrieved by different SEs have low correlation with *SemCor* with regard to sense distribution. If we consider that the diversity of topics and domains in the web is close to that of most traditional open domain balanced corpora, we may infer from our experiments that the sense distribution bias is due to the fact that web engines rank their pages using non-linguistic criteria. It should be noted that the best correlation was achieved with SEs that only cover a part of the web (news and blogs), whose text sources are thus rather far, in terms of topic and genre distribution, from a traditional balanced corpus. By contrast, the worse correlation was achieved by the search engine (*WebCorp*) using the entire web as text source. We can surmise that ranking factors related to popularity or navigational queries introduce some non-linguistic bias in the concordances retrieved by general-purpose SEs. Furthermore, some SEs may retrieve concordances with serious problems concerning linguistic coherence (24% of concordances of *WebCorp* display problems of linguistic coherence). All these observations lead us to conclude that word sense information obtained by SEs used for knowledge extraction, word sense disambiguation, or lexicographic consultation might not be totally reliable.

References

1. Agirre, E., Olatz, A., Hovy, E. and Martinez, D. Enriching very large ontologies using the WWW. ECAI 2000, Workshop on Ontology Learning. Berlin (2000).
2. Agirre, E. and Martinez, D. The effect of bias on an automatically - built word sense corpus. In Proceedings of the 4rd International Conference on Language Resources and Evaluations (LREC). Lisbon, Portugal (2004).
3. Baroni, M., and Bernardini, S. (eds.) WaCky! Working papers on the Web as Corpus. Bologna: Gedit" (2006).
4. Baroni, M., and Bernardini, S. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC-2004 (2004).
5. Andrei Broder, "A taxonomy of web search," ACM SIGIR Forum 36, no. 2 (9, 2002): 3 (2002).
6. Ping Chen, David Brown, Andrew Tran, Noble Ozoka, Rafael Ortiz. Word Sense Distribution in a Web Corpus. IEEE Int. Conference on Cognitive Informatics (ICCI'10), 449-453 (2010).
7. Kilgarriff A. and Grefenstette G. Introduction to the special issue on the Web as corpus. Computational Linguistics vol. 29, p. 333-348 (2004).

8. Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT, pages 1048–1056 (2008).
9. Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04). Association for Computational Linguistics, Stroudsburg, PA, USA, Article 279 (2004).
10. Mihalcea, R. Semcor semantically tagged corpus. http://www.cse.unt.edu/rada/downloads.html (1998).
11. Morley, B. 'WebCorp: A Tool for Online Linguistic Information Retrieval and Analysis' in A. Renouf & A. Kehoe (eds.) The Changing Face of Corpus Linguistics, Amsterdam: Rodopi (2006).
12. Preslav Nakov, Svetlin Nakov, and Elena Paskaleva. Improved word alignments using the Web as a corpus. In Proceedigs of Recent Advances in Natural Language Processing (RANLP'07), pages 400– 405, Borovets, Bulgaria (2007).
13. Celina Santamaría, Julio Gonzalo, and Javier Artiles. Wikipedia as sense inventory to improve diversity in Web search results. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics (2010).
14. Skoutas, Dimitrios and Minack, Enrico and Nejdl, Wolfgang. Increasing Diversity in Web Search Results. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, 2010, Raleigh, NC: US (2010).
15. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast|but is it good? evaluating non-expert annotations for natural language tasks. In Proc. of EMNLP (2008).
16. Volk, M. Using the Web as Corpus for Linguistic Research. In Tähendusepüudja. Catcher of the Meaning. A Festschrift for Professor Haldu Õim. (eds. Pajusalu, R. and Hennoste, T.) Department of General Linguistics 3, University of Tartu (2002).