NHH

Kan analyser av nettbaserte tekster bidra til å forutsi samfunnsutviklingen?

Nordisk konferanse om språk og teknologi
Nasjonalbiblioteket, Oslo, 7. oktober 2013

Gisle Andersen og Marita Kristiansen

Emne og problemstilling

- Tekst-basert språkteknologi
- Korpuslingvistikk og samfunnsnytte
- Hvordan kan korpuslingvistikk og maskinell tekstanalyse bidra til å si noe om utviklingen av samfunnet vi lever i?
- Diakrone studier av publisert tekstdata samlet inn med korte tidsintervaller
  - Short-term diachronic corpus linguistics (Leech et al. 2009)

Innhold

- Språkressurser og teknologier utviklet i Norge
- Andres språkressurser og teknologier
- Aktuelle metoder/teknologier
  - Nyordsobservasjon
  - Frekvensanalyse
  - Observasjon av termdanning og -variasjon
  - Keyness-analyse (‘weirdness’, CDA)
  - Sentimentanalyse
- Eksempler fra politikk, finans o.a.

Korpuslingvistikk og kritisk diskursanalyse (CDA)

- Critical Discourse Analysis (CDA) ”characterised by the common interests in de-mystifying ideologies and power through the systematic … investigation of semiotic data (written, spoken or visual) …
- (CDA) analyses language as shaped (even in its grammar) by the social functions it has come to serve” (Wodak 2011)
- A corpus-based approach (to CDA) helps to provide quantitative evidence of the existence of discourses by enabling researchers to identify repetitive linguistic patterns of language use and to uncover hidden meanings in lexical items e.g. by examining collocations.
- … (analysis of) a large corpus is likely to show a range of ideological positions - something which an analysis of a single text may be less likely to reveal. (Baker 2006)
- Keyness analysis reveals lexical items that are salient in domain/genre-specific text

Et eksempel: islam/muslims

- Baker, P. (2010) 'Representations of Islam in British broadsheet and tabloid newspapers 1999-2005. *Language and Politics*. 9:2 310-338.
- 87 million word corpus of British newspaper articles
- comparative analysis: tabloid vs. broadsheet newspapers
- analyses of **words** and **collocation patterns**
- tabloids tend to focus more on British interests, writing about Muslims in a highly emotional style, in connection with terrorist attacks and religious extremism, focussing on a small number of high-profile Muslim "villains".
- broadsheets, a more restrained reporting stance, writing about Muslims in a wider range of contexts, covering more stories about Muslims engaged in international wars

Sentimentanalyse

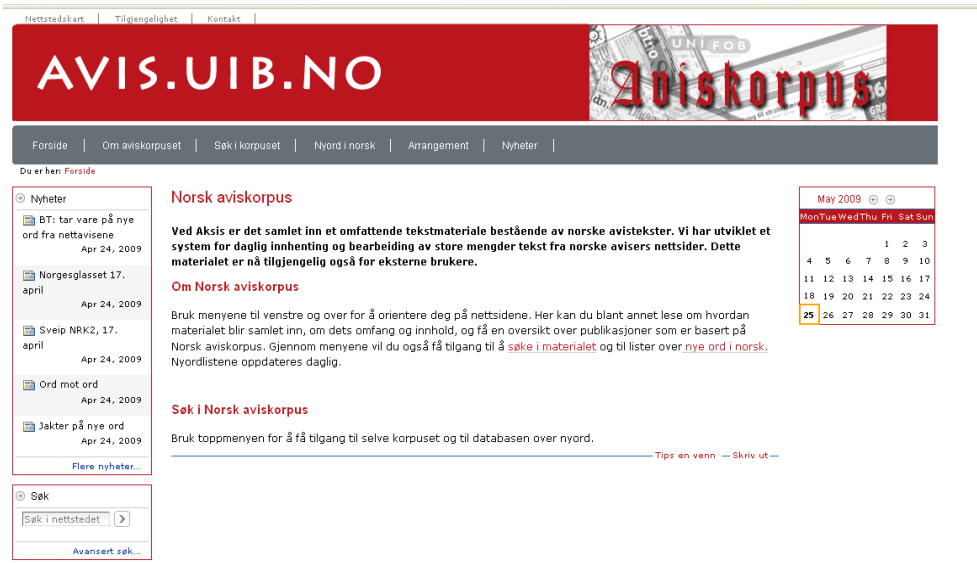
- Sentiment analysis uses the terminology of a specialist domain, say the terminology associated with financial instruments (stocks, currencies, bonds, market indeces) in conjunction with a thesaurus of words that are used in articulating sentiment (Ahmad 2008: 21)
- (Content analysis, news impact analysis)
- E.g. Namenwirth & Lasswell (1970) compare manifestos of Democrats and Republicans
- Showing that manifestos were becoming gradually more similar as time passed
- Leads to the assumption of decreasing political differences between the parties

Forskningsmiljø og ressurser

- Samarbeid om oppbygging av nettbaserte monitorkorpus
- Forskergruppe: NHH/Uni Computing/Univ. of Bergen
- Ressurser
  - Norsk aviskorpus (Andersen 2012; Andersen & Hofland 2012)
  - Forskning.no-korpuset
  - Finansbloggkorpus
  - Eksperimentelt norsk Twitter-korpus

Norsk aviskorpus

- Nettbasert monitorkorpus
- Norsk (bokmål, nynorsk)
- Høster inn tekster daglig fra norske avisers nettutgaver
- 1998-d.d.
- Størrelse: 1 147 944 425 ord
- Daglig vekst: 230 000 ord
- Dagens tekster sjekkes mot eksisterende ordtilfang



Fra starten i 1998	Senere lagt til
Adresseavisen	Klassekampen
Aftenposten	Dag og tid
Bergens Tidende	Firda
Dagsavisen	Hallingdølen
Dagbladet	Hordaland
Dagens Næringsliv	Vest-Telemark blad
Fædrelandsvennen	Sogn avis
Nordlys	Sunnhordland
Stavanger Aftenblad	Gudbrandsdølen Dagningen
Verdens Gang	Sunnmørsposten
	Morgenbladet
	Nationen
	Vikebladet
	Vårt land

Nyordsarbeid for Nynorsk ordliste

- Korpusdrevet nyordsarbeid (Aviskorpus)

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		

Frekvensprofilering

- Identifikasjon av ordformer som er relevante for leksikografisk registrering
- Seleksjonskriterium: frekvensutvikling over tid
- Statistisk standardmål: lineær regresjon med minste kvadraters metode
- Statistiske filtre skiller mellom
    1. ikke-frekvente ord
    2. frekvente ord med stabil/variabel frekvens over tid
    3. frekvente ord med stigende frekvens over tid.
- Frekvensfiltrene reduserer antallet nyordskandidater drastisk:
- Gruppe 3 anses som mest aktuelle nyordskandidater
- NB! Andre kriterier enn frekvens har betydning

Ikke aktuelle	alpindamene, bøllefri, dettestår, vikig, viestad
Aktuell men ikke nyord	akilles, bakfull, lukeparkere
Nyord	dyssosial, fistet, flatskjermen, foodprosessor, hacker, hajj, lagbygging, omrokeringer, zappe

Words of the 05.10.2013

Politikk	Erna · Erna Solberg · Høyre-leder · Jensen · Siv · Siv Jensen · Solberg · Sundvollen
Økonomi	Capital · Klingenberg · Norwegian · Pro · Rema · Twitter
Kultur	Agnete · Alexander · Band · BlackSheeps · Bob · Bø · Cohen · Daily · Emelie · Eriksen · Fox · Grand · John-John · Kennedy · Mimi · Mona · Teater · teater · Uka · Viktoria · « · »
Sport	Aksel · Berget · Brann · Carey · Crystal · Drillo · Frøya · Gudmund · Havnaa · Januzaj · Kenny · Klitsjko · Kongshav · Rekdal · Robin · Sandnes · Siggen · Skjølsvik · Stefan · Steven · Storm · Strømsgodset · Sunderland · Ulf · Viking · Villa · West · Zlat
Innenriks	Helse · Kleppa · kreft · Neda · Tveita
Utenriks	Berlusconi · Bildet · Dagsrevyen · Forsvar · Kongress · Obama · omskjæring · Representantenes · republikaner · Syria · Syrias
Regional	Bjørg · VIP
Forbruker	GT-E1150 · kamera · Stangvik · ull · øl
Teknologi	Nokia · telefon

«««04.10.2013««« Words of the Day [cluster view]

Kollokasjonsanalyse og korpus-basert termekstraksjon i Forskning.no-korpuset

NHH

uniComputing

eng | nob | Sign out *(gislea)*

Corpuscle :: Collocations

- Corpuscle Home
- Documentation
- Publications
- Corpus list
- Overview
- Query
- Concordance
- Collocations
- Distribution
- Word List
- Text
- Localization

Query: "nano.*" = wordlist

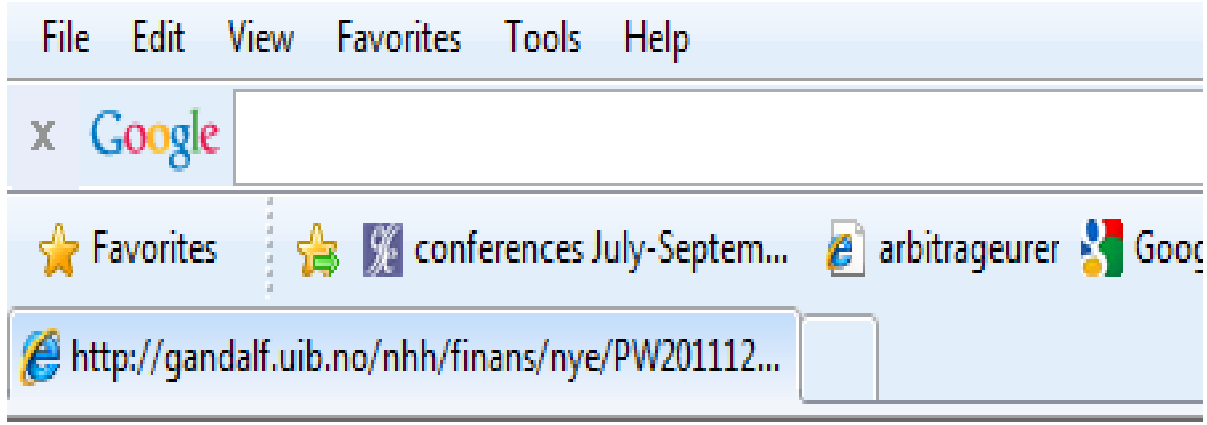
Show collocations by word, left context: 1, right context: 1, sorted by MI * log(Freq) | Download

217 collocations calculated; page 1 of 5. Previous Next | Show concordance for selection

Freq.	Rel. freq.	MI	LL	Delta	Collocate
1	0.2500	19.1794	23.5233	-1	*omhandlar* nanopartiklar
1	0.0435	17.9778	126.9900	1	nanopartiklers *giftighet*
1	0.0238	17.1090	266.7489	-1	*Kartlegger* nanopartiklers
3	0.0067	13.6972	3854.6467	1	nanopartiklenes *virkning*
1	1.0000	16.2919	1492.6072	-1	*10^{20}* nanopartikler
1	1.0000	16.2919	1492.6072	-1	*fareklassifisere* nanopartikler
1	1.0000	16.2919	1492.6072	-1	*plateformede* nanopartikler
1	0.0125	16.1794	597.3529	1	nanopartiklers *mobilitet*
1	0.0294	16.0919	132.5654	-1	*milliardar* nanopartiklar
3	0.1000	12.9699	424.6375	-1	*framstilte* nanopartikler
5	0.0388	11.6026	77.3414	-1	*produserte* nanopartikler
1	0.5000	15.2919	1277.4412	-1	*Hoecke* nanopartikler
1	0.5000	15.2919	1277.4412	-1	*ensartete* nanopartikler
1	0.0030	15.1350	3837.6367	-1	*generell* nanopartikkel-effekt
6	0.0282	11.1421	127.7598	-1	*kunstige* nanopartikler
1	0.0027	14.9699	4385.7750	1	nanopartikkelens *overflate*
1	0.0027	14.9699	4385.7750	1	nanopartikkels *overflate*
2	0.0263	12.7615	62.8505	-1	*selvlysende* nanopartiklene
1	0.1250	13.2919	831.5433	-1	*ensartede* nanopartikler
1	0.0028	12.7077	3000.5625	1	nanopartiklar *kunstig*

Nyordsobservasjoner fra en finansblogg

- Peter Warrens finansblogg (http://www.peterwarren.no/): hedging, finansmarkeder, råvarehandel, makroøkonomiske tema etc.
  - Altså tema med høy samfunnsrelevans



Arkiv

andregangsryktet Europa.. Merkozy Middelalder-politikk Petrescu's
POBC rykter- støttekjøpes valutaopsjonstrader

Bloggen førstemann ut…

PWs blogg – juni 2010

[…] kan det være greit å se om denne stressindeksen endrer retning før man blindt omfavner det ovennevnte kursmål. (PW2010/06)

vs.

Norsk aviskorpus – februar 2012

Norges Banks **stressindeks** er utviklet med utgangspunkt i den Hanschel og Monnin konstruerte for sveitsisk banksektor i 2005… (DN070212)

vs.

Norges Bank – april 2012

«Kan finansiell stabilitet måles? En stressindeks for den norske banksektoren»

(http://www.norges-bank.no/no/om/publisert/publikasjoner/staff-memo/2012/4/)

Sentimentanalyse av endringer i finansmarkedet (Daly et al. 2009)

- Basert på to mål på endringer i finansmarkedene: avkastning (return) og markedsutslag (market volatility)
  - forbrukernes tillit til nasjonaløkonomien (Michigan Consumer Condence Index & Irish Economic and Social Research Institute's Consumer Condence survey
  - uttalte 'sentimenter' i avistekster om finans (Irish Times 1995-2005)

- Sammenlignet med reelle endringer i viktige aksjeindekser (S&P 500, ISEQ og Nikkei) over to år eller mer

- Studiene viser at det er en viss sammenheng mellom perioder med vesentlige markedsutslag i finansmarkedene, eller i viktige aksjeindekser og variasjonen (utslag) av negative nyhetsutsagn i løpet av et år

Markedssentiment…

- Polariserte uttrykk:

Veksten i verdensøkonomien har vært svak [sterk] siden finanskrisen i 2008, […]

Ubalansene har vært selvforsterkende. Høyere [lavere] prising av risiko har gitt økte [reduserte] renter på statspapirer i land med høy [lav] gjeld. […]

Det forsterker nedgangen [oppgang] i den økonomiske aktiviteten. […]

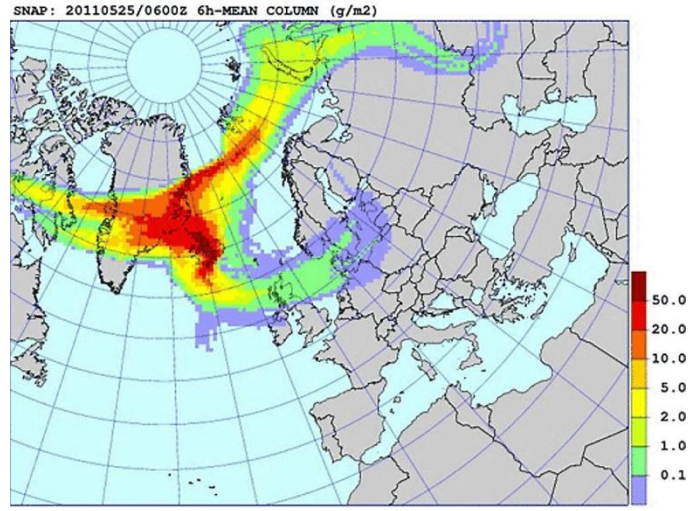
Lavere [høyere] verdsetting av statspapirer har også skapt usikkerhet om bankenes stilling. […] Den sveitsiske sentralbanken mente at en ytterligere styrking [svekkelse] av valutakursen kunne ført til nedgangstid [oppgangstid] med deflasjon [inflasjon].

(http://www.norges-bank.no/no/om/publisert/foredrag-og-taler/2012/cme-foredrag/)

Veien videre… (1/2)

- språket påvirker hvordan vi oppfatter og reagerer på verden rundt oss, både som individer og som samfunn ved at vi bruker språket til å strukturere tankene våre, retorikken og handlinger

- gjør det relevant å analysere og visualisere nettverk for å finne ut hvor termer oppstår, hvordan de over tid spres mellom aktører og på tvers av ulike grupper og teksttyper, og å identifisere hvilke aktører som påvirker spredningen av terminologi/neologismer

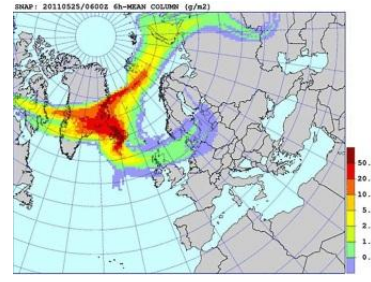


Veien videre… (2/2)

- hvem (hvilke individer, grupper og institusjoner) påvirker termdanning og spredning som en reaksjon på nye hendelser, slik som askekrisen i april 2010 (De Smedt 2012)
- hvordan og hvilke termer oppstår, hvilken termvariasjon kan vi spore, hvilke uttrykk vinner terreng eller dør ut
- hvilken rolle spiller sosiale medier og brukerne, inkludert politikere og journalister som tvitrer, nettaviser og leserkommentarer til nyhetsoppslagene

Eksempler:

- hvordan *ytringsansvar* har oppstått som et parallelt uttrykk til *ytringsfrihet* i debatten etter 22. juli-angrepet;
- hvordan uttrykket *askefast* har ført til en ny produktiv komponent (-fast) som har spredt seg til andre bruksområder;
- hvordan begrepet medmor har tatt steget fra offentlig diskusjon til å bli et lovregulert, juridisk begrep
- fenomenet *gråblogg* som nå har oppstått som en mot-reaksjon til *rosablogg*



Oppsummering

- Tidsaktuelle samfunnsspørsmål kommuniseres ofte gjennom neologismer/nyord
- Ulike sosiale medier har ulik grad av faglighet/faglig status og representerer ulike brukergrupper/samfunnsaktører
- Ved språkobservasjon av sosiale medier ved hjelp av korpusbaserte metoder og ved å fange opp nyord vil en kunne fange opp strømninger i samfunnet; aktuelle begreper og temaer
- Sentimentanalyse og teknologistøttet diskursanalyse kan gi viktige bidra til å beskrive utviklingen av samfunnet vi lever i.

Referanser (1/2)

Ahmad, K. 2011. The ‘return’ and ‘volatility’ of sentiments: An attempt to quantify the behaviour of the markets? I Ahmad, K. (red.). Affective Computing and Sentiment Analysis: Metaphor, Ontology, Affect and Terminology. Heidelberg: Springer.

Ahmad, Kurshid (2008), 'Edderkoppspinn eller nettverk: News media and the use of polar words in emotive contexts', Synaps, 21, 19-35.

AHRC ICT Methods Network Expert Seminar on Linguistics.

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

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.

Baker, P. (2010) 'Representations of Islam in British broadsheet and tabloid newspapers 1999-2005. Language and Politics. 9:2 310-338.

Baker, Paul (2010), Sociolinguistics and Corpus Linguistics (Edinburgh: Edinburgh University Press).

Daly, N., K. Ahmad & C. Kearney. 2009. Correlating Market Movements With Sentiments: A Longitudinal Study, https://www.cs.tcd.ie/Khurshid.Ahmad/Research/Sentiments/2009_Ahmad_Leipzig_Paper.pdf

Referanser (2/2)

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.

Kristiansen, M. 2012.Using web-based corpora to find Norwegian specialised neologies. I Communication and Language at Work 1/2012, 10-19.

Leech, Geoffrey, et al. (eds.) (2009), Change in contemporary English (Cambridge: Cambridge University Press).

Namenwirth, Zvi and Lasswell, Harald D. (1970), The changing language of American values: A computer study of selected party platforms (Beverly Hills: Sage Publications).

Norges Bank, http://www.norges-bank.no/no/om/publisert/publikasjoner/staff-memo/2012/4.

Norsk aviskorpus, http://avis.uib.no.

Peter Warrens finansblogg, http://www.peterwarren.no/.

Wodak, Ruth (2012), 'Critical discourse analysis: overview, challenges, and perspectives', in Gisle Andersen (ed.), Pragmatics of society (Berlin: De Gruyter Mouton), 627-50.