Exploring Newspaper Language

Using the web to create and investigate a large corpus of modern Norwegian

Edited by Gisle Andersen

Studies in Corpus Linguistics 49

JOHN BENJAMINS PUBLISHING COMPANY

Exploring Newspaper Language

Using the web to create and investigate a large corpus of modern Norwegian

Edited by

Gisle Andersen

Norwegian School of Economics, Bergen

John Benjamins Publishing Company

Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

Exploring newspaper language : using the web to create and investigate a large corpus of modern Norwegian / edited by Gisle Andersen.
p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 49)
Includes bibliographical references and index.
1. Norwegian language (Nynorsk)--Usage. 2. Norwegian language (Nynorsk)--Syntax. 3. Newspapers--Norway. 4. Mass media--Norway. 5. Information technology--Norway. I. Andersen, Gisle.
PD2914E97 2012
439.8'20188--dc23 2011045662
ISBN 978 90 272 0354 0 (Hb ; alk. paper)
ISBN 978 90 272 7499 1 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Exploring Newspaper Language

Studies in Corpus Linguistics (SCL)

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see *http/benjamins.com/catalog/scl*

Volume 49

Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian
Edited by Gisle Andersen

Table of contents

Building a large corpus based on newspapers from the web*

Gisle Andersen & Knut Hofland
NHH Norwegian School of Economics / Uni Computing

The Norwegian Newspaper Corpus (NNC) is an initiative to create a large monitor corpus representing contemporary Norwegian language in both its written varieties, Bokmål and Nynorsk. The corpus is compiled through daily harvesting and processing of published texts from the web edition of Norwegian newspapers. This introductory chapter gives a survey of work on corpus building, tool development and research in connection with the NNC project. It provides an overview of the corpus and its system architecture, describing the work flow, tools and methods used in the data processing. The chapter also gives a presentation of the individual research contributions to this volume.

1. Introduction

The Norwegian Newspaper Corpus (NNC) is an undertaking which started in 1998 with the set-up of an automatic system for extracting Norwegian newspaper text from the web (Wangensteen 2002; Hofland 2000; Andersen 2005, 2010). There were originally 10 newspapers in the corpus, but from that point this effort has been extended in various ways, in terms of the number of newspapers retrieved (now 24), the technologies applied, the range of topics researched, the number of users, etc. This book serves the twofold purpose of describing the procedures, methods and tools developed in the NNC project, as well as presenting new research based on the corpus. This introductory chapter first provides an overview of the corpus and its system architecture (Section 2) and then gives a brief presentation of the individual contributions to this volume (Section 3).

* We are grateful to a number of institutions for their contributions to the Norwegian Newspaper Corpus project, especially the Research Council of Norway (the AVIT programme), the University of Bergen (Faculty of Humanities), NHH Norwegian School of Economics and Uni Research.

In corpus linguistics there has been a tremendous development of new language resources, technologies and methods over the past few decades. We consider the emergence of dynamic corpora and web-based corpora to be among the most significant events in recent corpus development (Fairon et al. 2007; Hundt, Nesselhauf & Biewer 2007; Renouf 2007). Equally crucial is the exploitation of large corpora in lexicography and terminology (Grefenstette 2002; Atkins & Rundell 2008; Pulcini 2008). The Norwegian Newspaper Corpus project is both inspired and directly influenced by these developments and represents the first initative to create a large monitor corpus of contemporary Norwegian language in both its written varieties, Bokmål and Nynorsk. The NNC is compiled through daily harvesting and processing of published texts from the web edition of Norwegian newspapers. It is a 'modern diachronic corpus', in the sense described by Renouf (2007), enabling the study of language change, neologistic usage and lexical productivity and creativity as it unfolds in written language. This open-ended monitor corpus can be used to study language as an evolving phenomenon at the diachronic micro-level, by comparing time frames calibrated on a daily, weekly, monthly or yearly basis within the time-span the corpus represents – from 1998 to the present. This makes it particularly well suited for studying ongoing innovation in contemporary language, such as lexical, morphological, syntactic and pragmatic innovation (Hundt, Nesselhauf & Biewer 2007: 3). Naturally, it can also be used as a comparison with other synchronic or diachronic Norwegian corpora or as a basis for cross-linguistic comparison with similar corpora representing other languages.

As the contributions to this book will reveal, the corpus has been used as the basis for a variety of case studies. However, we should emphasise that the corpus is particularly tailored for lexicographical work (Hofland 2000; Wangensteen 2002; Andersen 2005, 2010). An important source of inspiration has been the Collins COBUILD project and subsequent corpus development projects using a similar methodology. Initiated by Atkins and Sinclair in the 1970s, this was the first lexicography initiative which systematically used a corpus as its main source of knowledge about words and their use in the language (Sinclair 1987). This effort 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 & Tugwell 2002: 125). An important driving force behind our efforts has therefore been to enable systematic corpus resources for the study of lexical neography in Norwegian (Fjeld & Nygaard this volume). Indeed the project team has consisted of lexicographers who utilise the corpus for this purpose and several of the language processing tools to be described in Section 2 are meant to make life easier for the lexicographer

and, more generally, to stimulate research in lexicology, terminology, morphology and related fields. Until the compilation of our corpus, Norwegian lexicography had been largely based on manual extraction of new words, but now it is possible to use the method of corpus lexicography "for measuring hard evidence of the lexical behaviour of words since this will (arguably) result in a more representative, coherent and consistent output than a lexicon produced from conventional means" (Ooi 1998: 37).

We also draw on methods used in the AVIATOR project, which developed "the first 'dynamic' corpus of unbroken chronological text" (Renouf 2007: 36). This was established in 1990 using the *Times* newspaper as its source. The efforts of the Birmingham-based RDUES group were followed by other efforts to monitor and classify new words in the APRIL and ACRONYM projects (Renouf 1996, 2007). Since the turn of the millennium, it has also become increasingly common to develop and explore web-based corpora, aka. 'cyber-corpora' (Renouf 2007), resulting in a growing body of corpus-based studies using the web as their prime source of data (Kilgarriff & Grefenstette 2003; Fletcher 2007; Hundt, Nesselhauf & Biewer 2007). Corpus-building efforts such as WebCorp (Renouf, Kehoe & Banerjee 2005), the Corpus of Contemporary American English (Davies 2009) and the WaCky initiative (Baroni et al. 2009) use web crawler technology to make the web serve as a linguistic corpus. In line with this approach, we regard the World Wide Web as "a mine of language data of unprecedented richness and ease of access" (Lüdeling, Evert & Baroni 2007: 7) and our overall aim has been to exploit the web as a language resource for the compilation of a large corpus.

Further, we consider newspapers to be a relevant source on which to base a large corpus of Norwegian. Web newspapers lend themselves to use in a corpus for a number of reasons, such as their ease of access, their large and regular text production, their wide range of topics, their systematicity and coherence in text categorisation, their coverage of local, regional, national and global events, etc. (van Dijk 1988; Cotter 2010). Unlike many other sections of the World Wide Web, newspapers contain journalistic text which is professionally written and edited. This entails a certain level of quality and standardisation in terms of spelling, language use, genre conventions and other features. The initiation of a project of compiling a web-based corpus might nevertheless raise the question "Why not just use Google?" In response to this we would argue, like Kilgarriff and Grefenstette (2003: 12) that for most scientific purposes a systematically compiled and annotated corpus is preferable to relying on Google searches for a number of reasons. First, a corpus such as the NNC enables the extraction of all instances of a token within a set period of time and within a certain genre, in our case web newspapers. Second, it allows for the presentation of more context than Google gives, and for the restriction of the search to predefined categories such as newspaper type,

date of publication, type of article, author's gender, etc. Third, a structured corpus allows the researcher to sort and select tokens according to transparent linguistic criteria, such as word form, lemma, word class, collocational features and the like. Fourth, it serves as a basis for extracting reliable usage and distribution statistics according to linguistic and non-linguistic criteria such as the ones mentioned.[1]

Note that the NNC is a *web-based* newspaper corpus. We have restricted the inclusion of newspapers to the web edition of newspapers that also have a *printed counterpart*. This means that fully web-based media, such as the web newspaper *Nettavisen* or the web news service of NRK, the Norwegian national broadcasting corporation, are not included. Given this restriction, we have aimed at a maximally broad selection of regionally and distributionally different newspapers, and the collection includes local, regional, national and special interest (niche) newspapers. Table 1 gives a survey of the newspapers included in the corpus and their sizes (measured in total number of words in the corpus) as of April 2011.

Table 1. Survey of newspapers in the Norwegian Newspaper Corpus

Newspaper	Code	Distribution	Category	Written standard	Million words	%
1. Adresseavisen*	AA	Regional – middle	General	BM	100.32	9.3%
2. Aftenposten*	AP	National	General	BM	136.62	12.7%
3. Bergens Tidende*	BT	Regional – west	General	BM/NN	58.35	5.4%
4. Dagsavisen*	DA	National	General	BM	46.23	4.3%
5. Dagbladet*	DB	National	General	BM	128.71	12.0%
6. Dagens Næringsliv*	DN	National	Financial	BM	59.404	5.5%
7. Dag og tid	DT	National	Niche	NN	4.389	0.4%
8. Firda	FI	Local – west	General	BM/NN	16.29	1.5%
9. Fædrelandsvennen*	FV	Regional – south	General	BM	101.78	9.5%
10. Gudbrandsdølen Dagningen	GD	Local – east	General	BM/NN	11	1.0%
11. Hallingdølen	HD	Local – east	General	BM/NN	8.79	0.8%
12. Hordaland	HO	Local – west	General	NN	0.88	0.1%

(*Continued*)

1. Most of these objections can also be raised against using a commercial text archive such as Atekst, which, although large and useful, is not tailored for corpus linguistic research (like the WWW itself).

Table 1. (Continued)

Newspaper	Code	Distribution	Category	Written standard	Million words	%
13. Klassekampen	KK	National	Niche	BM/NN	8.18	0.8%
14. Morgenbladet	MB	National	Niche	BM/NN	25.6	2.4%
15. Nationen	NA	National	Niche	BM/NN	20.06	1.9%
16. Nordlys*	NL	Regional – north	General	BM	55.86	5.2%
17. Stavanger Aftenblad*	SA	Regional – southwest	General	BM/NN	101.31	9.4%
18. Sunnhordland	SH	Local – southwest	General	BM/NN	3.99	0.4%
19. Sogn avis	SO	Local – west	General	NN	11.8	1.1%
20. Sunnmørsposten	SP	Local – northwest	General	BM/NN	30.95	2.9%
21. Vikebladet	VB	Local – northwest	General	NN	2.43	0.2%
22. Verdens Gang*	VG	National	General	BM	130.35	12.1%
23. Vårt land	VL	National	Niche	BM	6.8	0.6%
24. Vest-Telemark blad	VT	Local – east	General	NN	2.9	0.3%

The asterisk indicates that the newspaper is among the original ten newspapers that were included from the start of the harvesting in 1998, while the other newspapers are later additions; hence their text material does not necessarily cover the whole period (cf. Section 2). Note that some national newspapers such as *Aftenposten* also serve regional functions in their respective areas (Oslo/Central Eastern Norway in this case). The categorisation in Table 1 follows the official classification used by the Norwegian Ministry of Culture in their calculation of the annual public funding for Norwegian newspapers, which distinguishes between *lokalaviser* 'local newspapers', *regionaviser* 'regional newspapers', *riksaviser* 'national newspapers' and *nisjeaviser* 'niche newspapers'. The niche newspapers are also known as *rikskringkastede meningsbærende aviser* 'nationally distributed opinion-making newspapers'. The ones included in the corpus are: *Klassekampen*, the newspaper of the political left; *Dag og Tid*, a weekly Nynorsk newspaper for a academic readership; *Morgenbladet*, a weekly newspaper primarily for an academic readership; *Nationen*, a newspaper representing agro-pastoral/agricultural communities and readers; and *Vårt Land*, a newspaper for a Christian readership. Our policy in compiling the corpus has been to be *maximally inclusive* and to achieve a balance between different types of newspapers. And, indeed, the corpus comprises all the national newspapers; all major regional

newspapers from Norway's largest cities (Oslo, Bergen, Trondheim, Stavanger, Tromsø, Kristiansand); the five most important opinion-making/niche newspapers, each with its own relatively well-defined readership; as well as a number of regional newspapers representing most parts of the country, though not all due to time restrictions of the NNC project. The full political spectrum is also represented, ranging from the business newspaper *Dagens Næringsliv* to *Klassekampen* representing the political left. Of course, any newspaper corpus based on printed newspapers may end up having a different composition than the NNC in terms of content, text types, distribution, etc. Nevertheless, we are convinced that the NNC is a rich resource that covers a very broad range of topics and vocabulary, with a number of different stylistic levels, from information-dense technical terminology to highly informal slang (G. Andersen this volume).

Before the compilation of the NNC, no large corpus of Norwegian was available. Given the experience of previous national corpus-building projects in other countries, we knew that shipments of texts at regular intervals from newspapers or publishers is costly and involves manual work at both ends. The overall aim of our venture has thus been to compile the largest possible corpus of the best possible quality, while minimising the need for manual work. As of April 2011, the NNC is a little short of 1 billion words, which makes it the largest searchable corpus of Norwegian. Our motivation for collecting a large corpus can be viewed against the experience of the British National Corpus (BNC). In this 100 million word corpus "the bulk of the lexical stock occurs less than 50 times in it, which is not enough to draw statistically stable conclusions about the word. For rarer words, and combinations of words, we frequently find no evidence at all" (Kilgarriff & Grefenstette 2003:336). Concerning representativity, we concur with Kilgarriff and Grefenstette, who state that "the web is not representative of anything other than itself" (Kilgarriff & Grefenstette 2003:333). We do not conceptualise the NNC as a representative sample of a larger population, but have rather aimed at including the whole population of Norwegian web-based newspapers in the corpus, in agreement with Manninng & Schütze who state that "one should simply use all the text that is available" (Manning & Schütze 1999:120).

2. An overview of the Norwegian Newspaper Corpus and its system architecture

The text collection for the Norwegian Newspaper Corpus began in 1998. The growth of this dynamic corpus is on average approximately 230,000 words per day. The corpus currently consists of the full web version of 24 Norwegian

newspapers. This includes large, national newspapers like *Aftenposten, Dagbladet* and *VG*, major regional newspapers like *Bergens Tidende* and *Stavanger Aftenblad* and local newspapers like *Sogn Avis*. Most of the newspapers are of general interest, but important niche publications are also included. The selection of newspapers has resulted in a large corpus with broad topical coverage containing relatively homogeneous data, despite the fact that it is harvested from the web. Although maximal efforts have been made to ensure a balance between the two language varieties, the Bokmål variety is massively larger than Nynorsk (cf. Section 2.3). This discrepancy reflects the degree to which Bokmål and Nynorsk are used in newspapers on the web, but it is not necessarily representative of the use of the two varieties in other contexts.

The system involves several stages of processing, most of which run automatically due to a self-executing cron job. Its architecture is visualised as a data flow diagram in Figure 1 and involves the following main steps:

1. *harvesting*: a web-crawler programme (*w3mir* or *wget*) downloads the full internet versions of Norwegian newspapers
2. *boilerplate and duplicate removal*: a set of specifically designed programs automatically selects the core text, including the body text, headline, lead paragraph and picture caption, but discarding advertisements, navigation menus, etc.
3. *language classification*: the texts are classified as either Bokmål or Nynorsk, while English and other foreign texts are discarded
4. *text annotation*: metadata concerning date, author and source are extracted from the source texts, and the texts are machine-classified according to topic and morphosyntactically tagged by the Oslo-Bergen tagger
5. *user interface*: the annotated texts are made available for search via Corpus Workbench and Corpuscle, a new in-house search system
6. *neology extraction*: the inventory of word forms of newly harvested texts is compared with an accumulated list of word forms, and a list of forms not previously recorded is extracted and added to the accumulated word list
7. *neology classification*: new word forms are classified according to orthographic criteria, and anglicism candidates are identified
8. *frequency profiling and lexical database entry*: statistical filters are used to identify neologisms that are most relevant for lexicography and registered in the Norwegian Word Bank and subsequently used by the Oslo-Bergen tagger
9. *extraction of multiword expressions*: sequences of words with a strong tendency to co-occur are extracted from the corpus and added to the lexical database

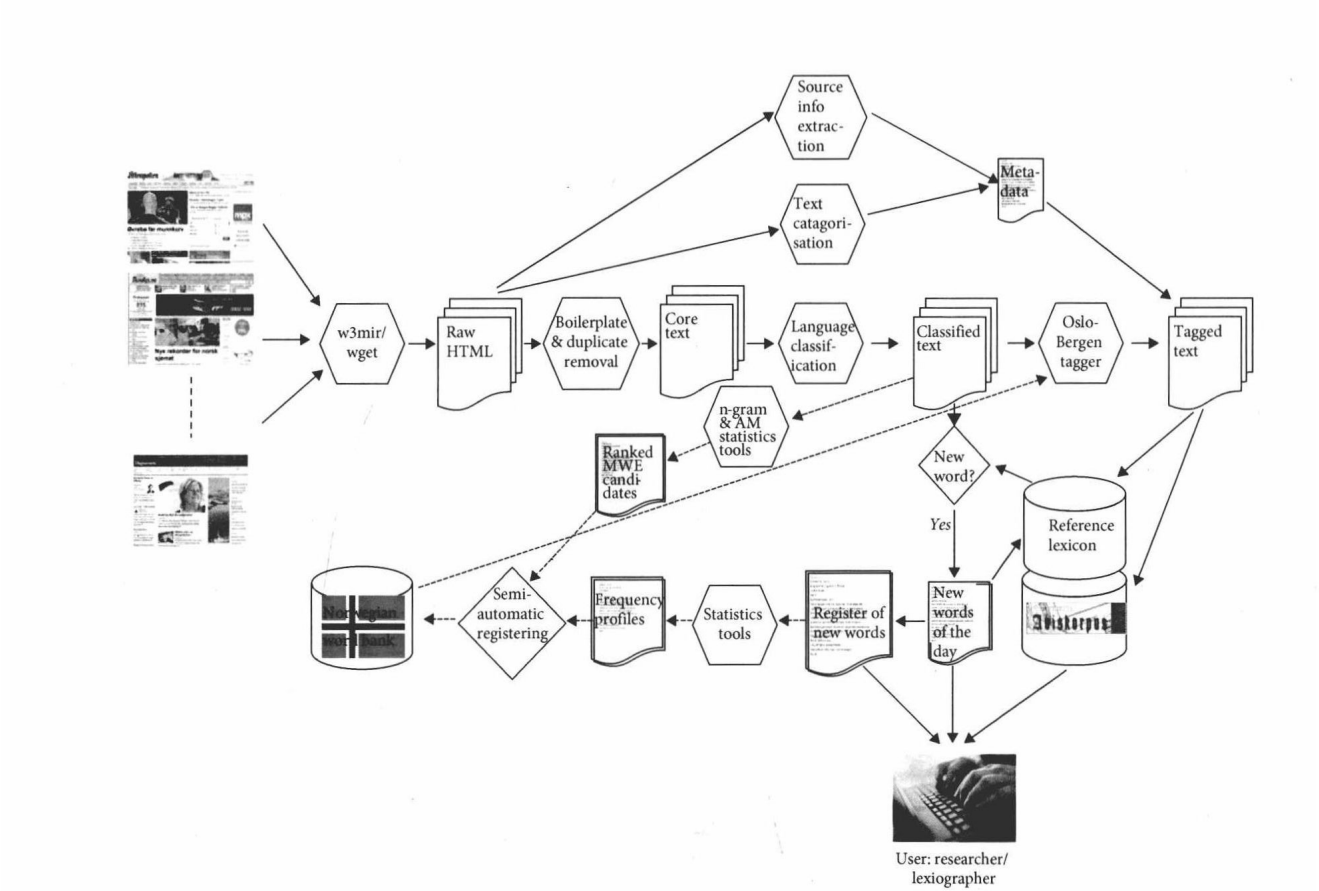

Figure 1. System architecture for the Norwegian Newspaper Corpus

Parts of this system have been running fully automatically since 1998, but some of the modules are of a more recent date. The system runs on a Sun Unix system and uses a number of custom-made programs and shell scripts. Steps 1–7 are performed automatically on a daily basis, while steps 8–9 require manual intervention and are performed in less regular batches (hence the dotted arrows in Figure 1). In the following, we describe each of these steps in more detail. Each of the subsections below corresponds to the numbering of procedural steps in the list above.

2.1 Text harvesting

The NNC uses a web crawler in combination with a variety of different specifically developed scripts (cron, grep, UNIX shells, perl, etc.) in a system which is fully automatic from web text harvesting to corpus text indexing and neology extraction. When the project was started in the late 1990s, a web mirroring and batch download program, w3mir, was chosen.[2] This was one of the few web crawler programs available at the time. W3mir is a set of Perl scripts which can be run from the command line, and it is thus easy to call from other scripts. It has an advantageous option to return a list of all the web-references in a document in addition to the document itself. For technical reasons the web crawler was later replaced by wget, which is a more well-known Linux software package.[3] Originally, the system downloaded all referenced pages recursively, but we soon realised that recursive web crawling led to unduly large downloaded files and a lot of duplication, so the system was adjusted so as to download selected files non-recursively. When a newspaper site is visited by the web crawler program, a list of all URLs referred to on this page is compiled. This list is sorted by means of a series of grep patterns and compared with an accumulated list of previously retrieved URLs from the newspaper, to ensure that only unique and corpus-relevant references are selected. A shell script is used to retrieve these individual pages and to update the accumulated list of retreived URLs. All the articles retreived from an individual newspaper on a particular day are collected in one file. A cron job is set to run automatically every day at 10 PM, and this activates a series of shell scripts which collect the articles and process them further. Since URLs may refer to older newspaper articles, the corpus contains some text that is written before the text collection started.

As part of the administrative project work, each of the newspaper editors was contacted in writing for permission to use the material. In our correspondence, we emphasised that the project represented a non-profit, coordinated, national effort

2. http://www.langfeldt.net/w3mir/

3. http://www.gnu.org/software/wget/

(a text collection *dugnad* 'voluntary effort') which was meant for non-commercial research purposes only. Further, we explained that only registered users would have access to the corpus and its derived data, that they would only be given limited access to the data via concordance views, and that no transfer of the material to any third party would take place. Most importantly, we stressed that the inclusion of their texts in the corpus required no effort on their part, except for an e-mail granting us the right to use their material under the conditions stated in the letter. We received only positive responses from the editors.

2.2 Boilerplate and duplicate removal

The development of tools for the removal of boilerplates and duplicates turned out to be a particularly time-consuming and complicated task. Boilerplate is the name given to repetitive, marginal text which is not part of the core text of the newspaper article (Fairon et al. 2007). Anyone who has visited a web newspaper has seen the large amount of such marginal text. Consider Figure 2:

Figure 2. Sample web newspaper article (*Dagbladet* 5 April 2011)