What kind of corpus is a web corpus?

Janne Bondi Johannessen
Tektslab, ILN, University of Oslo
jannebj@iln.uio.no

Emiliano Raul Guevara
Tektslab, ILN, University of Oslo
e.r.guevara@iln.uio.no

Abstract

This paper discusses an investigation into the Norwegian NoWaC corpus. We have compared this web corpus with one corpus of spoken language and one of written language. For nearly all variables that we look at, the web corpus sides with the written corpus, not the spoken one. Thus, despite including language samples from blogs and web forums, NoWaC does not appear to be more speech-like. One exception is interjections, which it does have to some larger extent than the written corpus. It also has taboo words, lacking in the other two. The comparisons have been made purely on the basis of frequency lists, showing that this is a possible and simple way of comparing corpora. We use both a qualitative and quantitative method. In the latter, (log) relative frequency plots show an almost linear relation between NoWaC and the written corpus.

1 Credits

This work depends on the existence of four Norwegian corpora. We are grateful to those who have been central in compiling and developing them. These are: Anne Marit Bødal, Kristin Hagen, Signe Laake, Anders Nøklestad, Joel Priestley (all are former or present staff at the Text Laboratory), Øystein Alexander Vangsnes and Tor Anders Åfarli (central in the compilation of the Nordic Dialect Corpus, and in the network ScanDiaSyn), and the VD group at USIT, UiO, for computing assistance. Finally we want to acknowledge the Keywords for Language Learning for Young and adults alike, KELLY, with project leader Sofie Johansson Kokkinakis, which provided the funding for the compilation and processing of the NoWaC lemma frequency list.

2 Introduction

It is well known that spoken and written language differ from each other in a variety of ways. This has been discussed and shown in, e.g. Akinnaso (1982), Chafe and Tannen (1987), McCarthy (1998), Miller and Weinert (1998), Biber et al. (1999), a special issue of *Studia Linguistica* for spoken language (Johannessen 2008), for Swedish, Allwood (1998) and for Norwegian, Johannessen and Hagen (2008). For the same reason corpora based on written and on spoken language also differ from each other, and are thereby useful for different purposes.

Since the web started to be used as a corpus, and now when there are actual corpora built by mining it (cf. Baroni and Bernardini 2006, Baroni and Kilgariff 2006, Kilgariff 2007, and the workshop series Web as a Corpus), it is interesting to investigate what kind of language the web actually contains. No doubt this will differ from language to language, so we stress here that our investigation is based on Norwegian.

Our aim is two-fold. On the one hand we want to investigate the Norwegian web corpus NoWaC to try to uncover what kind of a corpus it is w.r.t. the spoken/written dimension. The other is to see whether using frequency lists is a useful method in order to determine the differences in genre/register between different corpora.

We will compare NoWaC with one written and one spoken language corpus. Our hypothesis is that NoWaC will be somewhat closer to the spoken language corpus than the written corpus. This is due to the fact that Internet is widely available (88 % of people aged 16/79 used Internet at home in 2010, according to Statistics Norway). In fact Norway was the first country after the USA to have Internet. One would therefore assume that a lot of the contents on the Norwegian web pages would be informal forums and blogs that are by hypothesis speech-like.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 122–129

However, as we shall see, surprisingly, NoWaC is not like a spoken corpus, in fact it is in many ways even more formal linguistically than the written corpus.

3 Three corpora – three frequency lists

In order to investigate what kind of texts NoWaC contains, we will compare it with two other corpora, which we will describe in turn. But first, let us start with NoWaC.

NoWaC (for details about the methodological steps taken during the contruction of the corpus, see Guevara 2010) is a corpus mined from the web using the bootstrapping guidelines as described in Baroni and Bernardini (2006). It contains about 700 million words of Bokmål Norwegian. With the strict legislation in Norway (the Personal Data Act) and its implementation by the Privacy Ombudsman for Research, mining the web and keeping the data is not legal without special permission, which this corpus obtained from the Ministry of Culture. The version we have used here was tagged with a statistical tagger (Treetagger), trained on a small manually annotated corpus available at the Text Laboratory, UiO.

In addition we have used what we here choose to call the Written Language Corpus, although its official name is the Oslo Corpus of Written Norwegian Bokmål Texts (Johannessen et al. 2000). It has 18 million words, is tagged with the rule-based Oslo Bergen Tagger, and contains about 10 % fiction (in addition to 50 % newspaper and magazine texts and 40 % non-fiction).

Finally we use a combination of two speech corpora, here called the Spoken Language Corpus. These are the Nordic Dialect Corpus (Johannessen et al. 2009) and the NoTa-Oslo: Norwegian Speech Corpus - Oslo part (Johannessen et al. 2007). These contain approximately 2.2 million words from recordings of spontaneous speech in dialogue. They have been transcribed orthographically, and are thus immediately comparable with the written language corpus, and have been grammatically tagged with a statistical tagger.

From these corpora we have compiled frequency lists. Each list contains the 6000 most frequent lemmas of that corpus. It is these lists that we will use in the corpus comparisons in this paper.

Comparing corpora has been a common concern in computational linguistics and in corpus linguistics since the advent of large electronic corpora in the 1990's (see, among others, Hofland and Johansson 1982, Kilgarriff 1997, Rayson and Garside 2000). However, all the measures and methods proposed in the literature concentrate on one of the following two cases:

- comparing a sample (specialistic) corpus to a large(r) reference corpus
- comparing two corpora of roughly the same size

Our situation is different. We have a very large corpus of which we want to determine the linguistic variety, and two smaller reference corpora. In addition, some of the previous methods (e.g. Kilgarriff's 1997 "Known Similarity Corpora") cannot be applied to our data without leaving out a substantial part of NoWaC. In what follows we will present a combination of qualitative and quantitative methods addressing our research question.

4 Comparing corpora through frequency lists: a qualitative point of view

Biber et al. (1999) use corpus analysis to distinguish the styles of spoken and written language. They focus on syntactic structures and collocations, which we will not do in this task. However, some syntactic constructions are accompanied by certain words. Biber et al. (op.cit, p.691) find that clauses introduced by *whether* are typically less common in spoken language than *if*.

Incidentally, Norwegian, too, has two words for the introduction of interrogative subordinate clauses, *om* and *hvorvidt*. The results from our three corpora are given in table 1. (Throughout this section, the number represents relative frequency, obtained by dividing the frequency count by corpus size, figures rounded for convenience.)

	Spoken	Written	NoWaC
om 'if'	0.00198	**0.00718**	0.00495
hvorvidt 'whether'	–	**0.00006**	0.00003

Table 1: Two subordinating conjunctions meaning 'whether'.

Hvorvidt has a more bookish feel to it, and this intuition is confirmed by the frequency lists, where it is absent in the spoken language corpus.

But notice that in NoWaC its frequency is relatively lower than in the written text corpus. *Om* confirms this point, showing a strong difference between the spoken corpus on one side, and the written corpus and NoWaC on the other side, but with NoWaC closer to the speech corpus.

Bick (2010) compares five English corpora going from chat, e-mail, to one formal and one informal speech corpus, plus a written text corpus, and shows that many features expected to be more typically informal are indeed so. For example, there is relatively little subordination in the chat corpus. Looking at the subordinator *at* ('that') our results again show that NoWaC is closer to the spoken corpus by a small margin. The written corpus is the most formal of all.

	Spoken	Written	NoWaC
At (‘that’)	0.00461	**0.01463**	0.00887

Table 2: The subjunction *at* ‘that’.

A very interesting finding by Bick (2010:727) regards the distribution of pronouns. He finds that the chat corpus, with live (written) dialogue, has nearly three times as many 3p pronouns as 1p ones, and that the spoken corpus also scores high here, with about twice as many. This is also true of his written corpus, which contains a lot of fiction. The monologues that the e-mails consist of and the formal speeches in his formal speech corpus have very different figures, where the 1p pronouns are more frequent than the 3p ones.

We have tested our three corpora for the singular pronouns of all three grammatical persons. From Bick’s paper we would expect our spoken corpus to show the same distribution as the chat corpus, and possibly the written corpus to be quite different, given what we have seen for the other categories above. For NoWaC we would expect it to be closer to the written corpus, on the basis of what we have seen above. The results are given in table 3.

	Spoken	Written	NoWaC
jeg (‘I)	**0.02193**	0.01130	0.00875
du (‘you’)	**0.01147**	0.00429	0.00507
han (‘he’)	0.00409	**0.01300**	0.00290
hun (‘she’)	0.00214	**0.00717**	0.00144

Table 3: Personal pronouns

For ease of exposition we illustrate the numbers in the following chart, with the 3p shown cumulatively.

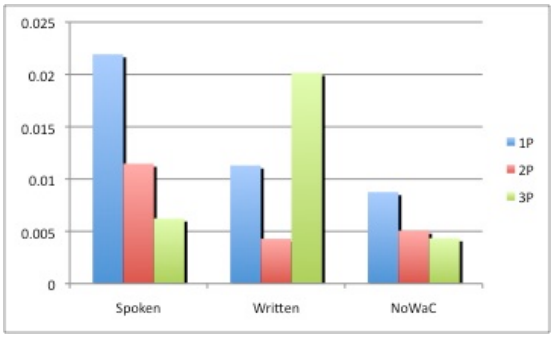

Fig 1: Personal pronouns as chart

However, we do not find quite the same as Bick. The spoken corpus, as expected, has very a high frequency for 1p and a relatively high frequency for 2p. On the contrary, the written corpus that has a relatively high number of 3p pronouns w.r.t. 1p and a lower frequency for 2p. NoWaC shows a mixed picture: it seems to pattern with the spoken corpus regarding 3p, but with the written corpus regarding 1p and 2p.

Why would this be the case? For the spoken corpus the answer has to be found in the way it was recorded. The informants were instructed carefully at the beginning of each session: they were told to avoid (for reasons to do with legal protection of individuals) talking about other people they knew – although they were for the most part also told that it would be acceptable to talk about people who were already in the public eye. This would be the reason that 3p pronouns are not as frequent as they would otherwise have been.

Several researchers look at the difference between spoken and written Norwegian language in Johannessen and Hagen (2008). Both Svein Lie’s paper and Søfteland and Nøklestad’s paper focus on the particular word *sånn* ‘such’, which is very typical of spoken language. Vangsnes argues that there is a systematic difference between the status of questions words, with *åssen* ‘how’ as the most colloquial. Fjeld’s paper discusses the many lexical items that are hard to find in dictionaries, which often have a written bias, however implicit. In the table below we investigate words from these papers.

	Spoken	Written	NoWaC
sånn (‘such’)	**0.00781**	0.00037	0.00024
åssen (‘how’)	**0.00049**	0.00002	0.000004
do (‘loo’)	0.00003	0.00002	**0.00004**
dass (‘loo’)	**0.000004**	-	0.000003

Table 4: Norwegian words typical of spoken language

The first two words in table 4, which are grammatical words, show the same as we have seen earlier, too. NoWaC tends to pair off with the written corpus, while the spoken language corpus is in a different league. The last two words are typical words used in the spoken domain. It is surprising that they occur in the written corpus and NoWaC. We had to further investigate this, and found that both were mixed with other other lemmas that happened to be nouns according to our tags. For *do* 'loo', the less offensive of the two, about half the occurrences in the written corpus are parts of Vietnamese names (*Tranh Thi Le Do*), English phrases (*back/ so I gotta do now*), names of sports (*Tae Kwon Do*), which means that it really should have a much lower relative frequency.

The word *dass* 'loo' is very colloquial and it is not among the 6000 most frequent words in the written corpus. A closer look at the hits in NoWaC reveals that some of the occurrences refer to the surname of the hymn writer Petter Dass, which should of course have been excluded from the lists. It also has many examples of *dass* used in a metaphoric sense (in expressions like 'go down the drain'). To conclude, we find NoWaC to have a bit in common with spoken language w.r.t. *dass*, which is the most colloquial word, although the relative frequencies are rather low. For the other typical spoken words, it sides more with the written corpus.

Interjections are a category that obviously belongs to the spoken domain, so this a category worth looking into.

	Spoken	Written	NoWaC
Ja (‘yes’)	**0.02701**	0.00056	0.00025
Nei (‘no’)	**0.00737**	0.00042	0.00021
Oi (‘oh’)	**0.00019**	-	-
Uff (‘oh’)	**0.0001**	-	0.000004
Nja (‘well’)	**0.00007**	-	-
Jøss (‘Jesus’)	**0.00005**	-	-
Fy (‘bad’)	**0.00004**	-	-
Æsj (‘yuck’)	**0.00003**	-	-
Au (‘auch’)	**0.00003**	-	-

Table 5: Interjections

The interjections show that NoWaC does have some speech-like contents, w.r.t. ‘yes’ and ‘no’, but so does the written corpus, both with low figures, as does NoWaC for the one additional internjection it has: *uff* ‘oh’. However, NoWaC does contain a lot of different interjections which did not make it into our list of the 6000 most frequent words in the corpus. Looking at the concordances, it is obvious that it is without a doubt blogs and forums, i.e. kinds of dialogue, which are the text types where these interjections are found. An example is given in (1).

(1) anonym : **æsj** , det suger
anonymous: yuck, it sucks

Swearing and taboo words are very rare in serious, written texts, so this is an interesting test for NoWaC. The results are given in table 6.

	Spoken	Written	NoWaC
Faen (‘devil’)	0.00004	**0.00006**	0.00003
Herregud (‘Lord God’)	**0.00007**	0.00002	0.000009
Fitte (‘female sexual organ’)	-	-	**0.000007**
Pikk (‘male sexual organ)	-	-	**0.000019**

Table 6: Swearing and taboo words

Unexpectedly, the first swear word has a relatively high position in the written material and in NoWaC, while the second one is much more common in the spoken corpus.

However, for taboo words – the last two rows in the table – the NoWaC corpus really differs from both corpora. It is the only corpus containing them, which makes it more informal than even the spoken language corpus.

One must ask why that is. For the spoken corpus the answer is easy: All recordings were done under controlled circumstances. All informants knew that they were being recorded and had a camera directed towards them. While it may be somewhat ‘cool’ to use swear words by some people (notably young men), swear words and taboo words are still embarrassing. For NoWaC, we had to investigate a bit further. Although we have not looked at all the examples, we have checked some, including googled them for their original context, and found that although clearly the taboo words sometimes occur on informal forums they very often are from porn sites.

To conclude: For typically written language variables, such as formal subordinators, NoWaC is like a written corpus. Looking at variables that will say something about the extent to which spoken topics are concerned, such as subordination, NoWaC is still like a written corpus, although by a small margin. Checking for spoken version words of those that have several variants, NoWaC is still also like a written corpus.

Typically colloquial elements like interjections, swear words and taboo words do not provide us with clear ways to distinguish the corpora and, actually, point out issues with the reliability and "naturalness" of the spoken corpus.

However, we should add that there are also many other interjections among the hundred most frequent words in the spoken corpus, which are not in NoWaC or the written corpus. These are typical discourse markers, especially OCMs (own communication management) used with turn-taking, which do not have a regular spelling, but they have to be regarded as words, indeed interjections, since they have a clearly identifiable phonology and semantics. One example is *mm* (disyllabic, with toneme 2), meaning: 'yes, I agree', which is the 30th most frequent word in the spoken corpus.

The relatively high proportion of taboo words does not necessarily show a speech-like quality of the corpus, but does show that the corpus contains some material not usually contained in carefully compiled corpora.

5 Comparing corpora through frequency lists: a quantitative point of view

As we pointed out in the introductory section, previous methods to compare linguistic corpora rely on the assumption that the comparison is made with corpora of at least the same size. In our case, however, the target of the study is over 300 times larger than the spoken reference corpus. We instead propose to systematically extend the kind of comparison that was presented in the previous section, that is, a correlation analysis between the relative frequencies of words in the different corpora.

In order to accomplish this, we first apply a logarithmic transformation to the frequency counts in the various corpora and create a data frame containing only the lemmas which are present in all the lists with the same POS tag (1810 items). Let us start by simply plotting the obtained data (see Fig. 2 and 3).

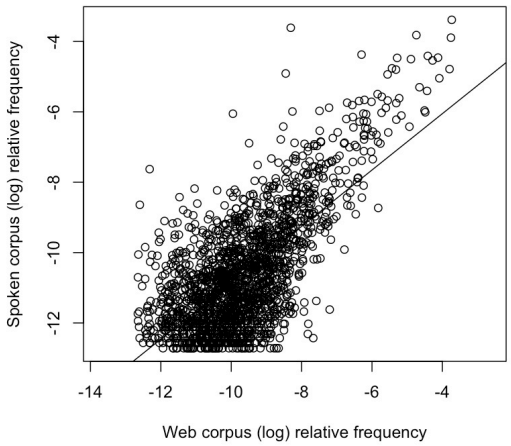

Fig. 2: A plot of corresponding frequencies in the spoken and web-based corpora

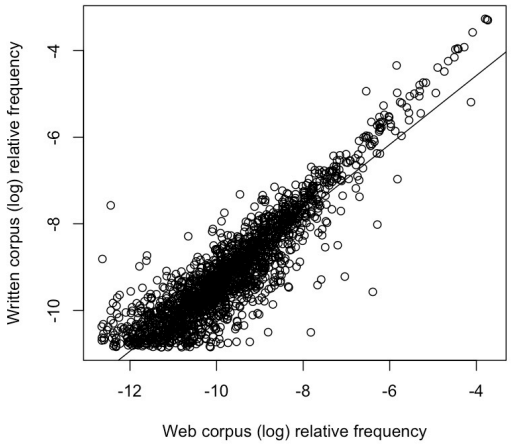

Fig. 3: A plot of corresponding frequencies in the written and web-based corpora

These simple plots already show that the frequencies from the NoWaC and the written corpus have an almost linear relation, while the comparison between web-data and the spoken corpus is much sparser. For an even more striking difference, compare the sparseness of the first cloud of data points with respect to the indicated regression line (simple linear regression).

In addition, the relation of the spoken data with our reference written corpus resembles very closely the cloud that we can see in the first figure above:

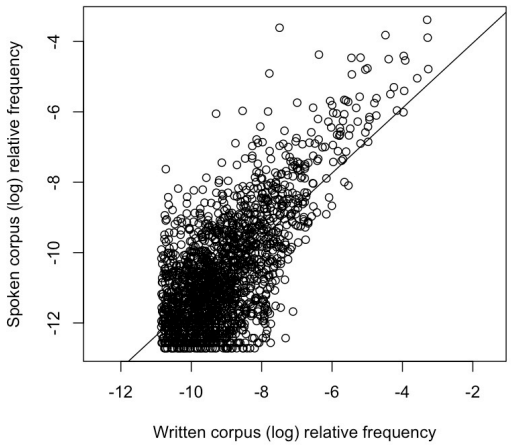

Fig. 4: A plot of corresponding frequencies in the spoken and written corpora

In other words, the spoken corpus seems to bear the same type of loosely linear relation to both the written corpus and NoWaC. On the other hand, NoWaC and the written corpus show a much tighter, linear correspondence which actually approximates quite closely a linear regression line.

To study these relations in depth we calculated Kendall's *tau* coefficient of correlation between the log frequencies obtained from NoWaC and each of the reference corpora. Kendall's *tau* is a robust non-parametric hypothesis test for statistical dependence that does not make any assumption about the distribution of the data. The test is a measure of rank correlation (related to Spearman's *rho*) that is able to deal with tied ranks in the data. We summarize the results in the following table:

Data	Correlation	p-value
NoWaC ~ Spoken	0.4098755	< 0.001
NoWaC ~ Written	**0.705881**	< 0.001
Spoken ~ Written	0.3955764	< 0.001

Table 7: Rank correlation

Clearly, the frequencies from all the used corpora are statistically correlated: this comes as no surprise, given that they share the same language and a great part of the vocabulary. However, the kind of language that was sampled in NoWaC is more closely correlated to the written corpus than to the spoken corpus (differences between the correlation coefficients statistically significant, two tailed p-value < 0.001).

In addition, we also calculated the correlation between the two reference corpora, whose result shows that there is less correlation between them than for each with NoWaC.

We interpret these results as indicating that our web corpus contains primarily a kind of language that is typical of the written register. However, some of its traits are in common with data from the spoken corpus.

In other words, if we consider the spoken and written reference corpora as two endpoints, NoWaC stands between them but not exactly half-way: it is significantly closer to the written end of language.

6 Conclusion

From a purely qualitative point of view, NoWaC is a written language corpus when it comes to proportions of typically written variables. Wherever a word exists in a formal and informal style, the formal style is by far the most common.

However, NoWaC does have interjections that are typical of dialogue, revealing this way that it does have some qualities shared with the spoken corpus. They are also considerably less frequent than in the spoken corpus. NoWaC also has a share of informal words and taboo words, showing that it contains texts that are not common in manually and carefully crafted written corpora, i.e. those that are based on texts from established publishing houses, legal documents etc.

These findings were further corroborated by a test of correlation between the frequencies from all three corpora. NoWaC shows a relatively stronger correlation to the written reference corpus, although it is also correlated significantly to spoken data.

The simple ideas and methods put forward in this paper provide us with plenty of novel insight. The web is the largest source of linguistic data, mostly text. We must be prepared to deal with the intricacies and peculiarities of this source. Although the web's register is primarily written (at least as it has been sampled in NoWaC), we might be witnessing the birth of a distinct register which contains elements of colloquiality and vocabulary from the spoken language.

References

Allwood, Jens. 1998. Some Frequency based Differences between Spoken and Written Swedish. In *Proceedings from the XVI:th Scandinavian Conference of Linguistics*, Department of Linguistics, University of Turku, Finland.

Akinnaso, F. 1982. On the differences between spoken and written language. *Language and Speech*, 25, 2, 97–125.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data*. Cambridge, Cambridge University Press.

Baroni, Marco and Silvia Bernardini (eds.) 2006. *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT.

Baroni, Marco and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of European ACL*, Trento, Italy.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd, Harlow, Essex, UK.

Bick, Eckhard. 2010. Degrees of Orality in Speech-like Corpora: Comparative Annonatio of Chat and E-mail Corpora. In Otoguro, Ryo; Ishikawa, Kiyoshi; Umemoto, Hiroshi; Yoshimoto, Kei; Harada, Yasunari (eds.): *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Waseda University, Sendai, Japan: 721–729.

Chafe, William and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology* 16, 383–407.

Fjeld, Ruth E. Vatvedt. 2008. Talespråksforskningens betydning for leksikografien. In Johannessen and Hagen (eds.), 15–28.

Guevara, Emiliano. 2010. NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Los Angeles, California, 1–7.

Hofland, Knut and Stig Johansson. 1982. *Word Frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Johannessen, Janne Bondi, Anders Nøklestad and Kristin Hagen. 2000. A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts. In Gavrilidou, M., G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (red.) *Proceedings, Second International Conference on Language Resources and Evaluation (LREC 2000)*, Aten, 1725-1729.

Johannessen, Janne Bondi; Hagen, Kristin; Priestley, Joel; Nygaard, Lars. 2007. An Advanced Speech Corpus for Norwegian. In: *NODALIDA 2007 PROCEEDINGS*. Tartu: University of Tartu 2007. s. 29-36.

Johannessen, Janne Bondi. 2008. *Studia Linguistica: Special issue on spoken language.* Vol. 62, issue 1.

Johannessen, Janne Bondi and Kristin Hagen (eds.). 2008. *Språk i Oslo. Ny forskning om talespråk.* [Language in Oslo. New Research on Spoken Language] Novus forlag, Oslo, Norway.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.*

Kilgarriff, Adam. 1997. Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. In "Proceedings of the Fifth Workshop on Very Large Corpora", Beijing and Hong Kong, China. Association for Computational Linguistics.

Kilgariff, Adam. 2007. Googleology is Bad Science. *Computational Linguistics* 33 (1): 147-151.

Lie, Svein. 2008. Veldig sånn festejente. In Johannessen and Hagen (eds.), 78–95.

McCarthy, Michael. 1998. *Spoken language and applied linguistics.* Cambridge University Press, Cambridge, UK.

Miller, Jim and Regina Weinert. 1998. *Spontaneous Spoken Language. Syntax and Discourse.* Oxford University Press, Oxford, UK.

Norway Statistics: http://www.ssb.no/

Rayson, Paul and Roger Garside. 2000. Comparing Corpora using Frequency Profiling. In Adam Kilgarriff and Tony Berber Sardinha (eds.) "Proceedings of the Workshop on Comparing Corpora", Hong Kong, China. Association for Computational Linguistics.

Søfteland, Åshild and Anders Nøklestad. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Johannessen and Hagen (eds), 226–234.

Vangsnes, Øystein Alexander Vangsnes. 2008. Omkring adnominalt åssen/hvordan i Oslo-målet. In Johannessen and Hagen (eds), 50–62.

Corpora

Nordic Dialect Corpus: http://www.tekstlab.uio.no/nota/scandiasyn/index.html Read about it in Johannessen et al. (2009).

NoWaC: http://www.hf.uio.no/iln/tjenester/sprak/korpus/skriftspraksk orpus/nowac/ Read about it in Guevara (2010).

NoTa-Oslo: Norwegian Speech Corpus - Oslo part: http://www.tekstlab.uio.no/nota/oslo/english.html Read about it in Johannessen et al. (2007).

The Oslo Corpus of Tagged Norwegian Texts: http://www.tekstlab.uio.no/norsk/bokmaal/english.html Read about it in Johannessen et al. (2000).

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955