Forty years of working with corpora: from Ibsen to Twitter, and beyond

Knut Hofland, Paul Meurer and Andrew Salway*

Computational Language Unit, Uni Research

Abstract

We provide an overview of forty years of work with language corpora by the research group that started in 1972 as the Norwegian Computing Centre for the Humanities. A brief history highlights major corpora and tools that have been developed in numerous collaborations, including corpora of literature, dialect recordings, learner language, parallel texts, newspaper articles, blog posts and tweets. Current activities are also described, with a focus on corpus analysis tools, treebanks and social media analysis.

Keywords: corpus building; corpus analysis tools; treebanks; social media analysis

*** Principal contact:**
Andrew Salway, Group Leader (Computational Language Unit)
Uni Computing, Uni Research, Thormøhlensgate 55, N-5008 Bergen, Norway.
Tel.: +47 55584170
E-mail: andrew.salway@uni.no

1. Introduction

In 1972 the Norwegian Computing Centre for the Humanities was formed in Bergen. Over the subsequent years this group morphed into the Humanities Data Centre, the Humanities and Information Technology Centre, Unifob AKSIS and Uni Digital. The latest incarnation, since 2011, is the Computational Language Unit (CLU) at Uni Computing, which is a department of Uni Research AS, Bergen. The constant thread through these groups has been research and development at the intersection of language, culture and information technology. Since the beginning, researchers in these groups have been pioneering techniques for the compilation and analysis of text corpora, both for academic research and for more commercially-oriented purposes[1]. Here we survey some highlights of this work, which has included the creation and analysis of corpora of literature, dialect recordings, learner language, parallel texts, newspaper articles, blog posts and tweets. Through this brief history we comment on the impact of earlier work on current activities in Bergen and elsewhere, and summarise how developments in computing technologies have afforded, and continue to afford, opportunities for new kinds of corpus-based research. Then, looking to the future, we focus on three on-going strands of work at CLU: (i) the development of a corpus analysis system to best exploit the recent availability of large-scale and richly annotated corpora; (ii) the development of a system for building and using treebanks, and their application to linguistics and computational linguistics; (iii) corpus-based approaches to extracting information from massive and heterogeneous corpora of social media.

2. From Ibsen to Twitter

Since its early days, computer-based text analysis has been deployed for investigations into the work of leading literary figures. In the mid-1970s the Norwegian Computing Centre for the Humanities, in collaboration with researchers from University of Bergen, began work on a project to digitise the collected works of Henrik Ibsen, comprising 26 plays and a volume of poetry[2]. The resulting corpus of some 630,000 words was very large considering the data storage technology of the time; then a typical hard disc held 10MB of data. A major output from the project in 1987 was a PC diskette version of the corpus, with an index and a search program. An analysis of the corpus, along with a complete frequency list, was also produced (Noreng, Hofland and Natvig 1987): the list was organised to cross-reference alternate spellings used by Ibsen from 1870 onwards (largely due to new Scandinavian orthography rules), and also to include an index with modern spellings. Later, a complete concordance, minus the 100 most frequent words, was distributed as 3200 pages in six volumes (Noreng, Hofland and Natvig 1993); this included annotations to indicate speakers, word senses and rhyming. The "Ibsen Corpus/Concordance" meant, for the first time, scholars could quickly locate instances of where Ibsen used certain words and phrases, in order to analyse his linguistic style and the themes of his work, and also to check quotations attributed to him. Later it became one starting point for the Henrik Ibsens Skrifter project that developed critical editions[3]. These days the corpus, and related resources, can be accessed via a web interface[4].

Corpus linguistics is also concerned with understanding general or everyday language. Towards the end of the 1970s and into the 1980s, the Norwegian Computing Centre for the Humanities contributed to the development of the Lancaster-Oslo/Bergen corpus. As is perhaps well known, the LOB corpus was created as a British English counterpart to the pioneering Brown Corpus of American English. Thus, it too comprised about a million words from texts published in 1961; 500 samples across 15 text categories (Johansson, Leech and Goodluck

[1] Beyond the topic of this volume are other long-standing areas of work relating to lexicography, terminology and the production of electronic editions.
[2] http://www.hd.uib.no/ibsenbt.html
[3] http://www.ibsen.uio.no/varia.xhtml
[4] http://www.hd.uib.no/ibsen/

1978). Work in Bergen concentrated on entering material, automatic part-of-speech tagging and manual error correction, and producing a frequency list comparing the LOB corpus to the Brown Corpus (Hofland and Johansson 1982). The tagged corpus was used for an extensive study of English based on detailed frequency information about vocabulary and tags (Johansson and Hofland 1989).

Of course, language does not exist only in written form, and the importance of spoken language corpora is well recognised. The widespread emergence of computers with multimedia capabilities in the early 1990s led to new possibilities in this area. The group was involved in creating the Bergen Corpus of London Teenage Language (COLT): this was the first large English Corpus focusing on the speech of teenagers. It comprises the spoken language of teenagers from different London boroughs and totals 500,000 words that have been orthographically transcribed and tagged with word classes[5]. Key technical challenges in the development of COLT related to the integration of digitized sound with text data. The novel search system associated with the corpus allowed users to access sound directly from the results of a concordance. The corpus has been analysed in diverse investigations of teenage speech, for example Stenström, Andersen and Hasund (2002), and was included as a constituent of the British National Corpus. The underlying technology has been used subsequently in other youth dialect projects, including for Norwegian (UNO) and Spanish (COLA), and for more general dialect studies, such as the Norwegian Dialect Corpus.

Also in the 1990s, work began on the creation of parallel text corpora which are of interest for contrastive analysis and translation studies. The group was a partner in the English-Norwegian Parallel Corpus (ENPC) project[6]; later the corpus was extended to include German, Dutch and Portuguese. Given the size of the corpus, some 2.5 million words of English and Norwegian, it was critical to develop a method to align originals and translations automatically. The Translation Corpus Aligner (TCA) tool was developed for this task: it takes an original and a translation, and produces versions of the texts where each sentence in each text is linked to one or more corresponding sentences in the other, based on a bilingual list of anchor words and other features (Hofland 1996, Hofland and Johansson 1998). The TCA2 tool added a graphical user interface that allows the user to interactively correct alignments[7]. Work with parallel corpora has continued to the present time in the group, especially with scholars visiting Bergen as part of the EU-funded Batmult and Multilingua programs.

An innovative kind of parallel corpus was developed to help researchers interested in second language acquisition and language testing. The ASK learner corpus (Tenfjord, Meurer and Hofland 2006) is made up of about 2000 essays written in examinations by students learning Norwegian; it includes essays by students with ten different native languages, as well as control essays written by native Norwegian speakers. Each text was manually annotated according to a coding scheme that includes lexical, morphological, syntactic and punctuation error codes. The annotations are associated with grammatically tagged versions of the original essays, which in turn are aligned with corrected versions. Personal data about the students is also stored, e.g. age, gender, native language, time spent in Norway, amount of Norwegian instruction. The texts, annotations and personal data were integrated in a client-server system, with a web-based user interface which allows researchers to search for interesting error patterns in terms of actual errors and intended use. The system can also generate statistics about errors made by different groups of students, e.g. students sharing with the same native language. The corpus is currently being used by researchers in the ASKeladden project (2010-13)[8], and was the basis for a recent volume of research on language testing (Carlsen 2012).

[5] http://gandalf.uib.no/colt/
[6] http://www.hf.uio.no/ilos/english/services/omc/enpc/
[7] This was developed by Øystein Reigem.
[8] http://www.uib.no/rg/askeladden

As our brief history moves into the late 1990s and early 2000s, so we encounter the emergence of the World Wide Web which has made many impacts on corpus-based work. Not least is the opportunity to harvest corpora from the web, though we recognise this can be problematic both technically and methodologically. Since 1998, Hofland has been harvesting newspaper articles for the Norwegian Newspaper Corpus. This resource stands as a large and continuously updated monitor corpus of contemporary Norwegian, in both its written varieties. As such, it is an invaluable resource for the "study of language change, neologistic usage, and lexical productivity and creativity" (Andersen and Hofland 2012); it is also used by researchers interested in how the media influences society's understanding of major issues, like climate change. Tools automatically check for new stories, download them and extract metadata to organise the texts, e.g. by date, newspaper author and topic. This means it is possible to analyse the emergence and uptake of new words, in different newspapers, at the granularity of days, months or years: a simple example of this is the sudden appearance and continued use of the word "tsunami", Figure 1. In the initial years, 10 newspapers were crawled: at the time of writing, the corpus encompasses 24 national and local publications, and is growing at the rate of some 230,000 words per day – the current total is over 1 billion words. In on-going work, the resource is being made more widely available as the *Norsk aviskorpus*, through the Norwegian Language Bank at the National Library[9]. A volume of research based on the corpus was published recently (Andersen 2012); this includes work on corpus compilation, neology detection, and the extraction of multiword expressions.

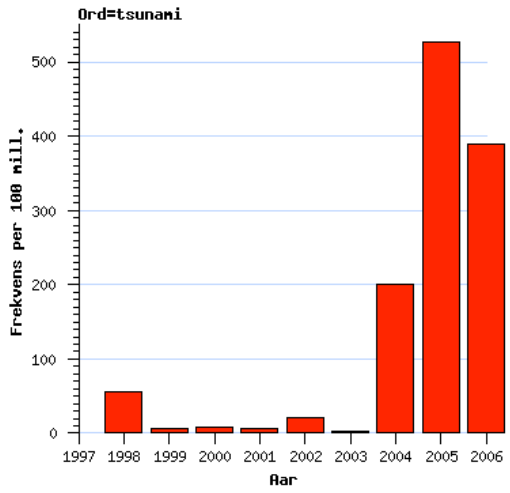


Figure 1. Relative frequency of "tsunami" in the Norwegian Newspaper Corpus.

A backdrop to the forty years of work with language corpora has been the rapid development of computing technologies. We note particular impacts on four aspects of corpus work.

(1) Rather than having to type or scan printed documents, texts are increasingly "born digital" which facilitates speedier corpus gathering. Thus, the quantity and diversity of material available in corpora has increased greatly in recent years, which in turn has stimulated ever greater interest in corpus-based research methods across the humanities and social sciences. One broad example of this is social media: for the last few years we have been gathering a corpus of Norwegian micro-blog posts ("tweets") from Twitter, and in the NTAP project (described later in this paper), we are gathering topic-specific corpora from the blogosphere.

[9] www.nb.no/English/Collection-and-Services/Spraakbanken/Available-resources/Text-Resources

(2) Cheaper computer storage means that it is now feasible to have corpora at the scale of billions of words. This challenges researchers to understand and exploit the possibilities of working with "big data", cf. the multinational "Digging into Data" initiative[10].

(3) Faster processors mean that more complex search and analysis of corpora is possible; much of the work discussed in the following three sections would have been inconceivable 5-10 years ago.

(4) In the past, the sharing of corpus data entailed mailing printed documents, magnetic tapes, floppy discs, CD's, etc. These days, networking technologies mean that corpora and the tools to analyse them can be located on one machine, or across many machines, and yet still be accessed seamlessly by researchers all over the world. CLU is part of the CLARINO initiative that is developing and populating such an infrastructure for language resources in Norway.

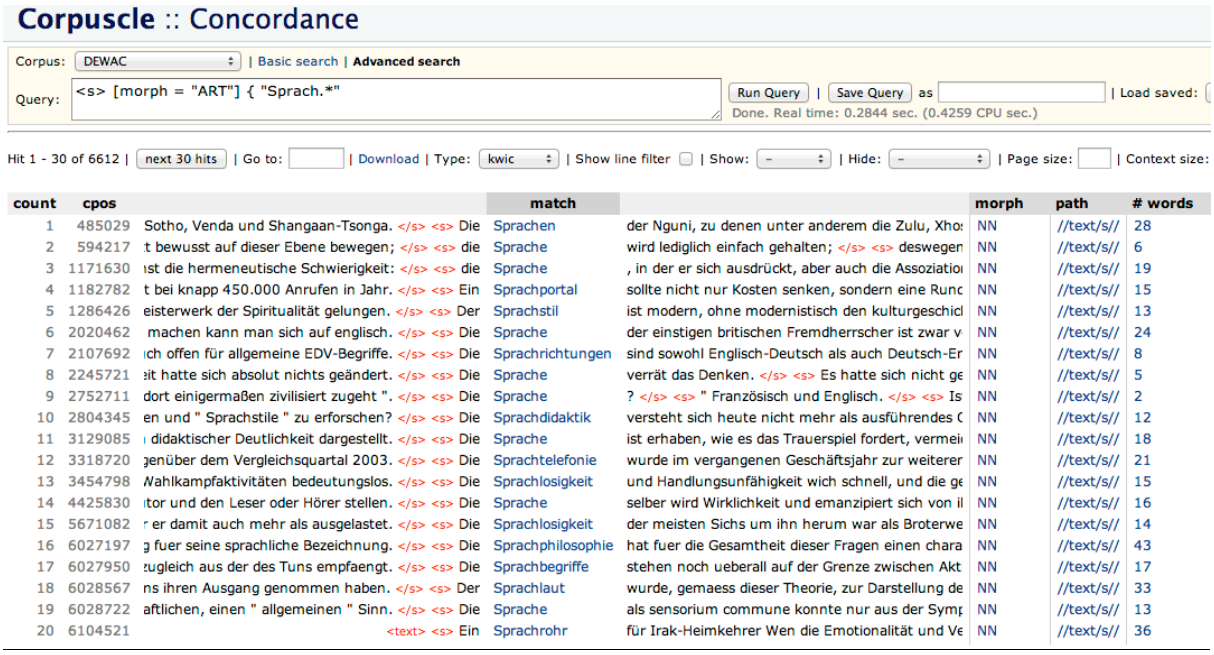
3. Corpus management and analysis software

New tools are needed to best exploit the potential of modern corpora containing billions of words, which may be written in a wide variety of orthographic systems, and which may be accompanied by rich annotation, e.g. using hierarchically structured XML data to describe document structure, error coding or paralinguistic features. The Corpuscle system has been under development at CLU since 2009 to allow users to manage such corpora, to assist them in making complex queries and to execute queries quickly (Meurer 2012a).

Corpuscle's query engine has been designed specifically to deal with very large corpora: it exploits the advent of cheap and plentiful data storage, so rather than being concerned with data compression, its indexing structures are optimised for the speedy execution of complex queries. The technical innovations here are: (i) the way that multiple word queries are evaluated first with respect to the least frequent word; (ii) the use of suffix arrays to store the lexicon which makes for fast look up of strings and regular expression matching; and, (iii) a transparent implementation of multi-valued attributes and set-valued attributes. The web-based interface of the system allows users to express queries either with a powerful query syntax, or through a graphical interface. It offers standard functionality, such as frequency lists, concordances, collocations and distribution statistics. Another important feature is the ability to search in parallel corpora. Furthermore, it facilitates manual corpus annotation with live querying, so that users can work in an interactive manner with their material. The system is freely available to use online[11]. Alternatively, a web server makes most of Corpuscle's functionality accessible via a REST API – a simple HTTP-based protocol, using *post*- or *get*- requests, with responses sent in XML or JSON formats. This means that other developers can incorporate Corpuscle's functionality into their systems and tool chains. In the longer-term, the plan is to release the software on an open source basis.

Here, we include some screenshots to show the system in action. Figure 2a shows a concordance of words that begin with "Sprach", are preceded by an article, and are in sentence initial position. The query was evaluated over the deWaC corpus that comprises 1.7 billion words from German webpages[12]. The query was executed here in less than one second, however this speed is dependent on similar queries having been made previously. Figure 2b shows the corresponding word frequency list for the query. Corpuscle's functionality for parallel corpora is shown in Figure 2c: the query specifies English sentences that contain "work" and that are aligned to Lithuanian sentences that do not contain "darb.*", i.e. the normal translation of "work". In Figure 2d Corpuscle's graphical query interface is being used to compose a query for two adjoining words, with the chosen sub-corpus excluded.

[10] www.diggingintodata.org
[11] http://iness.uib.no/corpuscle
[12] http://wacky.sslmit.unibo.it



Corpuscle :: Concordance

Corpus: DEWAC | Basic search | **Advanced search**

Query: <s> [morph = "ART"] { "Sprach.*"

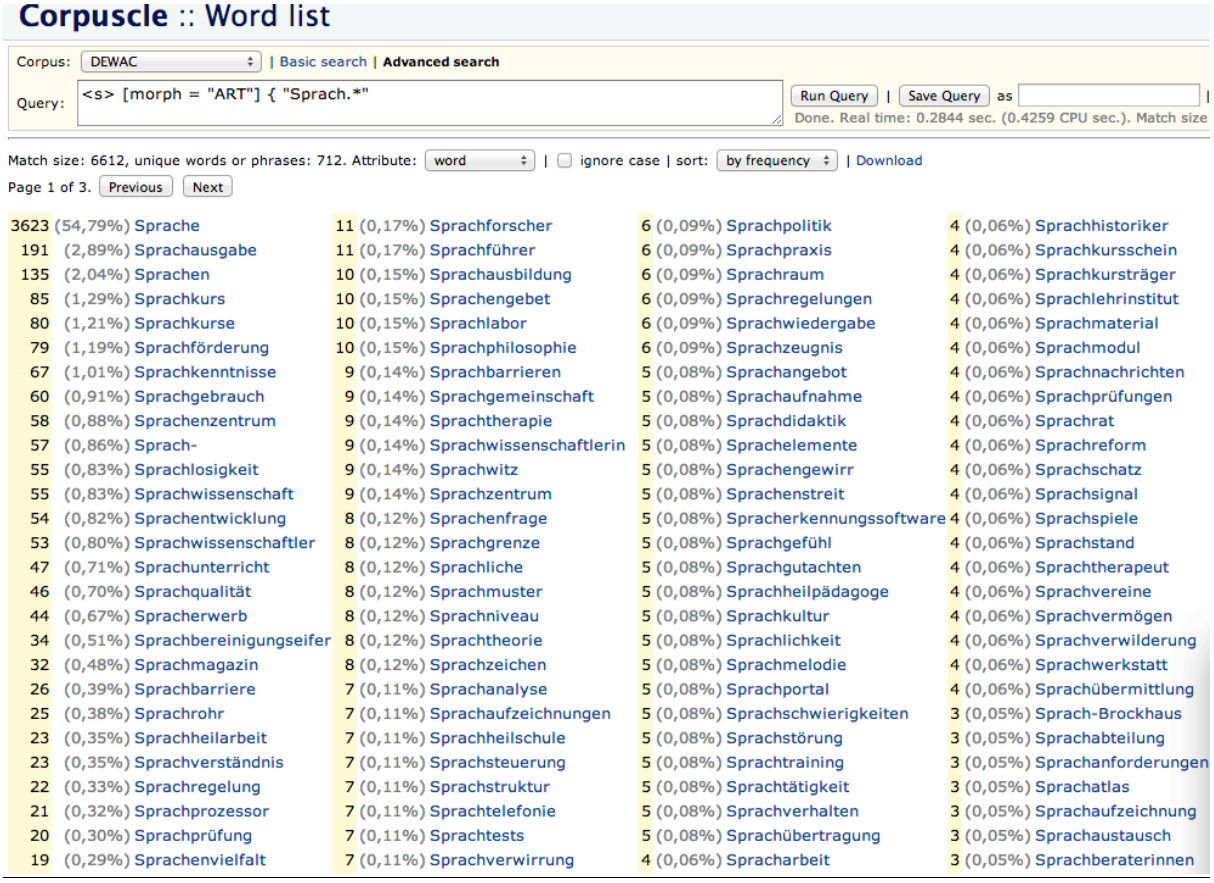
Run Query | Save Query as | Load saved:

Done. Real time: 0.2844 sec. (0.4259 CPU sec.)

Hit 1 - 30 of 6612 | next 30 hits | Go to: | Download | Type: kwic | Show line filter | Show: - | Hide: - | Page size: | Context size:

count	cpos		match		morph	path	# words
1	485029	Sotho, Venda und Shangaan-Tsonga. </s> <s> Die	Sprachen	der Nguni, zu denen unter anderem die Zulu, Xho:	NN	//text/s//	28
2	594217	t bewusst auf dieser Ebene bewegen; </s> <s> die	Sprache	wird lediglich einfach gehalten; </s> <s> deswegen	NN	//text/s//	6
3	1171630	ıst die hermeneutische Schwierigkeit: </s> <s> die	Sprache	, in der er sich ausdrückt, aber auch die Assoziatio	NN	//text/s//	19
4	1182782	t bei knapp 450.000 Anrufen in Jahr. </s> <s> Ein	Sprachportal	sollte nicht nur Kosten senken, sondern eine Runc	NN	//text/s//	15
5	1286426	eisterwerk der Spiritualität gelungen. </s> <s> Der	Sprachstil	ist modern, ohne modernistisch den kulturgeschicl	NN	//text/s//	13
6	2020462	machen kann man sich auf englisch. </s> <s> Die	Sprache	der einstigen britischen Fremdherrscher ist zwar v	NN	//text/s//	24
7	2107692	ıch offen für allgemeine EDV-Begriffe. </s> <s> Die	Sprachrichtungen	sind sowohl Englisch-Deutsch als auch Deutsch-Er	NN	//text/s//	8
8	2245721	ɨit hatte sich absolut nichts geändert. </s> <s> Die	Sprache	verrät das Denken. </s> <s> Es hatte sich nicht ge	NN	//text/s//	5
9	2752711	dort einigermaßen zivilisiert zugeht ". </s> <s> Die	Sprache	? </s> <s> " Französisch und Englisch. </s> <s> Is	NN	//text/s//	2
10	2804345	en und " Sprachstile " zu erforschen? </s> <s> Die	Sprachdidaktik	versteht sich heute nicht mehr als ausführendes (	NN	//text/s//	12
11	3129085	didaktischer Deutlichkeit dargestellt. </s> <s> Die	Sprache	ist erhaben, wie es das Trauerspiel fordert, vermei	NN	//text/s//	18
12	3318720	genüber dem Vergleichsquartal 2003. </s> <s> Die	Sprachtelefonie	wurde im vergangenen Geschäftsjahr zur weiterer	NN	//text/s//	21
13	3454798	Wahlkampfaktivitäten bedeutungslos. </s> <s> Die	Sprachlosigkeit	und Handlungsunfähigkeit wich schnell, und die ge	NN	//text/s//	15
14	4425830	ıtor und den Leser oder Hörer stellen. </s> <s> Die	Sprache	selber wird Wirklichkeit und emanzipiert sich von il	NN	//text/s//	16
15	5671082	r er damit auch mehr als ausgelastet. </s> <s> Die	Sprachlosigkeit	der meisten Sichs um ihn herum war als Broterwe	NN	//text/s//	14
16	6027197	g fuer seine sprachliche Bezeichnung. </s> <s> Die	Sprachphilosophie	hat fuer die Gesamtheit dieser Fragen einen chara	NN	//text/s//	43
17	6027950	zugleich aus der des Tuns empfaengt. </s> <s> Die	Sprachbegriffe	stehen noch ueberall auf der Grenze zwischen Akt	NN	//text/s//	17
18	6028567	ns ihren Ausgang genommen haben. </s> <s> Der	Sprachlaut	wurde, gemaess dieser Theorie, zur Darstellung de	NN	//text/s//	33
19	6028722	aftlichen, einen " allgemeinen " Sinn. </s> <s> Die	Sprache	als sensorium commune konnte nur aus der Symp	NN	//text/s//	13
20	6104521	<text> <s> Ein	Sprachrohr	für Irak-Heimkehrer Wen die Emotionalität und Ve	NN	//text/s//	36

Figure 2a. Corpuscle screenshot showing query and concordance.



Corpuscle :: Word list

Corpus: DEWAC | Basic search | **Advanced search**

Query: <s> [morph = "ART"] { "Sprach.*"

Run Query | Save Query as |

Done. Real time: 0.2844 sec. (0.4259 CPU sec.). Match size

Match size: 6612, unique words or phrases: 712. Attribute: word | ignore case | sort: by frequency | Download

Page 1 of 3. Previous Next

3623 (54,79%) Sprache	11 (0,17%) Sprachforscher	6 (0,09%) Sprachpolitik	4 (0,06%) Sprachhistoriker
191 (2,89%) Sprachausgabe	11 (0,17%) Sprachführer	6 (0,09%) Sprachpraxis	4 (0,06%) Sprachkursschein
135 (2,04%) Sprachen	10 (0,15%) Sprachausbildung	6 (0,09%) Sprachraum	4 (0,06%) Sprachkursträger
85 (1,29%) Sprachkurs	10 (0,15%) Sprachengebet	6 (0,09%) Sprachregelungen	4 (0,06%) Sprachlehrinstitut
80 (1,21%) Sprachkurse	10 (0,15%) Sprachlabor	6 (0,09%) Sprachwiedergabe	4 (0,06%) Sprachmaterial
79 (1,19%) Sprachförderung	10 (0,15%) Sprachphilosophie	6 (0,09%) Sprachzeugnis	4 (0,06%) Sprachmodul
67 (1,01%) Sprachkenntnisse	9 (0,14%) Sprachbarrieren	5 (0,08%) Sprachangebot	4 (0,06%) Sprachnachrichten
60 (0,91%) Sprachgebrauch	9 (0,14%) Sprachgemeinschaft	5 (0,08%) Sprachaufnahme	4 (0,06%) Sprachprüfungen
58 (0,88%) Sprachenzentrum	9 (0,14%) Sprachtherapie	5 (0,08%) Sprachdidaktik	4 (0,06%) Sprachrat
57 (0,86%) Sprach-	9 (0,14%) Sprachwissenschaftlerin	5 (0,08%) Sprachelemente	4 (0,06%) Sprachreform
55 (0,83%) Sprachlosigkeit	9 (0,14%) Sprachwitz	5 (0,08%) Sprachengewirr	4 (0,06%) Sprachschatz
55 (0,83%) Sprachwissenschaft	9 (0,14%) Sprachzentrum	5 (0,08%) Sprachenstreit	4 (0,06%) Sprachsignal
54 (0,82%) Sprachentwicklung	8 (0,12%) Sprachenfrage	5 (0,08%) Spracherkennungssoftware	4 (0,06%) Sprachspiele
53 (0,80%) Sprachwissenschaftler	8 (0,12%) Sprachgrenze	5 (0,08%) Sprachgefühl	4 (0,06%) Sprachstand
47 (0,71%) Sprachunterricht	8 (0,12%) Sprachliche	5 (0,08%) Sprachgutachten	4 (0,06%) Sprachtherapeut
46 (0,70%) Sprachqualität	8 (0,12%) Sprachmuster	5 (0,08%) Sprachheilpädagoge	4 (0,06%) Sprachvereine
44 (0,67%) Spracherwerb	8 (0,12%) Sprachniveau	5 (0,08%) Sprachkultur	4 (0,06%) Sprachvermögen
34 (0,51%) Sprachbereinigungseifer	8 (0,12%) Sprachtheorie	5 (0,08%) Sprachlichkeit	4 (0,06%) Sprachverwilderung
32 (0,48%) Sprachmagazin	8 (0,12%) Sprachzeichen	5 (0,08%) Sprachmelodie	4 (0,06%) Sprachwerkstatt
26 (0,39%) Sprachbarriere	7 (0,11%) Sprachanalyse	5 (0,08%) Sprachportal	4 (0,06%) Sprachübermittlung
25 (0,38%) Sprachrohr	7 (0,11%) Sprachaufzeichnungen	5 (0,08%) Sprachschwierigkeiten	3 (0,05%) Sprach-Brockhaus
23 (0,35%) Sprachheilarbeit	7 (0,11%) Sprachheilschule	5 (0,08%) Sprachstörung	3 (0,05%) Sprachabteilung
23 (0,35%) Sprachverständnis	7 (0,11%) Sprachsteuerung	5 (0,08%) Sprachtraining	3 (0,05%) Sprachanforderungen
22 (0,33%) Sprachregelung	7 (0,11%) Sprachstruktur	5 (0,08%) Sprachtätigkeit	3 (0,05%) Sprachatlas
21 (0,32%) Sprachprozessor	7 (0,11%) Sprachtelefonie	5 (0,08%) Sprachverhalten	3 (0,05%) Sprachaufzeichnung
20 (0,30%) Sprachprüfung	7 (0,11%) Sprachtests	5 (0,08%) Sprachübertragung	3 (0,05%) Sprachaustausch
19 (0,29%) Sprachenvielfalt	7 (0,11%) Sprachverwirrung	4 (0,06%) Spracharbeit	3 (0,05%) Sprachberaterinnen

Figure 2b. A frequency list corresponding to the query in Figure 2a.

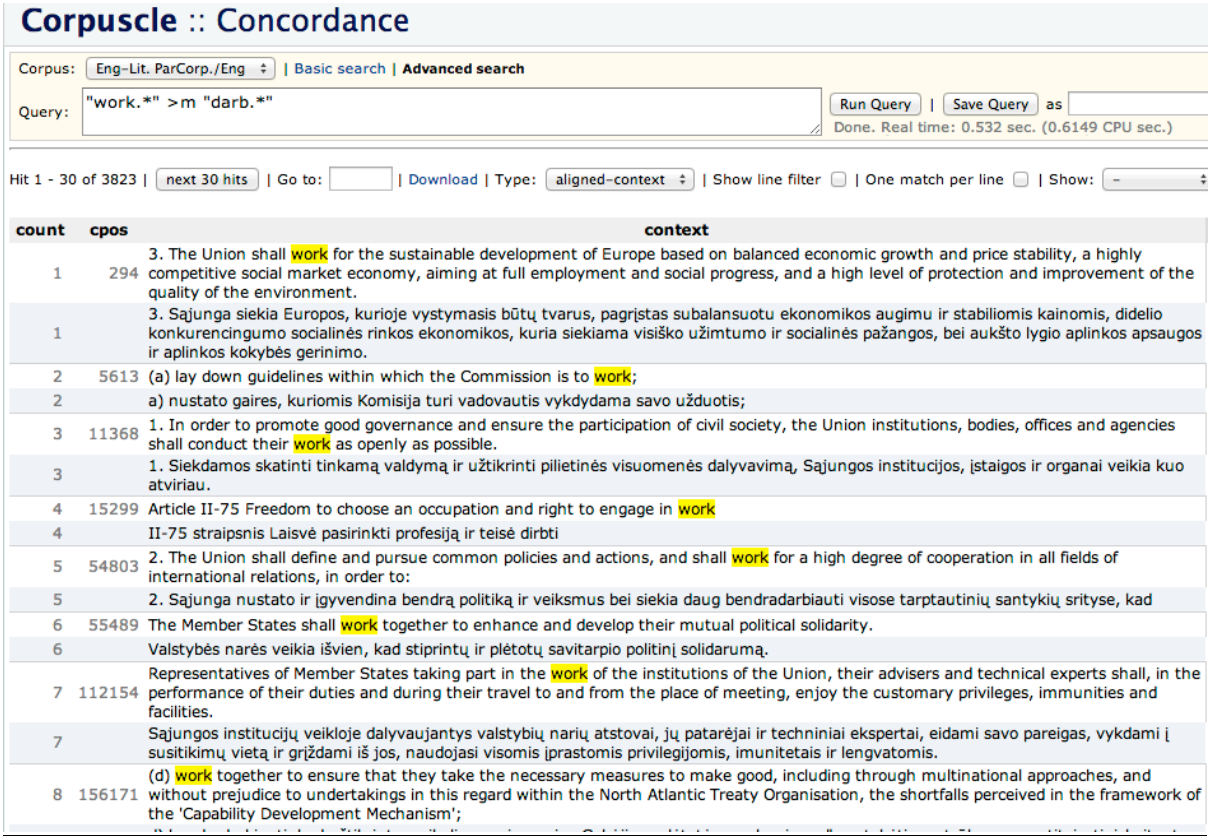


Figure 2c. A search over a parallel corpus of Lithuanian and English.

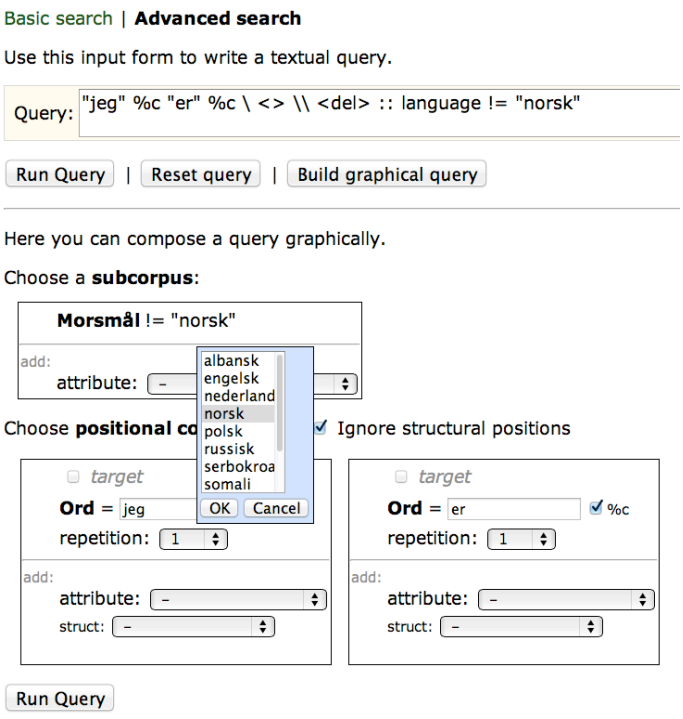


Figure 2d. Corpuscle's graphical query interface.

4. Turning corpora into treebanks

One way in which corpora can be enhanced is with the annotation of syntactic and semantic information. Such treebanks provide crucial evidence about the use of linguistic constructions and thus they support linguistic research, e.g. to test for the existence or frequency of predicted constructions. They also support the development of language technologies, e.g. as training data for automatic parsing systems. CLU is part of the NFR-funded INESS project (2010-16) to develop a fully web-based infrastructure for treebanking (Rosén et al 2012); INESS is closely connected to CLARINO in order to maximise access to the resources that are being developed.

The INESS project had several motivations, including: (i) the need to make the creation of treebanks easier, which means integrating the geographically distributed efforts of different human analysts, along with automated tools where they can make a contribution; (ii) the need to seamlessly manage and access treebanks that represent different languages and that are annotated according to different syntactic and semantic formalisms; and, (iii) the need to make the use of treebanks easier and more productive for researchers by enhancing their search functionality and enabling researchers to interact with visualisations of syntactic and semantic analyses of sentences.

The project currently hosts treebanks (some small) for 24 different languages with a variety of dependency, constituency and LFG structures annotated. A particular aim is to develop a 500 million word treebank of Norwegian (both written varieties, balanced over genres). The approach is to use an automatic LFG parser (NorGram) which may generate thousands of possible parses. Analysts, assisted by an intuitive interface, manually identify correct parses by selecting appropriate discriminants, in order to produce a gold standard treebank of some 5000 sentences; the technique is described in Rosén, Meurer and De Smedt (2009). This gold standard treebank is then used to train a stochastic disambiguation module, so that the remainder of the corpus can be parsed and disambiguated automatically. An initial step is the pre-processing of text to correct OCR spelling errors and to enter morphological information about previously unseen vocabulary items; the interface for facilitating these tasks is described in Rosén et al (2012). An important aspect of the infrastructure is that it supports the iterative improvement of the underlying grammar by handling feedback from analysts to those responsible for writing the grammar. Furthermore, once the grammar has been refined and used to parse the treebank again, there is some memory of previous discriminant choices which can be applied again, which saves the analysts a great deal of time.

With regards to search over treebanks, we are developing INESS-Search for constituency, dependency and LFG treebanks, with several novel features (Meurer 2012b). Firstly, it allows researchers to search not only in tree structures, but also over more general syntactic representations, e.g. f-structures which are an integral part of the LFG formalism. Secondly, its expressive power is equivalent to first order predicate logic, which means queries can include negation and universal quantification. For example, to find an NP node that does not dominate an NP via an intervening S node, the user can write "#x:NP & !(#x >* S >* NP)". An improvement over the popular TIGERSearch is that the query language is less verbose, without loss of expressivity; and, some additional features are available, e.g. explicit specification of quantification, and operators geared towards LFG structures. Search results are displayed with matching tree/c-structure and sub-f-structures highlighted, and the user can choose to see one sub-match at a time, or all possible matches at once. Figure 3 shows a match for the query "V >> (TNS-ASP TENSE) "pres"", which finds c-structure V (Verb) nodes that project to an f-structure whose TNS-ASP TENSE feature is "pres", in the German Tiger LFG treebank. In ongoing development, we are implementing solutions for querying in parallel treebanks, and in HPSG treebanks.

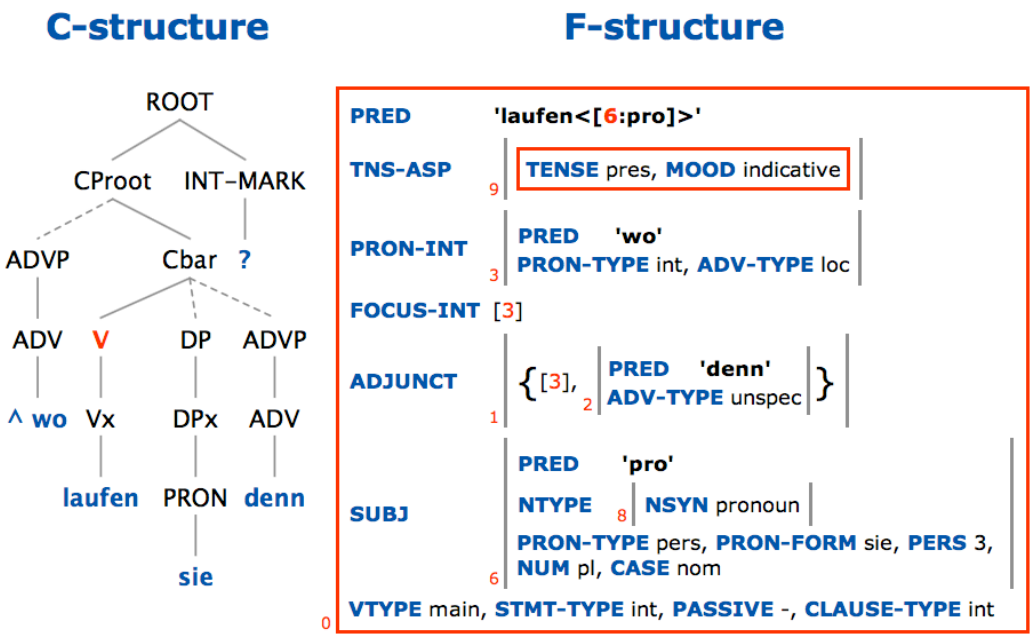


Figure 3. INESS-Search matching the query "V >>(TNS-ASP TENSE) "pres"" in the German Tiger LFG treebank.

5. Extracting information and information structures from corpora

The final area of work that we will describe is at the crossover between corpus linguistics and information management; specifically it is in the field of information extraction which aims to generate structured data from natural language. On the one hand, corpus analysis techniques can reveal linguistic patterning to inform the semi-automated adaptation of information extraction systems, e.g. by elucidating commonly written phrases in topic- and genre-specific corpora. Conversely, information extraction systems can automatically generate semantic annotations for text corpora, e.g. by extracting data about entities and events, which creates new possibilities for research in the humanities and social sciences.

Blogs and other social media are playing an increasingly central role in the public sphere as sources of information and as forums for debate, leading to vast and highly complex online discussions around socially important topics. Given massive and ever increasing text archives, e.g. collections of newspaper stories, blog posts on a popular topic, or the web at large, there is a critical need for users to be assisted in grasping the most important information relevant to their interests, be they citizens engaging a debate around a socially important topic, or researchers investigating the dynamics of such a debate. CLU is a partner in the NFR-funded NTAP project (2012-15)[13] that is developing methods and tools to detect, analyse and visualise the distribution, flow and development of knowledge and opinions across online social networks. A major case study in the project is concerned with the discourse about climate change in the English-language and Norwegian-language blogospheres.

One objective of NTAP is to characterise the content of blog posts as *key statements*, as a complement to the usual treatment of text content as keywords: a key statement is a semi-structured representation of a natural language fragment that renders it more amenable to further analysis. To this end we have adapted and applied a technique developed previously for extracting facts about tourist sites from the web using a web search engine's API (Salway et al 2010). Importantly, this technique does not rely on linguistic resources such as electronic

[13] www.ntap.no

dictionaries and grammars, so it ports easily and cheaply to new domains and languages. It requires only the simple specification of a template and some cues to characterise the ways in which information is typically expressed about the topic. Given the topic "climate change", a simple template (roughly "Subject Verb Object"), and about 15 cues for the "Verb" slot, the technique returned a set of over 1,000 statements. Figure 4 shows a small sample of these statements which characterize different points of view and different aspects of the climate change debate in the English-language blogosphere.

```
climate change…

1.  is_the_result_of, "natural causes"
2.  is_the_result_of, "market failure"
3.  is_the_result_of, "something called the greenhouse effect"
4.  is_the_result_of, "natural fluctuations"
5.  is_the_result_of, "man-made activities"
6.  is_caused_by, "climate-control and other energy-intensive practices"
7.  is_caused_by, "mankind's carbon and other gases output"
8.  is_caused_by, "overpopulation"
9.  is_caused_by, "long-range planetary trends"
10. is_caused_by, "geological factors"
11. is_caused_by, "burning of fossil fuels like coal"
12. is_caused_by, "the excessive amount of carbon emissions poured into
    the atmosphere"
13. is_caused_by, "the cyclical element of nature itself"
14. will_be, "most severe in Africa and South Asia"
15. will_be, "one of major reasons for migration across Asia in the years
    to come"
16. will_be, "the major electoral issue for the next 100 years"
17. will, "solve itself"
18. will, "just go away"
19. will, "boost the asthma rates of children in America"
20. is, "based on fraudulent science"
21. is, "destroying African farm land"
22. is_a, "gender issue because it affects men and women differently"
23. is_a, "liberal conspiracy"
24. is_an, "eco-fascist-communist-anarchist conspiracy"
25. is_an, "international hoax perpetrated by scientists"
26. is_an, "elaborate hoax"
27. is_the, "most critical challenge humanity has ever known"
28. is_one_of, "the challenges that faces the sustainability of
    agriculture"
```

Figure 4. A small selection of statements about climate change extracted automatically from the English-language blogosphere, January 2012.

A set of 1000 key statements may serve as a kind of a summary of a vast online debate, however to help users gain more insight into the issues and the bloggers who write about them, we are working on techniques to relate statements and to visualise how related statements appear over blog networks, and over time. Figure 5 is a mock-up that illustrates some of our thinking in this direction; for a more detailed discussion of it, see Salway, Diakopoulos and Elgesem (2012). The network display maps each blog in the corpus to a node, with the in-degree of that blog represented by the size of the node. Here, blue nodes are blogs that use the selected statement ("climate change is caused by natural causes") or a similar statement: orange nodes are blogs that use a disagreeing statement. To look at information diffusion, the user first selects a statement and then, using the temporal navigator, scrubs through the data over time to see how the network lights up with blue and orange as the statements diffuse.

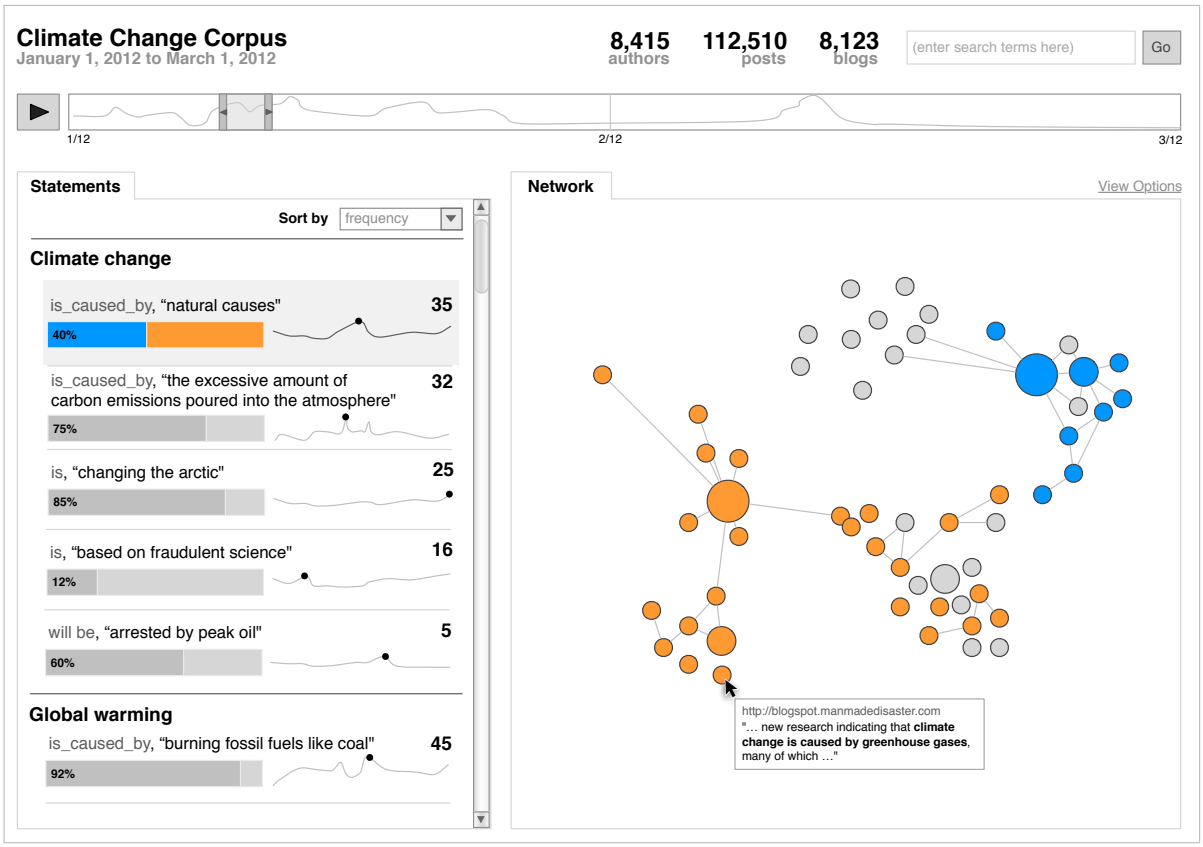


Figure 5. A mockup of an interface depicting a corpus of blog posts collected around the topic of climate change. (Reproduced from Salway, Diakopoulos and Elgesem 2012).

Currently the templates and cues we use for statement extraction are based on intuition and produced manually in a rather ad hoc manner. For this task, and more generally, it would be very desirable to be able to identify linguistic/information structures in a corpus more systematically, even from unannotated corpora of texts representing complex domains and discourses. Previous work showed how local grammar fragments around domain-specific keywords could be induced from an unannotated corpus of film scripts (Vassiliou 2006, Salway 2010). Inspired in part by the formal linguistic analysis of Harris (1954), and by the idea of local grammars (Gross 1997), this technique implements a distributional analysis in which collocation patterns around keywords in a corpus are merged and generalised, in order to induce a local grammar fragment around a word or term. In the NTAP project we are continuing work in this direction in order to automatically generate templates and cues for statement extraction. As a step towards the data-driven induction of salient information structures, we have been running automatic grammar induction techniques, like Solan et al. (2005), on sets of n-grams all containing the same keyword; these were obtained by iteratively querying a blog search engine. We believe that grammar induction techniques may also contribute to new corpus analysis methods by condensing and visualising concordance information.

6. Closing Remarks

We hope to have given the reader some flavour of the diverse corpus work carried out in the past forty years by researchers at the Computational Language Unit, and its predecessors. More importantly, we hope to have shown how the development of corpora and tools has contributed to new avenues of research in many fields. Of course, this was only possible due to countless collaborations with researchers in Bergen and beyond.

Acknowledgements

It is with great pleasure that we acknowledge the people, including some long-standing collaborators, who made significant contributions to the work described in this paper: Gisle Andersen, Koenraad De Smedt, Nick Diakopoulos, Helge Dyvik, Jarle Ebeling, Signe Oksefjell Ebeling, Dag Elgesem, Ingrid Kristine Hasund, Jostein Hauge, Hilde Johansen, Stig Johansson, Annette Myre Jørgensen, Gjert Kristoffersen, Anne Lindebjerg, Kristin Natvig, Harald Noreng, Silje Ragnhildstveit, Øystein Reigem, Victoria Rosén, Helge Sandøy, Anna-Brita Stenström, Sindre Sørensen, Lubos Steskal, Samia Touileb, Kari Tenfjord.

References

Andersen, G. (2012 ed.). *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*. John Benjamins.

Andersen, G. and Hofland, K. (2012). Building a large corpus based on newspapers from the web. In: G. Andersen (ed.), 1-30.

Carlsen, C. (2012 ed.). *Norsk Profil: Det felles europeiske rammerverket spesifisert for norsk.* Oslo: Novus Forlag.

Gross, M. (1997). The Construction of Local Grammars. In: E. Roche and Y. Schabes (eds.), *Finte-State Language Processing*. The MIT Press, 329-354.

Harris, Z. (1954). Distributional Structure. *Word, 10 (2/3),* 146-162.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In: S. Hockey, N. Ide, and G. Perissinotto (eds*.), Research in Humanities Computing*. Oxford: Oxford University Press, 165-178.

Hofland, K. and Johansson, S. (1982). *Word Frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen.

Hofland, K. and Johansson, S. (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In: S. Johansson and S. Oksefjell (eds.). *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies.* Amsterdam: Rodopi, 87-100.

Johansson, S., Leech, G. and Goodluck, H. (1978). Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers. Available: http://khnt.aksis.uib.no/icame/manuals/lob/index.htm

Johansson, S. and Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar.* Oxford: Clarendon Press.

Meurer, P. (2012a). Corpuscle – a new corpus management platform for annotated corpora. In: G. Andersen (ed.), 31-50.

Meurer, P. (2012b). INESS-Search: A Search System for LFG (and other) Treebanks. *Proceedings of the LFG12 Conference.* Available: http://csli-publications.stanford.edu

Noreng, H., Hofland, K. and Natvig, K. (1987). *Henrik Ibsens Ord Skatt.* Universitetsforlaget.

Noreng, H., Hofland, K. and Natvig, K. (1993). *Konkordans Over Henrik Ibsens Drammaer Og Dikt*. University Library Oslo.

Rosén, V., Meurer, P., and De Smedt, K. (2009). LFG PARSEBANKER: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. *Proceedings TLT7*. Available: http://lotos.library.uu.nl/publish/issues/12/index.html

Rosén, V., Meurer, P., Losnegaard, G. S., Lyse, G. I., De Smedt, K., Thunes, M. and Dyvik, H. (2012). An Integrated Web-based Treebank Annotation System. *Proceedings TLT11*. Available: http://tlt11.clul.ul.pt/

Salway, A. (2010). The Computer-based Analysis of Narrative and Multimodality. In: R. Page (ed.), *New Perspectives on Narrative and Multimodality*. London and New York: Routledge, 50-64.

Salway, A., Kelly, L., Skadiņa, I. and Jones, G. (2010), Portable Extraction of Partially Structured Facts from the Web. In H. Loftsson, E. Rögnvaldsson and S. Helgadóttir (eds.), *Lecture Notes in Artificial Intelligence 6233*. Heidelberg, Springer, 345-356.

Salway, A., Diakopoulos, N. and Elgesem, D. (2012). Visualizing Information Diffusion and Polarization with Key Statements. *Procs. SocMedVis 2012*, workshop at AAAI International Conference on Weblogs and Social Media, Dublin.

Solan, Z., D. Horn, E. Ruppin, and Edelman, S. (2005). Unsupervised learning of natural languages. *Procs. National Academy of Sciences*, *102 (33),* 11629-11634.

Stenström, A.-B., Andersen, G. and Hasund, I. K. (2002). *Trends in Teenage Talk: corpus compilation, analysis and findings*. John Benjamins.

Tenfjord, K., Meurer, P. and Hofland, K. (2006). The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language. *Procs LREC 2006*. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/573_pdf.pdf

Vassiliou, A. 2006. Analysing Film Content – A Text-Based Approach. Unpublished Ph.D. thesis, University of Surrey.