Working with the web as a source for dictionaries of informal vocabulary

Håkan Jansson
University of Gothenburg, Sweden

Informal vocabulary, e.g. slang, jargon and other forms of expression that are particular to different types of small or closed groups, is usually suppressed in writing that has passed an editorial process. That is to say with at least one important exception: the dialogues in works of fiction. This means that this type of vocabulary is not so readily gathered for the purpose of lexicon-making. Or this has nevertheless been the case up until recent years. But the constant stream of linguistic diversity on the Internet has given us new possibilities to tap into to the flow of colloquial and informal language.

The aim of this presentation is foremost to give a brief account of how the Internet could be 'harvested' for the purpose of creating corpora which include substantial amounts of informal language, and secondly, how to use these (in this case Swedish and Icelandic) corpora to gather candidates for headwords with informal markings such as **coll.**, **slang**, *and the like.*

1. The Internet as a corpus source

The Internet as a corpus source, or as a corpus itself, has been the object of many presentations, articles and even handbook chapters, so there will be no need to dwell further on this subject here – it will be enough to remind you of the references to a publication that gives an overview of the corpora of the *WaCky wide web* (Baroni et al. 2009) and the discussion of the topic in *The Oxford Guide to Practical Lexicography* (Atkins & Rundell 2008). Some interesting notes on the compilation of a web corpus as compared to the compilation of a traditional corpus, namely the BNC, are to be found in *Methods and tools for development of the Russian Reference Corpus* (Sharoff 2006). The question of the representativeness of a web corpus has been addressed in *Introducing and evaluating ukWaC, a very large web-derived corpus of English* (Ferraresi et al. 2008).

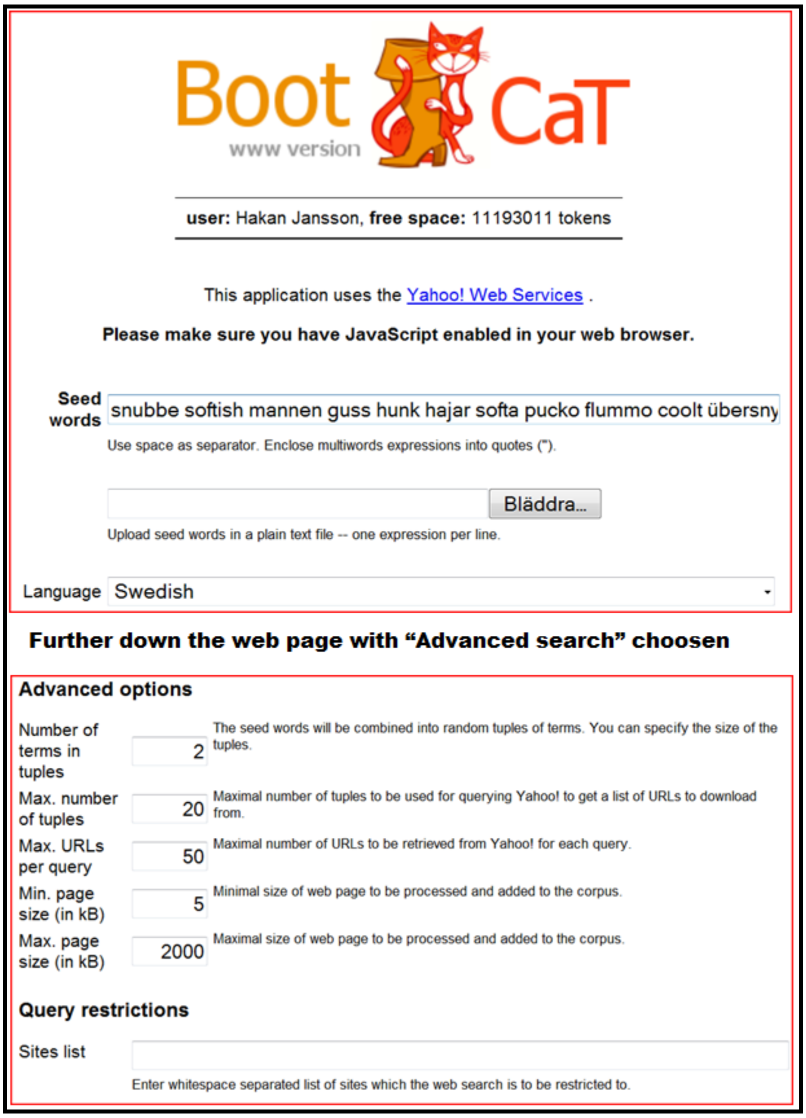
2. The corpus creation procedure

The work to be presented here has been carried out with the tools WebBootCaT and Sketch Engine, which both have been presented at Euralex and elsewhere – i.e. Euralex XI (Kilgarriff et al. 2004) and Euralex XII (Baroni et al. 2006). As presented in Baroni et al. (2006) the WebBootCaT scans the Web for the kind of vocabulary we set it to search for. Normally this is done by choosing so called *seed words*, that are expected to be typical of the type of vocabulary one is searching for.

In this study we will use a few Swedish and Icelandic slang words. In order to get a better control of the web-crawling process we will use WebBootCaT in the advanced mode, which gives us the opportunity of choosing the number of seed words required to be present in each web page which will be downloaded (cf. Fig.1, below).

In addition, Figure 1 shows that WebBootCaT in the advanced mode also gives the user a chance to choose the maximum number of combinations of seed words to be searched for, as well as the maximum number of URL:s to be downloaded. Only these three options in the advanced mode will be utilized here.

Since Swedish – and even more so Icelandic – are relatively small languages, in the sense that the amount of texts on the Internet is not overwhelmingly huge[1], the search procedures proposed here will not result in data overload. It is also convenient to be able to assume that the searches with Swedish or Icelandic seed words result in texts written by native speakers that have their roots in the local speech community (as opposed to English, but possibly also French, German and Spanish, which are used as *lingua franca* by non-native speakers).



user: Hakan Jansson, **free space:** 11193011 tokens

This application uses the Yahoo! Web Services .

Please make sure you have JavaScript enabled in your web browser.

Seed words snubbe softish mannen guss hunk hajar softa pucko flummo coolt übersny

Use space as separator. Enclose multiwords expressions into quotes (").

Bläddra...

Upload seed words in a plain text file -- one expression per line.

Language Swedish

Further down the web page with "Advanced search" choosen

Advanced options

Number of terms in tuples	2	The seed words will be combined into random tuples of terms. You can specify the size of the tuples.
Max. number of tuples	20	Maximal number of tuples to be used for querying Yahoo! to get a list of URLs to download from.
Max. URLs per query	50	Maximal number of URLs to be retrieved from Yahoo! for each query.
Min. page size (in kB)	5	Minimal size of web page to be processed and added to the corpus.
Max. page size (in kB)	2000	Maximal size of web page to be processed and added to the corpus.

Query restrictions

Sites list

Enter whitespace separated list of sites which the web search is to be restricted to.

Figure 1.

Another strategy, which also contributes to make the language size less important, is instead to set more restrictions on the domains to be downloaded. If the domains are restricted to, for example, websites of magazines that are closely tied to a certain type of trade, e.g.

[1] According to my estimate Google has indexed about 30 billion tokens of Swedish under the ***.se*** domain by mid november 2009. The approximate size was calculated by comparing the frequencies for 15 key words in *Swedish web corpus* with the frequencies for the same words at *Google.se*, utilizing the following formula:

$$\frac{\Sigma\ \textbf{token}\ \text{SweWebC}}{\boldsymbol{f}\ \textbf{key word}\ \text{SweWebC}} \approx \frac{\Sigma\ \textbf{token}\ \text{Google.se}}{\boldsymbol{f}\ \textbf{key word}\ \text{Google.se}} \rightarrow \Sigma\ \textbf{token}\ \text{Google.se} \approx \boldsymbol{f}\ \textbf{key word}\ \text{Google.se} * \frac{\Sigma\ \textbf{token}\ \text{SweWebC}}{\boldsymbol{f}\ \textbf{key word}\ \text{SweWebC}} \approx 30\ \text{billion}\ \textbf{token}$$

The chosen key words indicated that Google.se would have a size of 15 to 40 billion token, with an average of 30 billion.

construction business, you are bound to obtain a corpus that includes more than average of the vocabulary of that particular trade. My studies show, however, that in this case the structure of the general vocabulary is much the same as that of a bigger general corpus, that relies heavily on journalistic material, as in this case where *Press 97* of *Språkbanken* (the Language bank) of the University of Gothenburg was compared with an *ad hoc* gathered corpus, entirely composed of construction business magazines available on the Internet.

More surprising was to find that a corpus that was intended to replicate the corpora of *Språkbanken*, corpora which are characterized by a heavy bias towards journalistic texts, in the end resulted in a more informal structure. That corpus, the *Swedish web corpus*, available through *Sketch Engine*, turned out to be more akin to an *ad hoc corpus of informal web language* than it was to the corpora of *Språkbanken*. The Swedish web corpus was compiled by: 1. using some rather ordinary seed words, with special emphasis on verbs marking direct speech, and 2. subsequently limiting the search to a rather large number of specified domains. The domains were chosen among on-line magazines and newspapers as well as from websites which represent different organizations with the aim of influencing public debate, NGO's and others. The assumption was that the material being published there should have passed an editorial process of some kind, as well as have been composed with the intention of not being published on behalf of just the mere writer. During the work with the corpus it became evident that it contained a large number of texts typical of the informal web domain. This seemed to stem from the fact that a lot of the domains included in the corpus accumulation had discussion forums or blogs within them. This course of events resulted in a corpus with a different content than first intended, but since the changes seemed interesting the work was kept.

	Constr. mag.	**P97[2]**	**Swe wc[3]**	**Swe 1**
Highest ranking personal pronoun; rank in parenthesis	de [=they] (15)	de [=they] (15)	jag [=I] (15)	jag [=I] (12)
rank for *har* [=have, has]	12	14	17	17
deviance in occurance within 25 highest ranking types	jag [=I] missing	han [=he] only in this corpus		du [=you, sg] only in this corpus

Figure 2.

It should be kept in mind that the discovery of the patterns in structural characteristics within the corpora here mentioned, emanated from a comparison between corpora of type frequency within the top 100 segment. That comparison showed, as mentioned above, interesting similarities between, on the one hand, an *ad hoc construction business magazine corpus* (*Constr. mag.* in Figure 2) and the *Press 97* (*P97*) and, on the other hand, an *ad hoc corpus of informal language* (*Swe 1*) and the *Swedish web corpus* (*Swe wc*). Figure 2 shows that *Constr. mag.* and *P97* both have *de* [=they] as the most common personal pronoun referring to individuals (*det* [=it] is of course the most common pronoun in all corpora), while *jag* [=I] is the most common one in *Swe wc* and *Swe1*. It is widely accepted that use of the first person personal pronoun is typical of spoken language, but it is also more common in informal genres compared to other written genres, cf. Biber (1988).

The rank of the verb form *har* [=has, have] also shows a conspicuous pattern. Ranks 12 to 14 means that the type has a frequency that is 25 – 40% higher than a type with rank 17. *Har* is the second most common verb form in Swedish, second only to the copula *är* [=is]. This form

[2] P97 = Press 97, accessible at *Språkbanken*: http://spraakbanken.gu.se/ .

[3] Swe wc = Swedish web corpus, accessible at *Sketch Engine*: http://www.sketchengine.co.uk/ .

is, as it is in English, the finite constituent of any verb in perfect form, which according to Biber (in English) has ‘been associated with narrative/descriptive texts and with certain types of academic writing’ (Biber 1988: 223).

Finally, Figure 2 also gives some interesting examples of traits that further underlines the respective text structures: There is no ‘*I*’ present in the top ranks of the construction magazine, which is as expected, since it wishes to be descriptive and avoid the personal touch of first person narrative. The high ranking of *he* in the press corpus is yet another reminder of the fact that most of the news are about men in executive positions. There are first person pronouns present among the top 25 in this corpus, but they are outranked by the third person pronouns, as opposed to the web corpora where the first person ‘*I*’ is the most common person-referring pronoun. It is also noteworthy that it is only the corpus that was especially set up to collect informal web pages (*Swe 1*) that has the pronoun ‘du’ [=you, sg] among the top 25. This is probably a reflection of the ‘dialogue-like’ written exchange that is so typical of many of the different web forums that were represented in this corpus

3. Two corpora of informal language and what to be found in them

The approach taken for the corpora under this heading was to rely on the seed words only, and not specify any domains. There are, as is to be seen above, many ways to compile corpora from the Internet with tools like WebBootCaT. A tool like that makes it possible to create *ad hoc corpora* for different purposes, which is taken advantage of here. The aim of the study using these two corpora was to examine how informal endings are used in word formation in Icelandic and Swedish, with special interest taken in the degree of productivity in word formation and the grade of transparency of the compound. Examples of words and endings are presented in Figure 3 (below).

Language	Original word	PoS	Informal word	PoS	Ending
Swe.	Stallmästaregården (name of restaurant)	pn	Stallis	pn	-is
Swe.	flummig (=airy-fairy, unclear)	adj	flummo	n	-o
Icel.	Hressingarskálinn (name of restaurant)	pn	Hressó	pn	-ó

Figure 3.

The use of a corpus in this kind of a study is of course essential. There are numerous other ways of searching for any particular word, but it is only in a corpus you can search for words and forms that you don’t already know exist.

In this case the corpus was searched and filtered by applying regular expressions in the Sketch Engine. The return included of course quite a lot of words that happened to end with ***-is***, ***-o*** or ***-ó***, that were not at all informal, as the word searched for should be. The size of the material was, however, not bigger than that it was quite easy to go through the result manually. In a corpus of just over 3.6 million tokens only about 2.200 unique words (type) ended with ***-o***, of which less than 10% were of interest for the study.

The results of the study showed that all three endings are productive, with the Swedish ***-is*** most often connoting a general increase in the scale towards informal and/or intimate, while the Swedish ***–o*** suggests a derogatory or generally informal intention on behalf of the speaker (or writer). The Icelandic ***-ó*** has a function that is rather similar to the Swedish ***-is***.

The findings are probably less interesting in themselves, in a context such as this conference, but the methods could be applicable in any language. It should also be noted that many of the words in the corpora were new, even to a lexicographer with language habits characterized by diversity. One of the words that caught my attention through this kind of corpus work was rather common on the Internet but unknown to me and most of my colleagues. The word was *bloppis* derived from *blogg* (blog) and *loppis* (garage sale), meaning ‘garage sale on the blog’.

4. The likelihood of finding informal vocabulary in different corpora

Is informal language of lexicographical relevance? Well, since most dictionaries make use of labels like *informal*, *slang* or *offensive*, it would seem illogical to answer the question in any way but in the affirmative. In practice, however, many corpora that are used for dictionary making are probably rather likely to be biased against informal vocabulary, on the grounds of it being underrepresented in the kind of texts that make up most corpora. As has already been shown in Figure 2 (above), corpora that have been assembled by using *informal* or *slang* seed words are more likely to be akin to spoken language and then, as a consequence, also to contain vocabulary that are more characteristic of spoken language.

In Figure 4 (below) we can see this illustrated in a comparison between *PAROLE* and *GP04* of *Språkbanken*, two of my corpora, and Google under the .se domain. *PAROLE* is a morpho-syntactic tagged corpus that was compiled within the EU-project PAROLE that was finished in 1997. *GP04* is the entire published text of the leading Gothenburg newspaper *Göteborgs Posten* in 2004, cleansed from adverts and ‘table-like’ texts. *Swedish web corpus* was assembled by downloading files from the Internet using the functions in WebBootCaT that allows the user to specify seed words and domains to be downloaded from. The seed words were common quotation verbs and a few adverbs, with the idea of getting a slight bias for quoted speech. Since the seed words were that general, the delimiting factor was instead the domains. The download was limited to the domains of all leading newspapers and then an average spread of Swedish newspapers, all magazines that were freely available, all political parties, an average spread of ministries and public agencies and NGO:s. The *AdHoc ‘slang’ corpus* was collected through WebBootCaT, using typical slang words as seed words, with no limitation on domains. Google-frequencies must, as ever, be treated with caution, but this time for other reasons than the ones usually debated.

	PAROLE	**GP04**	**Swedish web corpus**	**AdHoc ‘slang’-corpus**	**Google.se**
	(19.4 million)	(19.4 million)	(18.0 million)	(3.6 million)	(c. 30,000 million)
dagis[4]	410	418	370	84	192 000
snackis[5]	--	5	12	10	108 000
flummo[6]	--	--	--	14	4 040
bloppis[7]	--	--	--	3	44 100

Figure 4.

Dagis is common parlance for nursery school in Swedish and not surprisingly present in all corpora. As expected *dagis* is also included in leading Swedish dictionaries such as Svenska Akademiens ordlista (SAOL) [the word list of the Swedish Academy] and Svensk Ordbok (SO) [Swedish dictionary]. None of the other words in Figure 4 are to my knowledge included

[4] *dagis* = nursery school.

[5] *snackis* = much discussed thing.

[6] *flummo* = airy-fairy person, dope head.

[7] *bloppis* = ‘garage sale’ on the Internet.

in any general Swedish dictionary[8], even though both *snackis* and *flummo* are well known and widely used words in Swedish. *Snackis* is often used referring to 'hot topics', and thence in connection to sports and in some radio talk shows. *Flummo* is mostly used among teenagers as a derogatory remark about someone that has difficulties in 'keeping his/her act together'.

Both *snackis* and *flummo* should be well known to lexicographers, but since they are not that common in most widely used corpora, they might easily be overlooked. *Bloppis* on the other hand was, as mentioned above, unknown to me and my colleagues until I found it in my corpus. It seems to be a well known and widely used (cf the Google-frequency) word – but only in certain segments of the 'blogosphere'. I found it to be especially common among the large group of young fashion-bloggers that constitute a very active community on the Swedish Internet. The use of that particular word in a very active Internet community would also explain why it would be that unknown among large groups of speakers and that scarcely represented in corpora. This of course also underlines the fact that a high Google frequency does not necessarily indicate widespread use of a word in large groups.

The exceptionally high Google-frequency for *snackis* in relation to the corpus-frequencies is another interesting example of how extralinguistic factors can upset Google statistics and thus make those figures unreliable. One of the big Swedish mobile operators has a mobile subscription named *snackis*. Quite a few Google hits are of course referring to that subscription. Another part of the explanation to the high frequency would be akin to the rationale behind *bloppis*: The word has become a bit *á la mode* among bloggers, which is quite logical since it refers to 'hot topic' – and blogs by definition more often then not are about hot topics.

5. In conclusion

If I were to summarize my experiences of creating and using web corpora for the purpose of doing vocabulary research, I would emphasise four main points:

- Others have shown that a balanced corpus is a relative concept. When there is a difference in content, it is not an easy task to prove *which part of the content* in *what corpus* it is that upsets the balance.
- Relying on specification of domains rather that using an extensive and specific set of seed words was by large a success. A caveat though: some domains may contain other subsets of language than expected.
- With exception for the very specialized part of the vocabulary, there is less variation between, on the one hand, texts and sources that we usually associate with LSP (Language for Special Purposes) and the edited part of newspaper websites, on the other, than it is within the newspaper websites themselves.
- When it comes to studying different aspects of informal vocabulary, using seed words to compile specialized *ad hoc* corpora is likely to yield more interesting findings than searches in even rather large corpora.

[8] *flummo* is included in Kotsinas (2000), the most comprehensive Swedish slang dictionary at present.

References

Internet resources

Sketch Engine. [Brighton: Lexical Computing Ltd]. http://www.sketchengine.co.uk/.

Språkbanken. [The Language Bank at University of Gothenburg]. [Göteborg: University of Gothenburg]. http://spraakbanken.gu.se/.

Dictionaries

Svenska Akademiens ordlista över svenska språket. (SAOL), 13 upplagan. Stockholm: Svenska akademien : Norstedts akademiska förlag [in distribution]. 2006.

Svensk ordbok utgiven av Svenska Akademien. (SO), Stockholm: Norstedt [in distribution]. 2009.

Books and other printed resources

Atkins, B. T. S.; Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Baroni, M. et al. (2009). 'The WaCky wide web: a collection of very large linguistically processed web-crawled corpora'. *Language Resources and Evaluation* 43 (3). http://dx.doi.org/10.1007/s10579-009-9081-4.

Baroni, M. et al. (2006). 'WebBootCaT: a web tool for instant corpora'. In: Corino, E.; Marello, C.; Onesti, C. (eds.). *Proceedings of the XII EURALEX International Congress.* Turin: Edizioni dell'Orso. Vol. 1. 123-131.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Ferraresi, A. et al. (2008). 'Introducing and evaluating ukWaC, a very large web-derived corpus of English'. In: *4th Web as Corpus Workshop (WAC-4); Can we beat Google?* Marrakech. http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf.

Kilgarriff, A. et al. (2004). 'The Sketch Engine'. In: Williams, G.; Vessier, S. (eds.). *Proceedings of the XI EURALEX International Congress.* Lorient: Université de Bretagne-sud. Vol. 1. 105-116. http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw.

Kotsinas, U-B. (2000). *Norstedts svenska slangordbok.* Stockholm: Norstedts ordbok.

Sharoff, S. (2006). 'Methods and tools for development of the Russian Reference Corpus'. In: Wilson, A.; Archer, D.; Rayson, P. (eds.). *Corpus linguistics around the world.* Amsterdam: Rodopi.