Recent developments in Norwegian corpus lexicography

Gisle Andersen[1] *

[1] NHH Norwegian School of Economics

Abstract

This paper gives an account of recent efforts within corpus-based lexicography in Norway. I explore the lexical neology database that has been developed in the Norwegian Newspaper Corpus project (Andersen and Hofland 2012). The aim of the paper is to show how this resource has been used for practical lexicographical work in two dictionary projects representing the Nynorsk and Bokmål/Riksmål varieties of Norwegian, respectively.

Keywords: lexicography; standardisation; corpora; neologisms; neology; Norwegian; Bokmål; Nynorsk; Riksmål

*** Principal contact:**
Gisle Andersen, Professor
Dept. of Professional and Intercultural Communication, NHH Norwegian School of Economics,
Helleveien 30, NO-5045 Bergen, Norway
Tel.: +47 55 95 93 26
E-mail: gisle.andersen@nhh.no

1. Introduction

Since the advent of corpus linguistics, one of the most useful ways of exploiting corpora has been for the purpose of standardising the vocabulary of a language through the making of dictionaries. The field of lexicography has gained substantially from the development of new corpora and tools for monitoring ongoing language development (Atkins and Rundell 2008; Grefenstette 2002; Ooi 1998; Pulcini 2008). Initiated by Atkins and Sinclair in the 1970s, the Collins COBUILD project was the first effort to use a corpus as its main source of knowledge about words and their use in the language (Sinclair 1987). This initiative has been described as a revolution which "changed the principles and methods of dictionary making" (Pulcini 2008: 189) and which enabled lexicographers to "view the evidence of how a word was used without the arbitrary filter of who thought what was an interesting example of a word" (Kilgarriff and Tugwell 2002: 125). Since the turn of the millennium, it has become increasingly common to develop web-based corpora (Fletcher 2007; Hundt et al. 2007; Kilgarriff and Grefenstette 2003; Renouf 2007a) and to use these as a basis for lexicographic work (Grefenstette 2002). As Atkins and Rundell (2008: 3) put it, in our day and age, "all good dictionaries take corpus data as their starting point".

Since its establishment in 1998, the development of the Norwegian Newspaper Corpus[1] (henceforth NNC) (Andersen and Hofland 2012; Hofland 2000) has had a major impact on corpus research in Norway. In brief, the NNC is a web-based monitor corpus of more than 1 billion words of newspaper text that grows on a daily basis. This resource has stimulated substantial corpus-based research in a variety of fields, as evidenced by the contributions to a recent collective volume (Andersen 2012c), which deal with topics such as anglicisms in Norwegian (Andersen 2012a), morphosyntactic variation (Dyvik 2012), terminology relating to the financial crisis (Kristiansen 2012), metonymy/vagueness (Halverson 2012), etc. Of particular salience has been the use of the corpus for investigating developments in the Norwegian lexicon (Andersen 2005, 2010, 2011a, 2011b; De Smedt 2012; Fjeld and Nygaard 2012; Kristiansen and Andersen 2012). The current paper gives a survey of some of the work that has been carried out, but, unlike previous accounts, it draws attention to two individual projects that have been commissioned by external bodies, notably dictionary projects representing Nynorsk and Bokmål/Riksmål. Thus, the current work accounts for applied NNC-based research which has served as input for lexicography and standardisation efforts.

In the following, I first give a brief description of the system for neology extraction which is inherent in the NNC infrastructure, next I describe the two applied case studies, before I give some concluding remarks. The studies in question are, firstly, work in connection with a dictionary project carried out by a group of lexicographers in cooperation with the Norwegian Language Council, namely the new edition of Nynorsk ordliste (Nynorsk dictionary), published by the Nynorsk publisher Samlaget. Secondly, I describe work carried out for the dictionary project called Det Norske Akademis Store Ordbok (The Norwegian Academy's Comprehensive Dictionary)[2]; henceforth NAOB. The NNC has served as input for NAOB's normative work on determining alternative morphological forms of a variety of lexemes in the varieties of Norwegian called Bokmål and Riksmål[3].

[1] http://avis.uib.no/

[2] http://www.naob.no/

[3] Riksmål is an unofficial written standard which, like Bokmål, originates in the Dano-Norwegian written language, but unlike Bokmål it is largely based on the standardised orthography of 1917, hence it is the more conservative of the two varieties (http://snl.no/riksm%C3%A5l./ikke-offisiell_m%C3%A5lform).

2. The NNC's infrastructure for neology research

As shown by Andersen and Hofland (2012), the NNC project is especially tailored for lexicography and the study of new words in the language, and one of the most important features of the NNC architecture is a system for monitoring the development of lexical neologisms. Each day the lexical inventory of all the harvested text in the corpus is checked against a comprehensive, accumulated word list which consists of some 5.3 million word forms. This reference list comprises all the previously recorded words in the NNC combined with the inventory of all other Norwegian language resources collected at Uni Computing[4] over the last 20 years, including a full-form lexicon derived from the comprehensive dictionary Bokmålsordboka. Of the c. 230,000 running words that are daily added to the text database, on average 1,300 are previously unrecorded word forms. It should be pointed out that a 'word' is any sequence of graphemes (letters, digits, punctuation or other symbols) found between two spaces in a running text. Thus, a 'new word' is any word that is not included in the accumulated reference list mentioned. The daily lists of neologisms can be viewed at the *Nyord i norsk* 'Norwegian Neologisms' section of the NNC's web page. Naturally, only a subset of the new forms retrieved on a given day are relevant for lexicographical purposes. Since it is not feasible to check manually all word forms on a daily basis, the project applies pattern-matching and statistical techniques for selecting what appear to be the most relevant word forms from a lexicographical point of view. As an initial classification, new words are automatically distinguished according to some of their orthographical features (Andersen and Hofland 2012).

For the purpose of lexicography, it is primarily words that are orthographically unmarked that are of main interest. They contain no capital letter, hyphen or the like but consist of lower-case letters only. This accounts for almost half of the new words. Real neologisms, that is, new linguistically motivated and authentic lexical items are typically found in this category. But orthographically unmarked words could also be lower-case spelling errors not previously recognised. These are irrelevant from a lexicographer's point of view, but could be relevant to the developer of spell checking systems or to the psycholinguist focusing on error patterns or the like. Other new words have special orthographic features which make them less relevant for inclusion in dictionaries. About 10 per cent of the new words are productive, hyphenated compounds, which would normally not be of interest to the lexicographer, unless they achieve some opaque (non-transparent) lexicalised meaning, and if so, the use of a hyphen would be much less likely. A substantial proportion of new words, about 30 per cent, are orthographically distinguishable as name candidates (including hyphenated/compound names), and other forms have orthographic patterns that suggest that they are abbreviations, digits, URLs and e-mail addresses. Jointly, the neology extraction tool and the neology classifier contribute to making life easier for the lexicographer and neology researcher, as it provides efficient filtering that eases the task of looking for the "lexical needle" in the "corpus haystack".

3. Two case studies in applied lexicography

Case 1: Frequency profiling of neologisms for the dictionary Nynorsk ordliste

The dictionary Nynorsk ordliste (Hellevik et al. 2012) is a general-purpose and popular dictionary which is especially common in Norwegian schools, where it is used as the standard reference for students who use Nynorsk as either their main or secondary written variety of Norwegian. In 2012 the dictionary came in its eleventh edition by the Nynorsk publisher Samlaget, and this was the first edition after the comprehensive reform of Nynorsk morphosyntax which was done by the Norwegian Language Council during 2010-2012 and which became an official standard as of 1 August 2012 (Hovdenak 2012). This was also the first edition of the dictionary which systematically used a corpus-driven approach to neology when building its headword list (Atkins and Rundell 2008). The new edition was a cooperation

[4] and its organisational predecessors the HIT Centre, Unifob AKSIS and Uni Digital

between a group of lexicographers and the Norwegian Language Council; hence the neologisms that have been included in the published version have simultaneously become officially recognised as part of standard Norwegian. I was contacted by one of the lexicographers who wanted to have a maximally updated dictionary that contained neologisms as part of its lexical inventory. This section describes the methods used to provide this input to the lexicographic process.

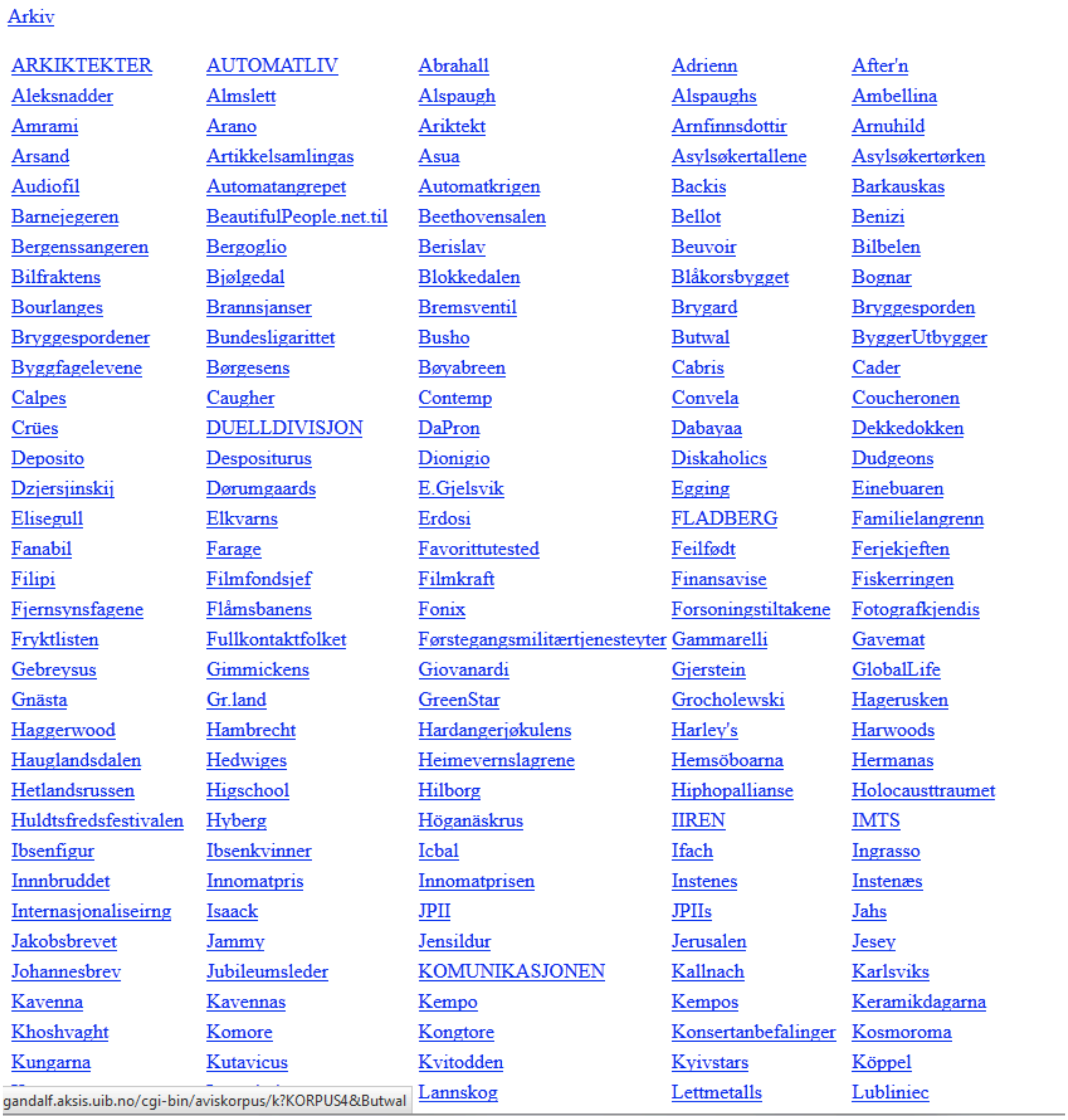
Naturally, the list of new word forms in the NNC is much too large to allow for any kind of manual inspection; the full neologism archive contains any previously unrecognised form that has occurred in any of the newspapers at least once during the corpus compilation period, i.e. from 1998 to the present. Therefore, the key to finding the most relevant words for inclusion in a dictionary is to consider frequency data. There could be several ways of calculating word frequencies, and ideally, one might wish to consider aspects such as the frequency profile of individual words, that is, its frequency development over time, its dispersion across different newspapers, the consistency of use of new words in various parts of the corpus, etc. However, given the time restriction of the commissioned task, and since the amount of neologisms to be included in this relatively short dictionary is limited, it was thought that a less sophisticated technique would be sufficient. It was decided that overall frequency statistics in the corpus as a whole would be enough to extract the most relevant word candidates that should be considered for inclusion by the lexicographers. In what follows I account briefly for the steps this task involved.

The starting point of the neology extraction was the archive of classified neologisms from the NNC (cf. section 2 and Andersen and Hofland 2012). For technical reasons, only neologisms recorded from 2005 onwards were included in the survey, i.e. over a period of eight years. This is because the classification tool was written in 2005 and only classified neologisms from that year to the present were available. If the full neologism archive from 1998 was to be included, one might consider extending the investigation to the whole period, either by classifying all the 1998-2004 words or by running the classifier on all unclassified neologisms from that period. This was not deemed necessary in the current project, since only a tiny fraction of all neologism candidates would eventually be included in the dictionary, and it was thought that the most relevant words would be highly recurrent in the data from 2005 onwards.

The classified data consist of a large set of html files which each contains the classified neologisms of a certain category on a given day. The relevant categories and rather self-explanatory file names are the following (cf. Andersen and Hofland 2012 for a full account):

```
anglicisms.html
contractions_and_inflections.html
digit_abbreviation.html
digit_compounds.html
digit.html
garbage.html
hyphenated_compound_acro_lex.html
hyphenated_compound_lex_acro.html
hyphenated_compound_lex_name.html
hyphenated_compound_name_lex.html
hyphenated_compound_name_name.html
hyphenated_compound_no_hyphen.html
hyphenated_compound_with_hyphen.html
multiwords_hyphenated.html
names.html
remaining_compounds_and_neologisms.html
urls_and_emails.html
```

To illustrate, Figure 1 shows the list of name candidates that were archived on the 18 April 2005.



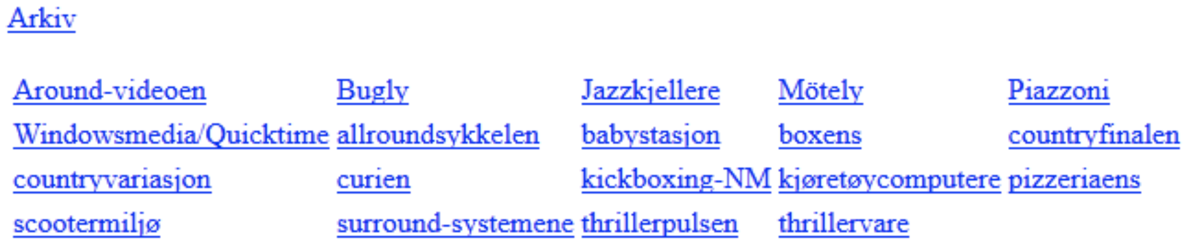
Arkiv

ARKIKTEKTER	AUTOMATLIV	Abrahall	Adrienn	After'n
Aleksnadder	Almslett	Alspaugh	Alspaughs	Ambellina
Amrami	Arano	Ariktekt	Arnfinnsdottir	Arnuhild
Arsand	Artikkelsamlingas	Asua	Asylsøkertallene	Asylsøkertørken
Audiofil	Automatangrepet	Automatkrigen	Backis	Barkauskas
Barnejegeren	BeautifulPeople.net.til	Beethovensalen	Bellot	Benizi
Bergenssangeren	Bergoglio	Berislav	Beuvoir	Bilbelen
Bilfraktens	Bjølgedal	Blokkedalen	Blåkorsbygget	Bognar
Bourlanges	Brannsjanser	Bremsventil	Brygard	Bryggesporden
Bryggespordener	Bundesligarittet	Busho	Butwal	ByggerUtbygger
Byggfagelevene	Børgesens	Bøyabreen	Cabris	Cader
Calpes	Caugher	Contemp	Convela	Coucheronen
Crües	DUELLDIVISJON	DaPron	Dabayaa	Dekkedokken
Deposito	Despositurus	Dionigio	Diskaholics	Dudgeons
Dzjersjinskij	Dørumgaards	E.Gjelsvik	Egging	Einebuaren
Elisegull	Elkvarns	Erdosi	FLADBERG	Familielangrenn
Fanabil	Farage	Favorittutested	Feilfødt	Ferjekjeften
Filipi	Filmfondsjef	Filmkraft	Finansavise	Fiskerringen
Fjernsynsfagene	Flåmsbanens	Fonix	Forsoningstiltakene	Fotografkjendis
Fryktlisten	Fullkontaktfolket	Førstegangsmilitærtjenesteyter	Gammarelli	Gavemat
Gebreysus	Gimmickens	Giovanardi	Gjerstein	GlobalLife
Gnästa	Gr.land	GreenStar	Grocholewski	Hagerusken
Haggerwood	Hambrecht	Hardangerjøkulens	Harley's	Harwoods
Hauglandsdalen	Hedwiges	Heimevernslagrene	Hemsöboarna	Hermanas
Hetlandsrussen	Higschool	Hilborg	Hiphopallianse	Holocausttraumet
Huldtsfredsfestivalen	Hyberg	Höganäskrus	IIREN	IMTS
Ibsenfigur	Ibsenkvinner	Icbal	Ifach	Ingrasso
Innnbruddet	Innomatpris	Innomatprisen	Instenes	Instenæs
Internasjonaliseirng	Isaack	JPII	JPIIs	Jahs
Jakobsbrevet	Jammy	Jensildur	Jerusalen	Jesey
Johannesbrev	Jubileumsleder	KOMUNIKASJONEN	Kallnach	Karlsviks
Kavenna	Kavennas	Kempo	Kempos	Keramikdagarna
Khoshvaght	Komore	Kongtore	Konsertanbefalinger	Kosmoroma
Kungarna	Kutavicus	Kvitodden	Kyivstars	Köppel
		Lannskog	Lettmetalls	Lubliniec

gandalf.aksis.uib.no/cgi-bin/aviskorpus/k?KORPUS4&Butwal

Figure 1. Name candidates archived on 18 April 2005

Similarly, Figure 2 gives a survey of the forms that have been singled out as anglicism candidates on the same day (Andersen 2005, 2012b).



Arkiv

Around-videoen	Bugly	Jazzkjellere	Mötely	Piazzoni
Windowsmedia/Quicktime	allroundsykkelen	babystasjon	boxens	countryfinalen
countryvariasjon	curien	kickboxing-NM	kjøretøycomputere	pizzeriaens
scootermiljø	surround-systemene	thrillerpulsen	thrillervare	

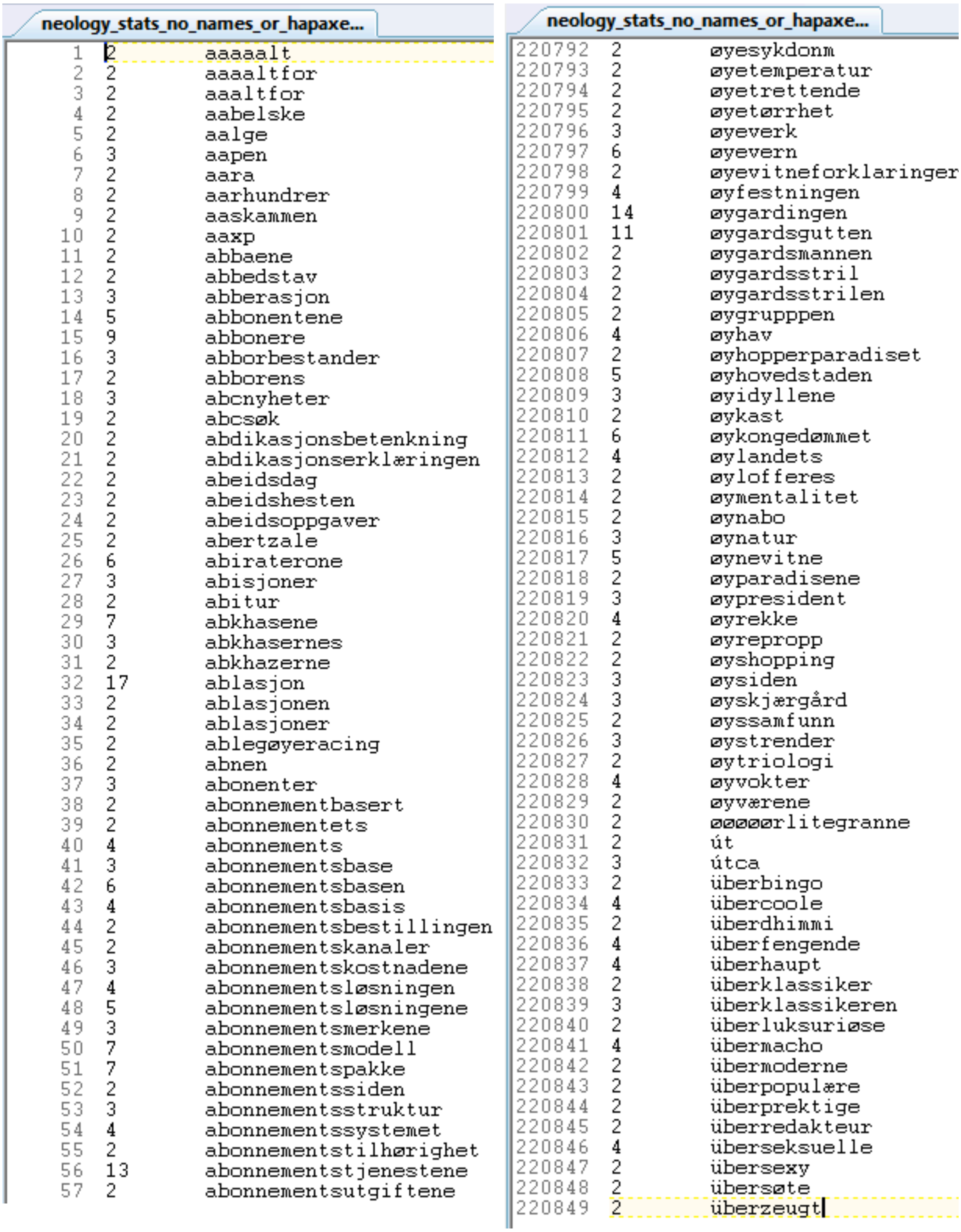
Figure 2. Anglicism candidates archived on 18 April 2005

Incidentally, these are the same files as the ones which appear in the neologism archive which is accessible to users of the corpus, and each word has a clickable link to the corpus location where a form is used. As can be seen, the neologism files merely establish the date of the first instance of a given form, but it contains no metadata on frequency or dispersion.

It is only two of the categories mentioned above that are considered applicable for lexical selection, namely `remaining_compounds_and_neologisms.html` and `anglicisms.html`. This is because they exclusively contain words that are orthographically unmarked, that is, they do not contain hyphens, digits, punctuation or control characters. A specifically written Perl script was used to traverse these lists and compare each form with a comprehensive word frequency list from the NNC containing 5.3 million unique word forms.[5] The script reads the non-hapax words and their frequencies, opens a catalogue of files, reads those files that are classified as neologism candidates (remaining words) or anglicism candidates and writes an alphabetically sorted list of neologism candidates and their frequencies to a single file. The output of this step was a file containing totally 897,131 entries. A problem arose in that the output file contained some (generally highly frequent) words that were clearly not neologisms, such as `646692 politiet` 'the police' and `1363 paraply` 'umbrella'. I was informed that this was due to a system error in the early stages of the project, which led to the inclusion of all words on a given day into the neologism archive. However, this was easily remedied by filtering the output list against the words which occur in the Scarrie lexicon, a multi-purpose full-form lexicon of Norwegian which is based on Bokmålsordboka (The Bokmål dictionary) and which is used as a reference in a variety of projects.[6] This step showed that the problem was rather marginal, as only 5,320 non-neologisms were found in the neologism archive, which amounts to a mere 0.6 per cent of the neologism candidates. The next step involved automatic filtering of the neologism candidates for hapax words and name candidates, containing an initial capital letter, as well as the manual filtering of some 20 forms that ought to have been removed by the previous filter but were not, since the Scarrie lexicon does not contain grammatical words, the genitive forms of lexical words or comparative and superlative forms of adverbs. These filtering steps reduced the dataset to 220,849 recurrent neologism candidates. For illustration, the first and last entries of this list are given in Figure 3.

[5] I am very grateful to Knut Hofland at Uni Computing for providing the neologism archive and frequency list. Again, a technical definition of 'word form' is understood here, i.e. any string that occurs between two spaces in the corpus.

[6] http://ling.b.uib.no/projects/scarrie/



neology_stats_no_names_or_hapaxe...

```
1   2         aaaaalt
2   2         aaaaltfor
3   2         aaaltfor
4   2         aabelske
5   2         aalge
6   3         aapen
7   2         aara
8   2         aarhundrer
9   2         aaskammen
10  2         aaxp
11  2         abbaene
12  2         abbedstav
13  3         abberasjon
14  5         abbonentene
15  9         abbonere
16  3         abborbestander
17  2         abborens
18  3         abcnyheter
19  2         abcsøk
20  2         abdikasjonsbetenkning
21  2         abdikasjonserklæringen
22  2         abeidsdag
23  2         abeidshesten
24  2         abeidsoppgaver
25  2         abertzale
26  6         abiraterone
27  3         abisjoner
28  2         abitur
29  7         abkhasene
30  3         abkhasernes
31  2         abkhazerne
32  17        ablasjon
33  2         ablasjonen
34  2         ablasjoner
35  2         ablegøyeracing
36  2         abnen
37  3         abonenter
38  2         abonnementbasert
39  2         abonnementets
40  4         abonnements
41  3         abonnementsbase
42  6         abonnementsbasen
43  4         abonnementsbasis
44  2         abonnementsbestillingen
45  2         abonnementskanaler
46  3         abonnementskostnadene
47  4         abonnementsløsningen
48  5         abonnementsløsningene
49  3         abonnementsmerkene
50  7         abonnementsmodell
51  7         abonnementspakke
52  2         abonnementssiden
53  3         abonnementsstruktur
54  4         abonnementssystemet
55  2         abonnementstilhørighet
56  13        abonnementstjenestene
57  2         abonnementsutgiftene
```

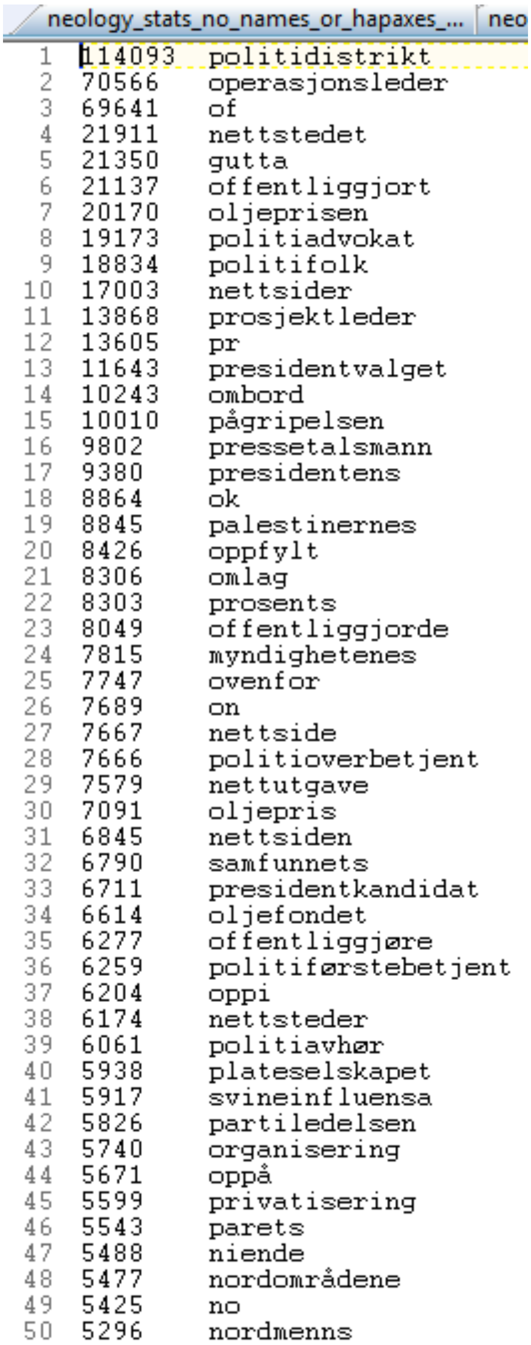
neology_stats_no_names_or_hapaxe...

```
220792  2         øyesykdonm
220793  2         øyetemperatur
220794  2         øyetrettende
220795  2         øyetørrhet
220796  3         øyeverk
220797  6         øyevern
220798  2         øyevitneforklaringer
220799  4         øyfestningen
220800  14        øygardingen
220801  11        øygardsgutten
220802  2         øygardsmannen
220803  2         øygardsstril
220804  2         øygardsstrilen
220805  2         øygrupppen
220806  4         øyhav
220807  2         øyhopperparadiset
220808  5         øyhovedstaden
220809  3         øyidyllene
220810  2         øykast
220811  6         øykongedømmet
220812  4         øylandets
220813  2         øylofferes
220814  2         øymentalitet
220815  2         øynabo
220816  3         øynatur
220817  5         øynevitne
220818  2         øyparadisene
220819  3         øypresident
220820  4         øyrekke
220821  2         øyrepropp
220822  2         øyshopping
220823  3         øysiden
220824  3         øyskjærgård
220825  2         øyssamfunn
220826  3         øystrender
220827  2         øytriologi
220828  4         øyvokter
220829  2         øyværene
220830  2         øøøøørlitegranne
220831  2         út
220832  3         útca
220833  2         überbingo
220834  4         übercoole
220835  2         überdhimmi
220836  4         überfengende
220837  4         überhaupt
220838  2         überklassiker
220839  3         überklassikeren
220840  2         überluksuriøse
220841  4         übermacho
220842  2         übermoderne
220843  2         überpopulære
220844  2         überprektige
220845  2         überredakteur
220846  4         überseksuelle
220847  2         übersexy
220848  2         übersøte
220849  2         überzeugt
```

Figure 3. Top and bottom entries of alphabetically sorted list of neologism candidates

Naturally, a flat list of some 220,000 words does not provide a very user-friendly resource for lexicographical work. Nevertheless it was deemed useful to include this list as part of the deliverable which was submitted to the lexicographers. The reason for this is that it gives a good survey of the productivity of individual forms, which is part of what the lexicographer needs to consider when deciding to include a particular word or not. As is salient in Figure 3, the word *abonnement* 'subscription' is a highly productive leftmost compound component, and the prefix *über-*, which originates in German, has come to be highly productive also in Norwegian, presumably via influence from English (Renouf 2007b). The extract includes several forms which are genuine neologisms and which should be considered for inclusion in dictionaries, such as *abitur*, *abkhasene* 'the Abkhazians', *ablasjon* 'ablation', etc. But this comprehensive list also includes a great many forms which are lexicographically irrelevant, such as transparent compounds like *øyfestningen* 'the island fortification', spontaneously creative spellings with elongated vowels like *aaaaltfor*, equivalent to 'faaaar too much', recurrent misspellings such as *abbonere* (*abonnere*) 'subscribe', and names and acronyms which happen to have been written without any capital letter, such as *aaxp*.

The final step involved the frequency-sorting of the list of neologism candidates. The top of the frequency list is shown in Figure 4.



```
114093  politidistrikt
70566   operasjonsleder
69641   of
21911   nettstedet
21350   gutta
21137   offentliggjort
20170   oljeprisen
19173   politiadvokat
18834   politifolk
17003   nettsider
13868   prosjektleder
13605   pr
11643   presidentvalget
10243   ombord
10010   pågripelsen
9802    pressetalsmann
9380    presidentens
8864    ok
8845    palestinernes
8426    oppfylt
8306    omlag
8303    prosents
8049    offentliggjorde
7815    myndighetenes
7747    ovenfor
7689    on
7667    nettside
7666    politioverbetjent
7579    nettutgave
7091    oljepris
6845    nettsiden
6790    samfunnets
6711    presidentkandidat
6614    oljefondet
6277    offentliggjøre
6259    politiførstebetjent
6204    oppi
6174    nettsteder
6061    politiavhør
5938    plateselskapet
5917    svineinfluensa
5826    partiledelsen
5740    organisering
5671    oppå
5599    privatisering
5543    parets
5488    niende
5477    nordområdene
5425    no
5296    nordmenns
```

Figure 4. Top and bottom entries of alphabetically sorted list of neologism candidates

Importantly, the listed words are neologism *candidates*, that is, there are many words which, albeit frequent, are not lexicographically relevant. This is usually either because they are fully transparent compounds, which do not belong in a dictionary due to their lack of lexicalization or idiomaticity (Atkins and Rundell 2008: 169ff), or because they are occasionalisms, i.e. words which are only pertaining to a certain news story or societal issue of limited duration.

In order to ease the lexicographers' work further, I split the frequency-sorted list into five files according to varying frequency thresholds of 10,000+ / 1,000+ / 100+ / 10+ / 2+ tokens. I also produced an accompanying comment file, in which examples of words extracted at the various frequency thresholds have been illustrated. This is reproduced here as Table 1.

Table 1

Survey of frequency thresholds for neologism candidates

Neologism frequency range	Words	File	Examples of neologisms from file
n ≥ 10,000	15	neology_stats_frq_10000_plus	*nettstedet, pr, nettsiden*
9,999 ≥ n ≥ 1,000	414	neology_stats_frq_1000_plus	*miljøkriminalitet, pressetalsmann, ok, venstreback, tastetrykk*
999 ≥ n ≥ 100	1,662	neology_stats_frq_100_plus	*halalmat, vuvuzelaene, remix, subprime, simkort*
99 ≥ n ≥ 10	8,819	neology_stats_frq_10_plus	*retusjering, medmor, serieforbryter, kitschy, blokkeringsfrie, eierskapsutøvelse, politihijab*
9 ≥ n ≥ 2	209,939	neology_stats_frq_2_plus	*vigselsliturgi, surfehastighet, surrogatfamilier, piggskate, polyamori, nyverdi*
TOTAL	220,849		

The words in the rightmost column are examples of words which are either genuine neologisms that the lexicographer should consider for inclusion, such as *nettstedet* and *nettsiden* 'the website', *tastetrykk* 'key stroke', *halalmat* 'halal food', *simkort* 'sim card', *subprime*, *medmor* 'co-mother', *polyamori* ' polyamory', etc., or words which are commonly used but have not been standardised in earlier general language dictionaries, such as the abbreviation *pr* 'per', the discourse marker *ok*, the compound *pressetalsmann* 'press officer', the fish name *piggskate* 'thornback ray', etc. Note also that the frequency lists are unlemmatised; thus *nettside* 'webpage' and *nettsiden* 'the webpage' will appear as two different entries. In sum, the deliverable for this subproject was the text file that consisted of some 220,000 neologism candidates, the frequency-sorted files and the comment file. The files sorted for frequency threshold give the most relevant place for lexicographers to start looking for new words to include in their headword list, and, according to the lexicographer who requested these data, the effort has provided highly valuable input for their manual lexicographical work.

Case 2: Investigating morphological variability for NAOB

The second case to be illustrated here also concerns applied research for lexicographical purposes and standardisation. However, unlike Case 1 above, the data were provided for a project for standardisation of the Riksmål and Bokmål varieties of Norwegian language. NAOB is a comprehensive dictionary project under the auspices of Det Norske Akademi for Sprog og Litteratur (The Norwegian Academy of Language and Literature) with funding from the Norwegian government. Its content is based on the existing Norsk Riksmålsordbok (Norwegian Riksmål Dictionary), but considerable new content is being added, and the existing content will be modernised and updated before its completion in 2017. It was in this connection that the NAOB lexicographers contacted the NNC project, as they wanted to use the corpus as a source of information about the degree of use of a range of different word forms with alternative spelling and morphology. In other words, it was not neology as such that was the concern here, but morphological and orthographical variation of already registered words.

Initially I was given two lists of words that the lexicographers wanted to use in this corpus-based investigation. The purpose was to assess the relevance of maintaining a standardised formal variation realised as two equally valid alternative forms in the new version of the dictionary. This included a range of forms that can have two alternative realisations in the stem. The first list contained words where the variation usually pertains to a 'conservative' and a 'moderate' or in some cases 'radical' alternative, as illustrated by the word pairs given in Table 2.

Table 2
Morphological variants in NAOB: conservative or moderate/radical stem

Variant 1 (conservative)	Variant 2 (moderate/radical)	English translation
ekenøtt	*eikenøtt*	acorn
gjetost	*geitost*	goat cheese
hjem	*heim*	home
hård	*hard*	hard
høk	*hauk*	hawk
peppermø	*peppermøy*	spinster
sne	*snø*	snow
iskold	*iskald*	ice cold
sorte får	*svarte får*	black sheep
tyve	*tjue*	twenty

The second list contained words where the alternative variants were due to the choice between an original or adapted spelling of foreign words, including some anglicisms (Andersen 2012a), or the variable representation of certain foreign phonemes such as the Greek diphthong *eu-*/ *ev-*. This is illustrated by the word pairs in Table 3.

Table 3

Morphological variants in NAOB: conservative or moderate/radical stem

Variant 1 (original spelling)	Variant 2 (adapted spelling)	English translation
abacus	*abakus*	abacus
alsacer	*alsaser*	Alsatian
apache	*apasje*	apache
baguette	*bagett*	baguette
blitz	*blits*	blitz
coda	*koda*	coda
eufemisme	*evfemisme*	euphemism
foyer	*foajé*	foyer
ghetto	*getto*	ghetto
hermeneutikk	*hermenevtikk*	hermeneutics
neutral	*nøytral*	neutral
rajah	*raja*	rajah
Talmud	*talmud*	Talmud
ton	*tonn*	ton
yoruba	*joruba*	Yoruba

So, the commissioned task was to provide reliable usage statistics in the NNC of each of the forms listed in two long lists. In effect, this amounted to finding an efficient way of searching for many words in one go, grouping and systematising the results and presenting statistics to the lexicographers. Note that some of the requested items were multiword expressions, such as *i bet for/i beit for* 'lacking', which required some additional processing.

The first step was to convert the word lists into machine-readable text files while maintaining the distinction between different categories of variation inherent in Tables 2–3. The inventory of forms to be investigated contained 436 word forms of the type illustrated in Table 2 and 1,032 word forms of the type in Table 3, totally 1,468 entries, of which 87 were multiword units.

One of the substantial advances of the NNC project is the development of the new search system and user interface called Corpuscle[7] (Andersen and Hofland 2012; Meurer 2012), which turns out to be ideal for this purpose. Besides its ability to handle very large amounts of data speedily, one of the main advantages is that the interface allows for regular expression-based searches of multiple and truncated word forms, and to have the result presented in an easily downloadable concordance or word list format with usage statistics for each retrieved form. At

[7] http://iness.uib.no/korpuskel/main-page

the time of investigation, the Corpuscle interface to the NNC searched in a little short of one billion words of newspaper text covering the period 1998-2009. By means of a Perl script the list of word forms was converted to a Corpuscle-compatible regular expression which can be seen below (shortened as shown by "…", which is not part of the search expression):

```
[word="(anstøtsstein|anstøtssten|bautastein|bautasten|benhard|benhardt|benh
ård|benhårdt|bleikeplass|blekeplass|blesand|blesbukk|blesgås|bleshøne|bless
|bliss|blissand|blissbukk|blissgås|blisshøne|blotstein|blotsten|bredslede|b
reislede|brostein|brosten|brynestein|brynesten|eikenøtt|eikeskog|eiketre|ei
ketresmøbel|einerbusk|einerbær|ekenøtt|ekeskog|…|værhard|værhård).*"]
```

As shown in the example, I consistently used right-truncated searches for this purpose, in order to investigate not merely the base form of the listed words but also their inflectional forms and compounding. Due to time restrictions, left-truncated searches were not performed (these are also assumed to give much fewer hits than right-truncations). Exact (un-truncated) searches were used to provide usage statistics for the 87 multiwords.

A series of similar searches as the one above were performed, and the results were saved as concordance lists. The efficiency of the search engine can be illustrated by the fact that it was unproblematic to search for as many as 387 right-truncated words in this 1 billion word corpus, which yielded a concordance list of 277,873 lines. Another search retrieved more than 2.2 million concordance lines, but it turned out that a few very short words had to be removed from this list and searched individually, as they gave rise to many irrelevant hits and hence low precision, notably *mø*, *gem*, *gir*, *hiv*, *kol*, *spe*, *ton* and *ufo*. This operation shortened the results of the search to a more manageable size. The output concordance lists were needed as a reference file in order to check the relevance of individual word forms (some, but not all; cf. below), and they were also part of the deliverable to the group of NAOB lexicographers. There was a need to write a Perl script to convert the concordance lists into a more manageable format in order to calculate usage statistics. The output file of this was simply a list of all the retrieved word forms and their frequencies, as illustrated by Figure 5, which gives all the retrieved types containing the form *aksent.**.

```
aksent             1059
aksent-grepet      1
aksent-syndromet   1
aksentbruk         1
aksenten           146
aksentene          14
aksenter           51
aksentfarge        2
aksentfarger       1
aksentfri          4
aksentfrie         1
aksentfritt        4
aksentkritikk      1
aksentpreget       2
aksentrik          1
aksentskifte       1
aksenttegn         4
aksentuere         16
aksentuerer        30
aksentueres        20
aksentuering       7
aksentueringen     3
aksentueringer     2
aksentuert         38
aksentuerte        6
```

Figure 5. Word forms containing the form *aksent.**

The next and most time-consuming stage of this project amounted to the manual inspection and lemmatisation of these lists. Altogether 18,609 word forms had to be checked and grouped according to the lemma and word pair in question. The need for this manual check is due to the inherent ambiguity of many word forms. This can be illustrated with reference to the word pair *trusel*/*trussel* 'threat', which yielded the following hits, among others:

```
truselen           204
truselene          7
truseler           1
truselforsikringer 1
*truselignende     1
*truselinningen    5
truselkategori     1
truselnivået       1
*truselogo         1
```

Figure 6. Extract of word forms containing the form *trusel.**

The forms marked by an asterisk are not relevant to the word pair in question, because they stem from other lemmas; the forms *truselignende*, *truselinningen* and *truselogo* are all compounds that contain the word *truse* 'panties' as their leftmost component.[8] I supplied the data with a comment field, which was used whenever relevant, in order to give the lexicographers the opportunity to alter decisions I had made regarding which forms to include and which to ignore, and redo the statistics. A pertinent example could be the comment regarding the form *blitz*

> valgte å ikke fjerne noe her, men enkelte kunne utgått da de vel har å gjøre med Kafe Blitz, dessuten er jeg usikker på om variasjon er mulig når det er snakk om blitz-krig
>
> I chose not to remove anything here, but some forms could have been deleted since they probably relate to Kafe Blitz, besides I am uncertain as to whether variation is possible in the word blitz war

In other words, the comment field information could be a useful resource for the subsequent quality-assurance of the data and statistics. The comments generally concerned the removal of irrelevant forms, such as name tokens and forms stemming from other lexemes than the targeted word pair, like *cigarette* and *facsimile* 'facsimile', which were not considered relevant for the variability of the word pairs *cigar/sigar* and *fax/faks*, respectively. In a few cases, I also commented on the need for further quality control by means of the manual inspection of the concordance lists for individual forms. A relevant example is the word *game*, which in most cases is used in the sense 'game', which is not relevant to the adjectival word pair *game/gem* 'pleasant, nice, sporty' (about person), and therefore in need of further inspection at the level of individual tokens than my time allowed for. Similarly, I reported the need for manual removal of irrelevant verb forms for the form *spe* 'dilute, thin', which are not relevant to the adjectival word pair *spe/sped* 'tiny, delicate, feeble', and I suspect that many of the forms of a word such as *force* represent use in the context of code switching into English or the multiword *force majeure* and are therefore not relevant for the variability of the nominal pair *force/forse* 'strength'. Other issues dealt with my decisions to include or exclude non-standardised forms

[8] In checking the individual word forms, I decided not to exclude forms which contained various idiosyncratic formatting errors, such as *trussel&quot*, but decided to keep these as valid tokens, provided that the form could unambiguously be assigned to one variant or the other of a relevant word pair.

such as *jigg* for the word pair *jig/gigg* (which was considered relevant and hence included), and more generally how ambiguous forms had been dealt with.

The final step of this investigation involved the calculation of percentages showing the degree of occurrence of forms relating to each word pair and presenting results both as a list and as graphical output, as seen from Figure 7.

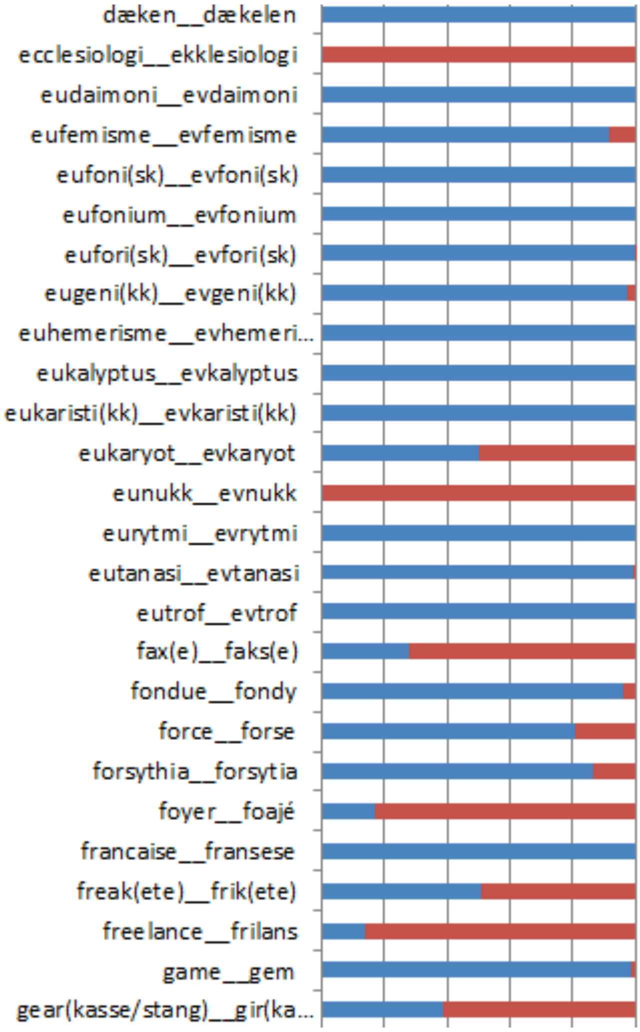


Figure 7. Extract of the statistical output of Case 2

As is seen from the figure, it varies considerably whether one or the other form is used; for instance, the spelling *eu-* for the Greek diphthong is clearly dominating in contemporary Norwegian, with the notable exception of the word *evnukk* 'eunuch', while users prefer the adapted orthography in words such as *faks*, *foajé* and *frilans*.

3. Concluding remarks

This study has reported on two individual projects within corpus lexicography that have been made possible due to the development of the Norwegian Newspaper Corpus. In many ways, the two studies illustrate the distinction between a *corpus-driven* and a *corpus-based* approach and their relevance for ongoing dictionary development. The first case, the Nynorsk ordliste project, relied on a corpus-driven approach, in which the corpus data were used in an inductive, bottom-up fashion as a basis for determining which neologisms to include in the new edition of this Nynorsk dictionary. According to the lexicographer, this was a highly appreciated effort which simplified their work and a clear improvement compared to earlier manual registration of neologisms. The second case, the NAOB project, relied on a corpus-based approach, in which a large set of variable realisations of already registered words were used as a basis for retrieving corpus tokens. These data formed the basis for the subsequent evaluation of the relevance of maintaining orthographic variants as officially recognised in the revised edition of the comprehensive NAOB dictionary for Bokmål and Riksmål.

Thus, the studies have shown that, by relatively simple means, it is possible to retrieve large amounts of pertinent data which may highlight observable usage patterns and systematically assist lexicographers in their editorial choices. However, it should also be added that the methods proposed may be much improved by the application of more sophisticated analytical approaches. Examples of this may be the use of frequency profiles of individual words and dispersion data across the various sections of the corpus rather than just overall frequencies, the use of lemmatised rather than unlemmatised word lists, and the use of collocational statistics as a means for weeding out unwanted tokens representing code switching, multiwords, or other contextually constrained uses where formal variation is not possible. Such approaches could also add valuable sociolinguistic information that could inform the analysis and further reduce the need for manual work on the part of the corpus linguist or the lexicographer.

References

Andersen, Gisle (2005), 'Assessing algorithms for automatic extraction of anglicisms in Norwegian texts', *Proceedings from Corpus Linguistics 2005,* 1. <http://www.corpus.bham.ac.uk/pclc/Birmingham_paper.doc>, accessed 2005.

Andersen, Gisle (2010), 'Halvautomatisk ekserpering av anglisismer i norsk', *Nordiska studier i leixcografi,* 10, 72-85.

Andersen, Gisle (2011a), 'Corpora as lexicographical basis: the case of anglicisms in Norwegian', *VARIENG - Studies in Variation, Contacts and Change in English,* 2011 (6). <http://www.helsinki.fi/varieng/journal/index.html>.

Andersen, Gisle (2011b), 'Finisj eller finish? Norvagisering femten år etter normeringsvedtaket', *Språknytt,* (1), 27-29.

Andersen, Gisle (2012a), 'A corpus-based study of the adaptation of English import words in Norwegian', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 157-92.

Andersen, Gisle (2012b), 'Semi-automatic approaches to Anglicism detection in Norwegian corpus data', in Cristiano Furiassi, Virginia Pulcini, and Félix Rodríguez Gonzáles (eds.), *The Anglicization of European Lexis* (Amsterdam: John Benjamins), 111-30.

Andersen, Gisle and Hofland, Knut (2012), 'Building a large monitor corpus based on newspapers on the web', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 1-30.

Andersen, Gisle (ed.), (2012c), *Exploring Newspaper Language - Using the web to create and investgate a large corpus of modern Norwegian* (Amsterdam: John Benjamins) 1-30.

Atkins, B. T. Sue and Rundell, Michael (2008), *The Oxford guide to practical lexicography* (Oxford: Oxford University Press) XII, 540 s.

De Smedt, Koenraad (2012), 'Ash compound frenzy: A case study in the Norwegian Newspaper Corpus', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create* *and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 241-56.

Dyvik, Helge (2012), 'Norm clusters in written Norwegian', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 193-220.

Fjeld, Ruth Vatvedt and Nygaard, Lars (2012), 'Lexical neography in modern Norwegian', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 221-40.

Fletcher, William H. (2007), 'Concordancing the web: promise and problems, tools and techniques', in Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer (eds.), *Corpus Linguistics and the Web* (Amsterdam/New York: John Benjamins), 25-45.

Grefenstette, Gregory (2002), 'The WWW as a resource for lexicography', in Marie-Hélène Corréard (ed.), *Lexicography and natural language processing* (Gothenburg: Euralex).

Halverson, Sandra (2012), 'Metonymic extension and vagueness: *Schengen* and *Kyoto* in Norwegian newspaper language', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 286-306.

Hellevik, Alf, Søyland, Aud, and Rauset, Margunn (2012), *Nynorsk ordliste* (Oslo: Samlaget).

Hofland, Knut (2000), 'A self-expanding corpus based on newspapers on the Web', *The Second International Language Resources and Evaluation Conference (LREC)* (Paris: European Language Resources Association (ELRA)).

Hovdenak, Marit (2012), 'Nynorsknorma - slik blir ho', *Språknytt,* (3), 18-21.

Hundt, Marianne, Biewer, Carolin, and Nesselhauf, Nadja (2007), *Corpus linguistics and the web* (Language and computers; Amsterdam: Rodopi) VI, 305 s. /.

Kilgarriff, Adam and Tugwell, David (2002), 'Sketching words', in Marie-Hélène Corréard (ed.), *Lexicography and Natural Language Processing* (Gothenburg: EURALEX).

Kilgarriff, Adam and Grefenstette, Gregory (2003), 'Introduction to the Special Issue on Web as Corpus', *Computational Linguistics,* 29 (3), 1-15.

Kristiansen, Marita (2012), 'Financial jargon in a general newspaper corpus', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (John Benjamins), 257-84.

Kristiansen, Marita and Andersen, Gisle (2012), 'Corpus approaches to terminology and their relevance for dynamic domains', *Neologica,* 13, 43-62.

Meurer, Paul (2012), 'Corpuscle – a new corpus management platform for annotated corpora', in Gisle Andersen (ed.), *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian* (Amsterdam: John Benjamins), 31-50.

Ooi, Vincent B. Y. (1998), *Computer corpus lexicography* (Edinburgh: Edinburgh University Press) X, 243 s.

Pulcini, Virginia (2008), 'Corpora and lexicography: the case of a dictionary of Anglicisms', in Aurelia Martelli and Virginia Pulcini (eds.), *Investigating English with corpora : studies in honour of Maria Teresa Prat* (Monza: Polimetrica), 189-203.

Renouf, Antoinette (2007a), 'Corpus development 25 years on: from super-corpus to cyber-corpus', in Roberta Facchinetti (ed.), *Corpus linguistics 25 years on* (Amsterdam/New York: Rodopi).

Renouf, Antoinette (2007b), 'Tracing lexical productivity and creativity in the British Media', in Judith Munat (ed.), *Lexical creativity, texts and contexts* (Amsterdam: John Benjamins), 61-90.

Sinclair, John McH. (ed.), (1987), *Looking up* (London/Glasgow: Collins ELT).