NHH

The corpus-driven approach to neology and its relevance for terminology

4th Termisti/CVC joint seminar
Haute École de Bruxelles

Gisle Andersen, gisle.andersen@nhh.no
29 April 2011

Aims of my presentation

Aims

- The corpus-based/corpus-driven approach to neology
- Argue for the relevance of web-based corpora for terminology and specialised translation purposes

- Describe research methodology and development of language resources (text corpora and associated tools)
  - The Norwegian Newspaper Corpus
  - The Forskning.no Corpus of Popular Science

- Specific focus on multiword expressions and extraction of term candidates

Background: Recent developments in corpus linguistics

Significant events in recent corpus development

- The emergence of mega-size corpora (100 m words +) (Renouf 2007)
- The use of the web as a corpus (Kilgarriff & Grefenstette 2003; Renouf et al. 2007; Hundt et al. 2007)
- The development of dynamic corpora (RDUES; Renouf 1996; Renouf 2007)
- The exploitation of large corpora in lexicography and terminology (e.g. Sinclair 1987; Kilgarriff & Tugwell 2002; Baroni & Bernardini 2004; Atkins & Rundell 2008; Pulcini 2008: 189).

- "for the study of certain phenomena, in particular neologisms, the web is and probably will be one of our best sources of information" (Hundt et al. 2007 :3)

The Norwegian Newspaper Corpus

- inspired and influenced by these developments
- represents the first initiative to create a large monitor corpus of contemporary Norwegian language (Bokmål and Nynorsk)

Background: The web as corpus

Why the www?

- The web: "a mine of language data of unprecedented richness and ease of access" (Kilgarriff & Grefenstette 2003)
- Compiling traditional corpora, costly, complicated and time-consuming (esp. true for LSP purposes)
- Continuously updated
- Wide coverage of topics
- Many languages

Two main strategies

1. Using the web itself as an object of study (Web *as* Corpus)
   - via a commercial search engine, e.g. Google
   - via a tailored search engine, e.g. WebCorp (Renouf et al. 2007)
2. Using specific sections of the www as text sources in the compilation of web-based corpora (Web *for* Corpus)

NHH

Relevance for specialised translation and terminology

- For most specialist fields, no carefully designed corpus exists
- Web-based searches may provide unreliable results
- Using a large, web-based corpus; what is there to gain for the translator faced with the task of translating a technical document, e.g. localisation of user manual from English to another language?

1. A general language corpus contains technical terminology
   - general level, popularised, restricted to certain domains
2. The Web-as-corpus method can be (re-)applied for the compilation of more technical LSP-like corpora

NHH

Relevant sources for the study of neology in Norwegian

General language	The Norwegian Newspaper Corpus (NNC)
Popular science	Forskning.no
Scientific texts	Planned web-based corpus on Business & Finance

The Norwegian Newspaper Corpus (NNC)

- Represents neology in general language
- Web-based **monitor corpus**
- Norwegian (bokmål, nynorsk)
- Daily harvesting and processing of published texts from web edition of Norwegian newspapers
- 1998-present
- Current size: 950 000 000 words; Daily growth: 230 000 words
- http://avis.uib.no/

Newspapers	
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
Nationen	Morgenbladet

w3mir/ wget

Raw HTML

Boilerplate & duplicate removal

Core text

Source info extrac-tion

Text catagori-sation

Meta-data

Language classif-ication

Classified text

Oslo-Bergen tagger

Tagged text

n-gram & AM statistics tools

Ranked MWE candi-dates

New word?

Yes

Reference lexicon

Norwegian word bank

Semi-automatic registering

Frequncy profiles

Statistics tools

Register of new words

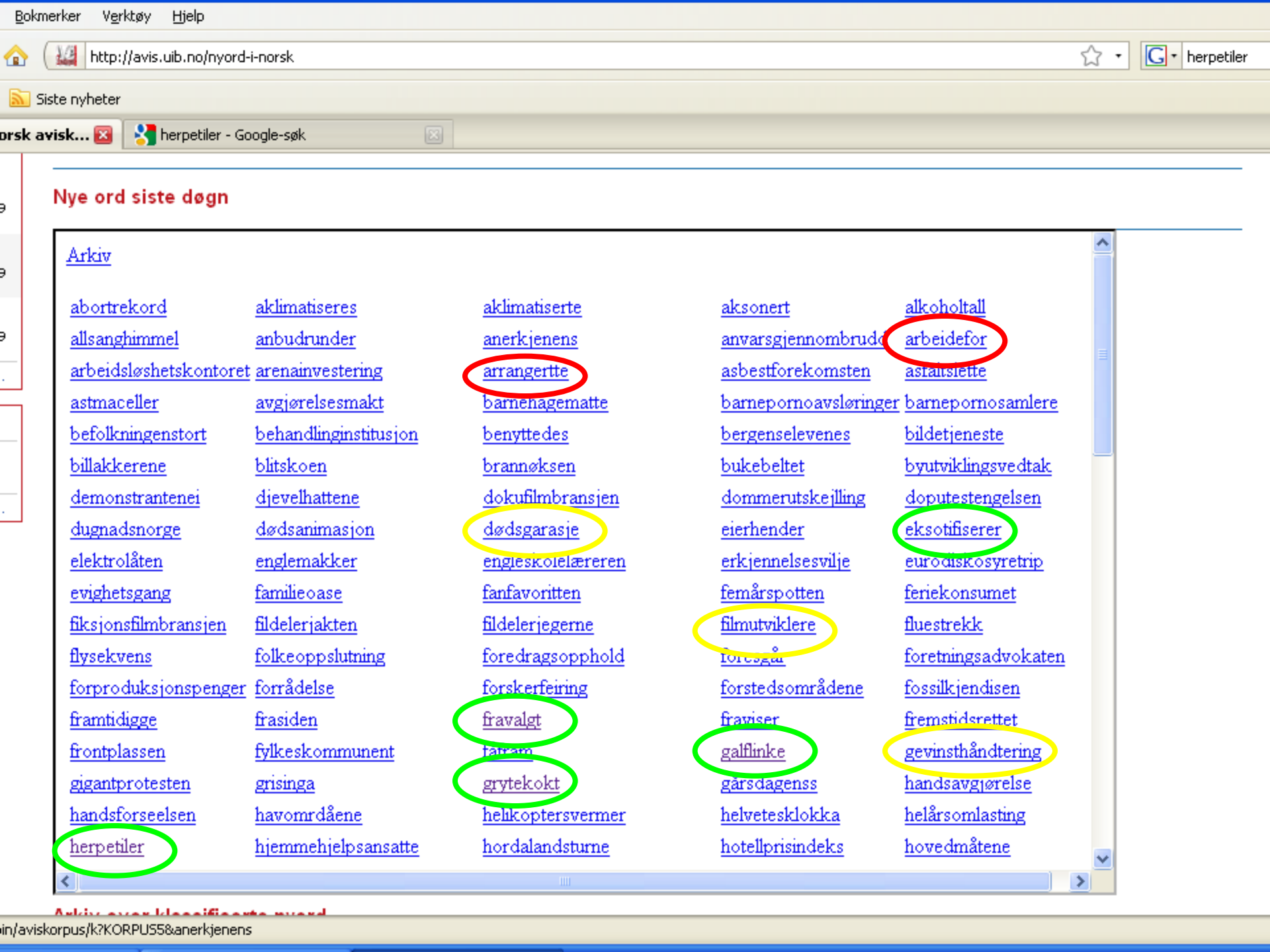
New words of the day

User: researcher / lexicographer

NHH

The corpus-driven approach to neology

- *harvesting*: two different web-crawler programmes, *w3mir* and *wget*, download the full internet version of Norwegian newspapers
- *boilerplate and duplicate removal*: a set of specifically designed programs automatically select the core text, including headlines, introductory and main text, and image texts, but discarding advertisements, navigation menus, etc.
- *language classification*: the texts are classified as either Bokmål or Nynorsk, while English and other foreign texts are discarded
- *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
- *user interface*: the annotated texts are made available for search via Corpus Workbench and Corpuscle, a new in-house search system
- *neology extraction*: the inventory of word forms of newly harvested texts are 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
- *frequency profiling and lexical database entry*: statistical filters are used to identify neologisms that are most relevant for lexicography/terminology and registered in the Norwegian Word Bank and subsequently used by the Oslo-Bergen tagger
- *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


Bokmerker Verktøy Hjelp
http://avis.uib.no/nyord-i-norsk
herpetiler
Siste nyheter
herpetiler - Google-søk
Nye ord siste døgn
Arkiv
abortrekord aklimatiseres aklimatiserte aksonert alkoholtall
allsanghimmel anbudrunder anerkjenens anvarsgjennombrudd arbeidefor
arbeidsløshetskontoret arenainvestering arrangertte asbestforekomsten asfaltslette
astmaceller avgjørelsesmakt barnehagematte barnepornoavsløringer barnepornosamlere
befolkningenstort behandlinginstitusjon benyttedes bergenselevenes bildetjeneste
billakkerene blitskoen brannøksen bukebeltet byutviklingsvedtak
demonstrantenei djevelhattene dokufilmbransjen dommerutskejlling doputestengelsen
dugnadsnorge dødsanimasjon dødsgarasje eierhender eksotifiserer
elektrolåten englemakker engleskolelæreren erkjennelsesvilje eurodiskosyretrip
evighetsgang familieoase fanfavoritten femårspotten feriekonsumet
fiksjonsfilmbransjen fildelerjakten fildelerjegerne filmutviklere fluestrekk
flysekvens folkeoppslutning foredragsopphold foresgå foretningsadvokaten
forproduksjonspenger forrådelse forskerfeiring forstedsområdene fossilkjendisen
framtidigge frasiden fravalgt fraviser fremstidsrettet
frontplassen fylkeskommunent fatram galflinke gevinsthåndtering
gigantprotesten grisinga grytekokt gårsdagenss handsavgjørelse
handsforseelsen havomrdåene helikoptersvermer helvetesklokka helårsomlasting
herpetiler hjemmehjelpsansatte hordalandsturne hotellprisindeks hovedmåtene
in/aviskorpus/k?KORPUS5&anerkjenens

NHH

Navigasjon

- Nyord i norsk
- Arkiv over nyord

Nyheter

BT: tar vare på nye ord fra nettavisene
24.04.2009

Norgesglasset 17. april
24.04.2009

Sveip NRK2, 17. april
24.04.2009

Ord mot ord
24.04.2009

Jakter på nye ord
24.04.2009

Flere nyheter...

Søk

Søk på nettstedet

Avansert søk...

Nye ord siste døgn

I prosjektet utvikles en nyordsdatabase hvor registrerte nyord klassifiseres og legges inn. Denne er under utvikling. man gå inn i prosjektets arkiv over nye ord ved å følge lenkene nedenfor.

Dagens nyord	Her finner du en liste over nyord som er lagt til databasen siste døgn.
Nyordsarkiv	Her finner du alle nyord som er lagt til databasen fra 2001 og frem til i dag, ordnet etter dato.

Den første tabellen nedenfor inneholder alle nyord som ble registrert siste døgn. Den andre tabellen inneholder alle nyord fra 2001 og frem til i dag (sist oppdatert 25. februar 2009).

Hvis du følger lenken nederst i hver tabell (merket "TOTALT ...") får du en alfabetisert liste med alle ordene. Hvis du følger får du alfabetiserte lister over nyord som er kategorisert etter hva slags type nyord det er, dvs. om ordet er et importord, med bindestrek, et navn, osv.

Nye ord siste døgn

Arkiv

billionersedlene	billionseddel	blasfemianklagede	coupeaktig	egenjobbing
fitnessfestival	fotballhore	fôrimportørene	gastropornografisk	klarthetskravet
kulinarisme	kulinarismen	kupeaktige	letthendte	middelklasseeliten
minnenotat	moneygrabbing	naring	nowisme	næringshimmelen
oddsmanageren	particulier	péniche	skipsinnsigantene	skiløperhytte
skotthyllag	skriveisolat	skybridges	solarvirksomhten	staffelimaleriet
takarmaturen	trenervykte	twitterfengsel	vinensusiaster	vorspielsendingen
wikigurk	wikigurker			

Application in terminology work

- Norwegian Language Council's WG for ICT Terminology (DTG)

EN	NO (Bokmål)
computer monitor, display, screen	skjerm
CRT monitor, CRT display, Cathode Ray Tube monitor/display	bilderørsskjerm, CRT-skjerm
flat screen, flat panel display	flatskjerm
Liquid Crystal Display (LCD)	LCD-skjerm
LED display	LED-skjerm
plasma display, plasma screen	plasmaskjerm
organic light emitting diode (OLED)	organisk lysdiode
OLED display	OLED-skjerm
touch screen	berøringsskjerm, trykkfølsom skjerm, trykkskjerm*, pekeskjerm, fingerskjerm
single-touch screen	enkeltberøringsskjerm, enkelttrykksskjerm*
multi-touch screen	flerberøringsskjerm, flertrykksskjerm*

NHH

Match size: 17578, unique words or phrases: 389. Attribute: word | sort: by frequency | Download

Page 1 of 2. Previous Next

5478 (31,16%) skjerm	18 (0,10%) touch-skjerm	5 (0,03%) Plasmaskjerm	3 (0,02%) kartskjerm	2 (0,01%) antirefleks-skjerm	2 (0,01%) visningsskjerm	1 (0,01%) LCD-faltskjerm
3294 (18,74%) fallskjerm	16 (0,09%) allround-skjerm	5 (0,03%) gigantfallskjerm	3 (0,02%) kasse-CRT-skjerm	2 (0,01%) bilskjerm	2 (0,01%) widescreen-LCD-skjerm	1 (0,01%) LDC-flatskjerm
2292 (13,04%) storskjerm	16 (0,09%) hovedskjerm	5 (0,03%) innerskjerm	3 (0,02%) laptop-skjerm	2 (0,01%) bremseskjerm	1 (0,01%) 13,3-tommers-skjerm	1 (0,01%) LED-baklysskjerm
1592 (9,06%) flatskjerm	15 (0,09%) million-fallskjerm	5 (0,03%) karaokeskjerm	3 (0,02%) laptopskjerm	2 (0,01%) brilleskjerm	1 (0,01%) 16:9-flatskjerm	1 (0,01%) Lcd-skjerm
483 (2,75%) dataskjerm	14 (0,08%) fremvisingsskjerm	5 (0,03%) touchskjerm	3 (0,02%) milliardfallskjerm	2 (0,01%) dagslysskjerm	1 (0,01%) 180-graders-skjerm	1 (0,01%) Mac-skjerm
362 (2,06%) fargeskjerm	14 (0,08%) lcd-skjerm	5 (0,03%) underholdnings-skjerm	3 (0,02%) miniskjerm	2 (0,01%) ekstrafallskjerm	1 (0,01%) 21:9-skjerm	1 (0,01%) Million-fallskjerm
358 (2,04%) TV-skjerm	13 (0,07%) LCD-flatskjerm	5 (0,03%) Storskjerm	3 (0,02%) mobiltelefonskjerm	2 (0,01%) ekstraskjerm	1 (0,01%) 3-D-skjerm	1 (0,01%) Nike-solskjerm
312 (1,77%) LCD-skjerm	13 (0,07%) kjempeskjerm	4 (0,02%) 103-tommersskjerm	3 (0,02%) overvåkningsskjerm	2 (0,01%) falskjerm	1 (0,01%) 3d-skjerm	1 (0,01%) Nokia-skjerm
286 (1,63%) berøringsskjerm	13 (0,07%) widescreen-skjerm	4 (0,02%) Bondevik-fallskjerm	3 (0,02%) parabolskjerm	2 (0,01%) farge-berøringsskjerm	1 (0,01%) 42-tommersskjerm	1 (0,01%) PC-storskjerm
185 (1,05%) trykkskjerm	12 (0,07%) blaskjerm	4 (0,02%) DVD-skjerm	3 (0,02%) paragliderskjerm	2 (0,01%) fargerskjerm	1 (0,01%) Almskog-fallskjerm	1 (0,01%) PDA-skjerm
166 (0,94%) Flatskjerm	12 (0,07%) plastskjerm	4 (0,02%) Millionfallskjerm	3 (0,02%) plasma-skjerm	2 (0,01%) fjellskjerm	1 (0,01%) Amoled-skjerm	1 (0,01%) PDP-flatskjerm
162 (0,92%) tv-skjerm	11 (0,06%) LED-skjerm	4 (0,02%) NRK-skjerm	3 (0,02%) plexiglasskjerm	2 (0,01%) flertrykkskjerm	1 (0,01%) Ansiktsskjerm	1 (0,01%) Pc-skjerm
157 (0,89%) støyskjerm	11 (0,06%) TFT-fargeskjerm	4 (0,02%) billedrørskjerm	3 (0,02%) reklameskjerm	2 (0,01%) forstørrelsesskjerm	1 (0,01%) Apple-skjerm	1 (0,01%) Plasma-flatskjerm
156 (0,89%) PC-skjerm	11 (0,06%) reserveskjerm	4 (0,02%) billedskjerm	3 (0,02%) sorthvitt-skjerm	2 (0,01%) gigant-fallskjerm	1 (0,01%) Berørinhgsskjerm	1 (0,01%) QVGA-fargeskjerm
106 (0,60%) videoskjerm	10 (0,06%) glasskjerm	4 (0,02%) framskjerm	3 (0,02%) spesialskjerm	2 (0,01%) gigantskjerm	1 (0,01%) Bildeskjerm	1 (0,01%) Quart-fallskjerm
99 (0,56%) plasmaskjerm	10 (0,06%) videokonferanse-skjerm	4 (0,02%) full-HD-skjerm	3 (0,02%) spillskjerm	2 (0,01%) hjemmekinoskjerm	1 (0,01%) Blåskjerm	1 (0,01%) Ready-skjerm
88 (0,50%) Fallskjerm	10 (0,06%) widescreenskjerm	4 (0,02%) funksjonerBerøringsskjerm	3 (0,02%) stormskjerm	2 (0,01%) infoskjerm	1 (0,01%) CSTN-fargeskjerm	1 (0,01%) Reuters-skjerm
84 (0,48%) millionfallskjerm	9 (0,05%) QVGA-skjerm	4 (0,02%) krystallskjerm	3 (0,02%) vidskjerm	2 (0,01%) innloggingsskjerm	1 (0,01%) Cinema-skjerm	1 (0,01%) Ryggeskjerm
83 (0,47%) fullskjerm	9 (0,05%) bildeskjerm	4 (0,02%) referanseskjerm	2 (0,01%) 103-tommersskjerm	2 (0,01%) kontorskjerm	1 (0,01%) Cinemascopeskjerm	1 (0,01%) SCOPO-skjerm
82 (0,47%) vindskjerm	9 (0,05%) informasjonsskjerm	4 (0,02%) slaveskjerm	2 (0,01%) 15-tommers-skjerm	2 (0,01%) kvalitetsflatskjerm	1 (0,01%) DLP-skjerm	1 (0,01%) SED-skjerm
73 (0,42%) forskjerm	8 (0,05%) Legefallskjerm	4 (0,02%) sort/hvitt-skjerm	2 (0,01%) 4:3-format-skjerm	2 (0,01%) lasteskjerm	1 (0,01%) Dagslys-skjerm	1 (0,01%) Sanktpeterskjerm
63 (0,36%) Storskjerm	8 (0,05%) Støyskjerm	4 (0,02%) startskjerm	2 (0,01%) Full-HD-skjerm	2 (0,01%) m/skjerm	1 (0,01%) Dataskjerm	1 (0,01%) Skandia-fallskjerm
60 (0,34%) Fargeskjerm	8 (0,05%) berøringskjerm	4 (0,02%) statusskjerm	2 (0,01%) Gigant-fallskjerm	2 (0,01%) mikrofonskjerm	1 (0,01%) Drømmeskjerm	1 (0,01%) Sponskjerm
59 (0,34%) pc-skjerm	8 (0,05%) bredformatskjerm	4 (0,02%) venstreskjerm	2 (0,01%) Gullskjerm	2 (0,01%) mini-fallskjerm	1 (0,01%) Flyfallskjerm	1 (0,01%) Stor-skjerm
56 (0,32%) bakskjerm	8 (0,05%) navigasjonsskjerm	3 (0,02%) 2-skjerm	2 (0,01%) HDTV-skjerm	2 (0,01%) pilotskjerm	1 (0,01%) GPS-skjerm	1 (0,01%) Superskjerm
54 (0,31%) solskjerm	8 (0,05%) superskjerm	3 (0,02%) Bredskjerm	2 (0,01%) IMAX-skjerm	2 (0,01%) plasma-flatskjerm	1 (0,01%) GameBoy-skjerm	1 (0,01%) Svart-hvitt-skjerm
52 (0,30%) bredskjerm	7 (0,04%) AMOLED-skjerm	3 (0,02%) LCD-fargeskjerm	2 (0,01%) Info-skjerm	2 (0,01%) ryggeskjerm	1 (0,01%) Gardin-fallskjerm	1 (0,01%) Tegneskjerm
51 (0,29%) TFT-skjerm	7 (0,04%) farveskjerm	3 (0,02%) VGA-skjerm	2 (0,01%) Kjempefallskjerm	2 (0,01%) sideproduksjonsskjerm	1 (0,01%) Giga-fallskjerm	1 (0,01%) Toppskjerm
44 (0,25%) Solskjerm	7 (0,04%) hovedfallskjerm	3 (0,02%) ansiktsskjerm	2 (0,01%) Knallskjerm	2 (0,01%) sluttfallskjerm	1 (0,01%) Gigantfallskjerm	1 (0,01%) Widescreen-skjerm
42 (0,24%) Berøringsskjerm	7 (0,04%) kjempefallskjerm	3 (0,02%) arbeidsskjerm	2 (0,01%) Kraby-fallskjerm	2 (0,01%) splitt-skjerm	1 (0,01%) Gigaskjerm	1 (0,01%) angrepsskjerm
42 (0,24%) OLED-skjerm	7 (0,04%) minifallskjerm	3 (0,02%) boligfallskjerm	2 (0,01%) LCD-berøringsskjerm	2 (0,01%) stjerneskjerm	1 (0,01%) Hughes-fallskjerm	1 (0,01%) ansiktskjerm
37 (0,21%) fjernsynsskjerm	7 (0,04%) småskjerm	3 (0,02%) breddeskjerm	2 (0,01%) Naviskjerm	2 (0,01%) tV-skjerm	1 (0,01%) Hybel-flatskjerm	1 (0,01%) asbestskjerm
37 (0,21%) lampeskjerm	7 (0,04%) ytterskjerm	3 (0,02%) bremsefallskjerm	2 (0,01%) OLED-fargeskjerm	2 (0,01%) tastatur/skjerm	1 (0,01%) Hz-skjerm	1 (0,01%) avskjedsfallskjerm
31 (0,18%) Hovedskjerm	6 (0,03%) 3D-skjerm	3 (0,02%) computerskjerm	2 (0,01%) Pekeskjerm	2 (0,01%) tekstskjerm	1 (0,01%) Høydeskjerm	1 (0,01%) berøringqskjerm
28 (0,16%) pekeskjerm	6 (0,03%) HD-skjerm	3 (0,02%) cruise-fallskjerm	2 (0,01%) Skandale-fallskjerm	2 (0,01%) tolinjers-skjerm	1 (0,01%) II-skjerm	1 (0,01%) blankskjerm
26 (0,15%) filmskjerm	6 (0,03%) Sandberg-fallskjerm	3 (0,02%) data-skjerm	2 (0,01%) Stjerneskjerm	2 (0,01%) tommersskjerm	1 (0,01%) Ink-skjerm	1 (0,01%) blyskjerm
25 (0,14%) mobilskjerm	6 (0,03%) XGA-skjerm	3 (0,02%) elevskjerm	2 (0,01%) TFT-berøringsskjerm	2 (0,01%) u/skjerm	1 (0,01%) JegserenTV.Enflatskjerm	1 (0,01%) bordflateskjerm
24 (0,14%) Sekundærskjerm	6 (0,03%) leskjerm	3 (0,02%) enkeltskjerm	2 (0,01%) TV/flatskjerm	2 (0,01%) ultralydskjerm	1 (0,01%) Kjempe-fallskjerm	1 (0,01%) bredformat-skjerm
23 (0,13%) radarskjerm	6 (0,03%) lysskjerm	3 (0,02%) flat-skjerm	2 (0,01%) Trykkskjerm	2 (0,01%) video-skjerm	1 (0,01%) Kuro-skjerm	1 (0,01%) cinema-skjerm
22 (0,13%) CRT-skjerm	5 (0,03%) 16:9-skjerm	3 (0,02%) gamingskjerm	2 (0,01%) allroundskjerm	2 (0,01%) viedoskjerm	1 (0,01%) LCD-TV-skjerm	1 (0,01%) crt-skjerm

Corpus-based method

- Systematic inspection of a large corpus provides relevant usage statistics of the various alternative term candidates proposed by the Language Council/WG
- But may also reveal other term candidates not thought of in the WG
- Moreover it reveals other concepts that should have been dealt with (standardised) in the same effort
  - *bredskjerm/vidskjerm, HD-skjerm, SED-skjerm, QVGA-skjerm*

berøringsskjerm	707
trykkskjerm*	417
trykkfølsom skjerm	110
pekeskjerm	63
fingerskjerm	4

Forskning.no

- National portal for research-related news
- The **forskning.no corpus**: Using same technology as NNC
- Web-based **monitor corpus**
- **Popular science** texts
- 10 million words
- 1998-present

Språket ut av Afrika

Genetiske metoder brukes for å finne språkets opprinnelse, og resultatene gir støtte til teorien om at språket oppstod kun én gang, på ett sted.

Hanne Jakobsen
journalist

Mandag 25. april 2011
kl. 05:00

At det moderne mennesket kommer fra Afrika, er forskerne stort sett enige om. Men hvor kommer egentlig språket fra?

Det er ikke like lett å avgjøre. Språkene forandrer seg nemlig fort, og da blir det vanskelig å spore dem langt bakover. Hittil har ingen klart å spore noe språk lenger enn cirka 9000 år tilbake i tid.

Men en språkforsker fra New Zealand har nå forsøkt seg på en ny metode for spore språkene mye lenger bakover.

Lingvisten har sett på hvor mange fonemer, altså hvor mange ulike lydenheter, forskjellige språk har, og funnet ut at språket kan ha fulgt samme utviklingsmønster som genene våre.

Dermed mener han at språket også oppsto der genene gjorde det: I Afrika.

San-folket, eller buskmennene, som de ofte kalles, har svært mange fonemer i språket. Kan deres språk være det originale målet? (Foto: Ian Beatty/Wikimedia Creative Commons)

Tema
- Språk
- Evolusjon

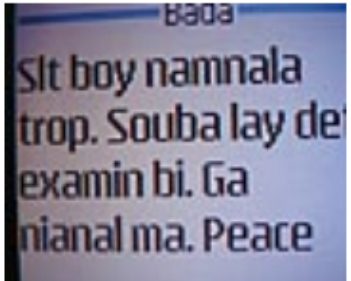
Bakgrunnsstoff

Evolusjon: Fra slik en enkel opprinnelse

- Det er storhet i dette synet på livet, skrev Charles Darwin på siste side av Artenes opprinnelse: - Fra slik en enkel begynnelse, er det utviklet uendeligheter av vakre og vidunderlige former.

Se også

SMS gir liv til truede språk



Relaterte saker
- Forklarer mystiske karbonvariasjoner
- SMS gir liv til truede språk
- Vil redde truet språk
- Vindu til middelalderen
- Planteeter med sabeltenner
- Kjempekanin herjet ferieøy i Middelhavet
- Hvor vill er villaksen?
- Nytt liv i stinkende ursuppe
- Partnerbytte mot parasittar
- Tror mennesket kommer fra det sørlige Afrika

På forsiden nå
- Hvem ga denne til sin elsker?
- Legger lokk over Tsjernobyl
- Til Mars i atomrakett
- Handel pressar utslepp opp
- Skriv ut din egen mat
- Skjebnesvangert møte
- – Verden må få nye atomregler
- 1 million allergikere

Siste blogginnlegg:

Search interface

Søk i Forskning.no med IMS CWB

RESET FORM

Korpus		År	Måned	Dag
Forskning.no				

	Ord 1		Ord 2		Ord 3	
SØK	Starten	klimamodell	Helt ord		Helt ord	
	Minusord		Minusord		Minusord	

Liste	KWIC kontekst (tegn) Venstre	Høyre	Antall linjer per side	Ikke skill Store/små bokstaver
KWIC-konkordans (høyresortert)	65	165	50	☑

NHH

Concordance view

Line 1 to 50 of 443

next | all | new search

Collocates

Lag liste

FN090907	modeller. § Beregninger med Bjerknessenterets nyutviklede	klimamodell	av denne typen tyder på at i en v
FN110204	§ Dette er i grunnen enkel matematikk : Hver enkelt	klimamodell	beskriver både naturlige variasjo
FN030628	kan vi få kritisk vannmangel om noen år. Og da kan en slik	klimamodell	bli skremmende aktuell. § - Probl
FN040610	og dette valget har mer å si enn for eksempel valg av	klimamodell	eller småskalaparameterisering me
FN091102	såkalte sensitivitetsanalyser, som undersøker hvor følsom en	klimamodell	er i forhold til en bestemt param
FN040610	påstanden er sterkt misvisende. Dersom man skulle bruke en	klimamodell	for å beskrive vinden over Sør-No
FN090502	middelverdien for flere simuleringer (' ensembler ') med samme	klimamodell	for å jevne ut slike 10-variasjon
FN100310	Bjerknessenteret samarbeidet for å utvikle og forbedre Bergen	klimamodell	fra 2003. Den er basert på proses
FN110120	- Beregningene i denne regionen er svært følsomme for hvilken	klimamodell	man bruker, understreker RadiÃ‡.
FN031124	en studie av Ben Santer og kollegaer, som sammenligner en	klimamodell	med enda et nytt datasett for sat
FN071229	som ingen klimamodell vil beskrive perfekt. Følgelig vil ingen	klimamodell	noen gang kunne beskrive naturen
FN040708	stemmer bedre overens både med de simulerte temperaturdata en	klimamodell	produserer og de bakkebaserte tem
FN091102	av havis rundt Arktis, mens det bare er en databasert	klimamodell	som kan estimere hvor tykk isen e
FN100914	år tilbake i tid. § Gruppen simulerte klimaet med Bergen	klimamodell	som kobler data fra atmosfære og
FN100805	vedkommende har valgt. § - Vi bruker for første gang en	klimamodell	som regner ut klimaeffekten av al
FN081231	§ 8. Global oppvarming utsatt § I mai skrev vi om en ny	klimamodell	som viser at naturlige klimafakto
FN071002	§ Modelleringen av skyer varierer dermed en hel del fra én	klimamodell	til en annen. § - Det betyr ikke

NHH

Word list view

```
[word="nano.*"%c];
 1687 hits

2 garbage collection(s) in 0.0 seconds.
      1 NANO
      1 NANOHy
     21 NANOMAT
      1 NANOMAT-konferanse
      7 NANOMAT-konferansen
      1 NANOMAT-program
      2 NANOMAT-programmet
      1 NANOMAT-programmets
      1 NANOMATpå
      1 NaNo
     24 Nano
      1 Nano-Sail
      1 Nano-Sail-D
      1 Nano-bling
      1 Nano-divide
      1 Nano-dragsteren
      1 Nano-drømmer
      1 Nano-konferanse
      1 Nano-maskinen
      1 Nano-mat
      1 Nano-robot
      1 Nano-samling
      1 Nano-sensorer
      1 Nano-sjanse
      1 Nano-teknologien
      1 Nano-vaskemaskin
      1 Nano-verdenen
```


```
[word="nano.*"%c];
 1687 hits

2 garbage collection(s) in 0.0 seconds.
```


```
      1 NANO
      1 NANOHy
     21 NANOMAT
      1 NANOMAT-konferanse
      7 NANOMAT-konferansen
      1 NANOMAT-program
      2 NANOMAT-programmet
      1 NANOMAT-programmets
      1 NANOMATpå
      1 NaNo
     24 Nano
      1 Nano-Sail
      1 Nano-Sail-D
      1 Nano-bling
      1 Nano-divide
      1 Nano-dragsteren
      1 Nano-drømmer
      1 Nano-konferanse
      1 Nano-maskinen
      1 Nano-mat
      1 Nano-robot
      1 Nano-samling
      1 Nano-sensorer
      1 Nano-sjanse
      1 Nano-teknologien
      1 Nano-vaskemaskin
      1 Nano-verdenen
      1 NanoCover
     22 NanoSail-D
      1 NanoSailD
      4 NanoSpace-1
      2 NanoVT
      1 Nanobakteriene
      1 Nanobakterier
      2 Nanobil
      1 Nanobotene
      3 Nanoboter
      1 Nanobusiness
```

Distribution across topic key words

```
[word="klimamodell.*"%c];

  442 treff der emneord er angitt

      3 Afrika
      1 Alkohol og narkotika
      5 Antarktis
      2 Arkeologi
     20 Arktis
      1 Bil og trafikk
      3 Biologi
      1 Biologisk mangfold
      1 Botanikk
      1 Data
      1 Dinosaurer
      1 Dyreverden
      2 Epidemier
      2 Evolusjon
      1 Fangst
      1 Fisk
      1 Fiskerifag
      2 Forebyggende helse
      8 Forskningsfinansiering
     15 Forskningsformidling
      6 Forskningspolitikk
     76 Forurensning
     33 Geofag
     19 Havforskning
      2 Informasjonsteknologi
      1 Internett
      1 Kjemi
    397 Klima
      1 Kommunikasjon
      1 Kulturlandskap
      7 Landbruk
      8 Marin geologi
      1 Matematikk
      1 Media
     13 Meteorologi
      2 Miljø
     48 Miljøovervåkning
     45 Miljøpolitikk
     12 Miljøvern
      1 Mobiltelefon
      6 Naturressursforvaltning
      6 Naturvern
     24 Om forskning
      3 Paleontologi
      1 Planteverden
     28 Polarforskning
      2 Politikk
      2 Romfart
      4 Romforskning
      9 Satellitter
      6 Skog
      5 Skogbruk
     59 Statistikk
      1 Svalbard
      2 Sykdommer
      1 Teknologi
     18 Uløste problemer
      2 Universet
     19 Vulkaner
    116 Vær og vind
```

NHH

Collocations

```
[word="b((æ|Æ))rekraftig.*"%c];
    390 {bærekraftig}
     90 {bærekraftig} utvikling
     75 en {bærekraftig}
     40 og {bærekraftig}
     39 {bærekraftige}
     39 for {bærekraftig}
     33 {Bærekraftig}
     24 om {bærekraftig}
     23 mer {bærekraftig}
     19 en {bærekraftig} utvikling
     18 | {Bærekraftig}
     18 på en {bærekraftig}
     18 for {bærekraftig} utvikling
     18 et {bærekraftig}
     17 § | {Bærekraftig}
     16 {bærekraftig} bruk
     16 {bærekraftig} ,
     15 {bærekraftig} utvikling ,
     15 {bærekraftig} mobilitet
     15 {bærekraftig} forvaltning
     14 {bærekraftig} måte
     14 {bærekraftighet}
     14 . § | {Bærekraftig}
     13 på en {bærekraftig} måte
     13 og {bærekraftig} utvikling
     13 en {bærekraftig} måte
     12 {bærekraftig} og
     12 {bærekraftig} forvaltning av
     12 på {bærekraftig}
     12 om {bærekraftig} utvikling
     11 til {bærekraftig}
     11 til en {bærekraftig}
     11 for en {bærekraftig}
     11 er {bærekraftig}
     10 {bærekraftig} bruk av
     10 en mer {bærekraftig}
      9 {bærekraftig} utnyttelse
      9 {bærekraftig} transport
      8 {bærekraftig} produksjon
      8 {Bærekraftig} utvikling
      8 ikke {bærekraftig}
      7 " {bærekraftig}
      7 {bærekraftig} utnyttelse av
      7 være {bærekraftig}
      7 er {bærekraftige}
      6 § {Bærekraftig}
      6 {bærekraftig} utvikling i
      6 {Bærekraftige}
      6 til {bærekraftig} utvikling
      6 og {bærekraftig} bruk
      6 miljømessig {bærekraftig}
      6 internasjonale instituttet for {bærekraftig} utvikling
      6 internasjonale instituttet for {bærekraftig}
      6 instituttet for {bærekraftig} utvikling
      6 instituttet for {bærekraftig}
      6 Det internasjonale instituttet for {bærekraftig} utvikling
      6 Det internasjonale instituttet for {bærekraftig}
      5 {bærekraftig} skogbruk
      5 {bærekraftig} retning
      5 {bærekraftig} måte ,
      5 {bærekraftig} fiske
     15 {bærekraftig} forvaltning
     14 {bærekraftig} måte
     14 {bærekraftighet}
     14 . § | {Bærekraftig}
     13 på en {bærekraftig} måte
     13 og {bærekraftig} utvikling
     13 en {bærekraftig} måte
     12 {bærekraftig} og
     12 {bærekraftig} forvaltning av
     12 på {bærekraftig}
     12 om {bærekraftig} utvikling
     11 til {bærekraftig}
     11 til en {bærekraftig}
     11 for en {bærekraftig}
     11 er {bærekraftig}
     10 {bærekraftig} bruk av
     10 en mer {bærekraftig}
      9 {bærekraftig} utnyttelse
      9 {bærekraftig} transport
      8 {bærekraftig} produksjon
      8 {Bærekraftig} utvikling
      8 ikke {bærekraftig}
      7 " {bærekraftig}
      7 {bærekraftig} utnyttelse av
      7 være {bærekraftig}
      7 er {bærekraftige}
      6 § {Bærekraftig}
      6 {bærekraftig} utvikling i
      6 {Bærekraftige}
      6 til {bærekraftig} utvikling
      6 og {bærekraftig} bruk
      6 miljømessig {bærekraftig}
      6 internasjonale instituttet for {bærekraftig} utvikling
      6 internasjonale instituttet for {bærekraftig}
      6 instituttet for {bærekraftig} utvikling
      6 instituttet for {bærekraftig}
      6 Det internasjonale instituttet for {bærekraftig} utvikling
      6 Det internasjonale instituttet for {bærekraftig}
      5 {bærekraftig} skogbruk
      5 {bærekraftig} retning
      5 {bærekraftig} måte ,
      5 {bærekraftig} fiske
```

Multiword Expressions

- Semi-automatic identification of collocations: two or more words that co-occur so often that they are perceived as a linguistic unit
- Collocativity – the tendency of words to collocate, co-occur more often than would be predicted by chance
- cf. Sinclair (1991); Renouf and Sinclair (1991); Renouf 1996; Moon (1998); Evert (2004); Granger and Meunier (2008)

- Task: identifying multiword expressions in a large corpus
- Assessing their relevance for terminology and lexicography

- Technical terms are commonly realised as multiword units
- MWEs important to identify for the purpose of term extraction

Method for MWE identification (NNC + FN)

1. Create statistics of all n-grams that occur in the corpus
   - bigrams, trigrams
2. Create lists of collocations, i.e. rank the n-grams in terms of their tendency to co-occur
   - using 10 + 4 different association measures (AMs)
3. Inspect the top 500 items of each category, identifying collocations that appear to have terminological/lexicographical relevance
4. Evaluate the usefulness of the different association measures for the purpose of term extraction

Association measures (bigrams)

AM-chi-sqrd-homogeneity-corrected.ranked.word	29.04.2009 15:23	WORD File
AM-dice.ranked.word	29.04.2009 15:23	WORD File
AM-jaccard.ranked.word	29.04.2009 15:23	WORD File
AM-likelihood-ratio.ranked.word	29.04.2009 15:23	WORD File
AM-odds-ratio-discriminative.ranked.word	29.04.2009 15:23	WORD File
AM-pointwise-MI.ranked.word	29.04.2009 15:23	WORD File
AM-t-score.ranked.word	29.04.2009 15:23	WORD File
AM-z-score.ranked.word	29.04.2009 15:24	WORD File
AM-z-score-corrected.ranked.word	29.04.2009 15:46	WORD File
AM-z-score-corrected-filtered.ranked.word	29.04.2009 15:46	WORD File
AM-z-score-corrected-filtered-out-n-grams.ranked.word	29.04.2009 15:46	WORD File
AM-z-score-filtered.ranked.word	29.04.2009 15:46	WORD File
AM-z-score-filtered-out-n-grams.ranked.word	29.04.2009 15:46	WORD File

Extracted MWEs

21.964594	whistle blowers	anglicism mwe
21.949667	patagonian toothfish	anglicism mwe
21.949667	haricots verts	foreign mwe
21.940052	nonfarm payrolls	anglicism mwe
21.93334	yom kippur	foreign mwe
21.84271	navns nevnelse	idiomatic phrase
21.838442	pueblos blancos	foreign mwe
21.838442	perfektum partisipp	term candidate
21.836512	garam masala	foreign mwe
21.78053	frode bjerkestrand	
21.772999	hyperactivity disorder	anglicism mwe
21.754522	amyotrofisk lateralsklerose	term candidate
21.748997	sveinung engeland	
21.71328	educational comix	anglicism mwe
21.71328	amuse bouche	foreign mwe
21.71328	sri lankiske	term candidate
21.71328	whiter shade	
21.71328	selvsvettende tectyl	
21.71328	gently weeps	
21.654438	rollon rolloff	anglicism mwe
21.654438	abra kadabra	idiomatic phrase
21.654438	shih tzu	
21.654438	gesamten norwegischen	
21.654438	derniers jours	
21.597767	curriculum vitae	foreign mwe

Classification scheme

anglicism MWE	*asset value*	asset value
foreign MWE	*alopecia areata*	*alopecia areata* (skin disease)
grammatical MWE	*i motsetning til*	as opposed to
idiomatic phrase	*abra kadabra*	abra cadabra
concept structure appositional phrase	*giftalgen prymnesium parvum*	the poisonous algae prymnesium parvum
term candidate	*amyotrofisk lateralsklerose*	amyotrophic lateral sclerosis

Term candidate MWEs

24.337948	tardive dyskinesier	term candidate
23.767403	patagonske tannfisken	term candidate
23.767403	kilhodet dvergkaiman	term candidate
23.636066	lipsum lorem	term candidate
23.60035	interpleural regionalanalgesi	term candidate
22.960312	retinitis pigmentosa	term candidate
22.888853	nucleus accumbens	term candidate
22.63496	solar plexus	term candidate
22.301065	respiratorisk syncytialt	term candidate
21.838442	perfektum partisipp	term candidate
21.754522	amyotrofisk lateralsklerose	term candidate
21.71328	sri lankiske	term candidate
21.538925	spinal muskelatrofi	term candidate
21.331665	erektil dysfunksjon	term candidate
21.202454	pankreas nekrose	term candidate
21.182045	vitro fertilisering	term candidate
21.035398	methyl isocyanat	term candidate
20.822964	kaffir limeblader	term candidate
20.813917	patagonsk tannfisk	term candidate
20.72288	elektrolytisk manganmetall	term candidate
20.643866	polyklorerte bifenyler	term candidate
20.548527	resultatbaserte omfordelingsmodellen	term candidate
20.471565	monokrystallinske silisiumskiver	term candidate
20.464855	kerrs pink	term candidate
20.330614	konjugert linolsyre	term candidate
20.31515	malignt melanom	term candidate

Anglicism MWEs

21.202454	shetland sheepdog	anglicism mwe
21.202454	respiratory infection	anglicism mwe
20.966064	kidney pie	anglicism mwe
20.928017	plea bargaining	anglicism mwe
20.851055	splendid isolation	anglicism mwe
20.784718	collateral damage	anglicism mwe
20.729548	extreme makeover	anglicism mwe
20.72288	collateralized debt	anglicism mwe
20.650385	impartial investigation	anglicism mwe
20.631908	buck stops	anglicism mwe
20.464855	noodle soup	anglicism mwe
20.39536	graphic novels	anglicism mwe
20.340229	compliance consultant	anglicism mwe
20.318953	stiff upper	anglicism mwe
20.212055	visibility corp	anglicism mwe
20.212055	predatory pricing	anglicism mwe
20.103842	flip flops	anglicism mwe
20.053831	clotted cream	anglicism mwe
19.936787	honky tonk	anglicism mwe
19.924793	chick flicks	anglicism mwe
19.668524	payroll taxes	anglicism mwe
19.642715	brain dysfunction	anglicism mwe
19.624268	bloody mary	anglicism mwe
19.593016	seabed logging	anglicism mwe
19.593016	polling stations	anglicism mwe
19.532187	cottage cheese	anglicism mwe

Foreign MWEs

21.202454	buon giorno	foreign mwe
21.121328	chop suey	foreign mwe
21.119072	planum temporale	foreign mwe
21.119072	pata negra	foreign mwe
21.035398	storia semplice	foreign mwe
20.934189	innocentum martyris	foreign mwe
20.834728	enfants terribles	foreign mwe
20.77501	basso continuo	foreign mwe
20.767136	mensa rotunda	foreign mwe
20.73709	culpa levissima	foreign mwe
20.707758	erat demonstrandum	foreign mwe
20.691628	vinho verde	foreign mwe
20.682875	notarius publicus	foreign mwe
20.631908	staphylococcus aureus	foreign mwe
20.614666	haemophilus influenzae	foreign mwe
20.614666	habeas corpus	foreign mwe
20.566465	vous plait	foreign mwe
20.531284	salade nicoise	foreign mwe
20.493305	vox populi	foreign mwe
20.471565	chronicarum cum	foreign mwe
20.400106	laterna magica	foreign mwe
20.355156	terra nullius	foreign mwe
20.304512	rattus norvegicus	foreign mwe
20.294897	carpe diem	foreign mwe
20.239643	suede shoes	foreign mwe

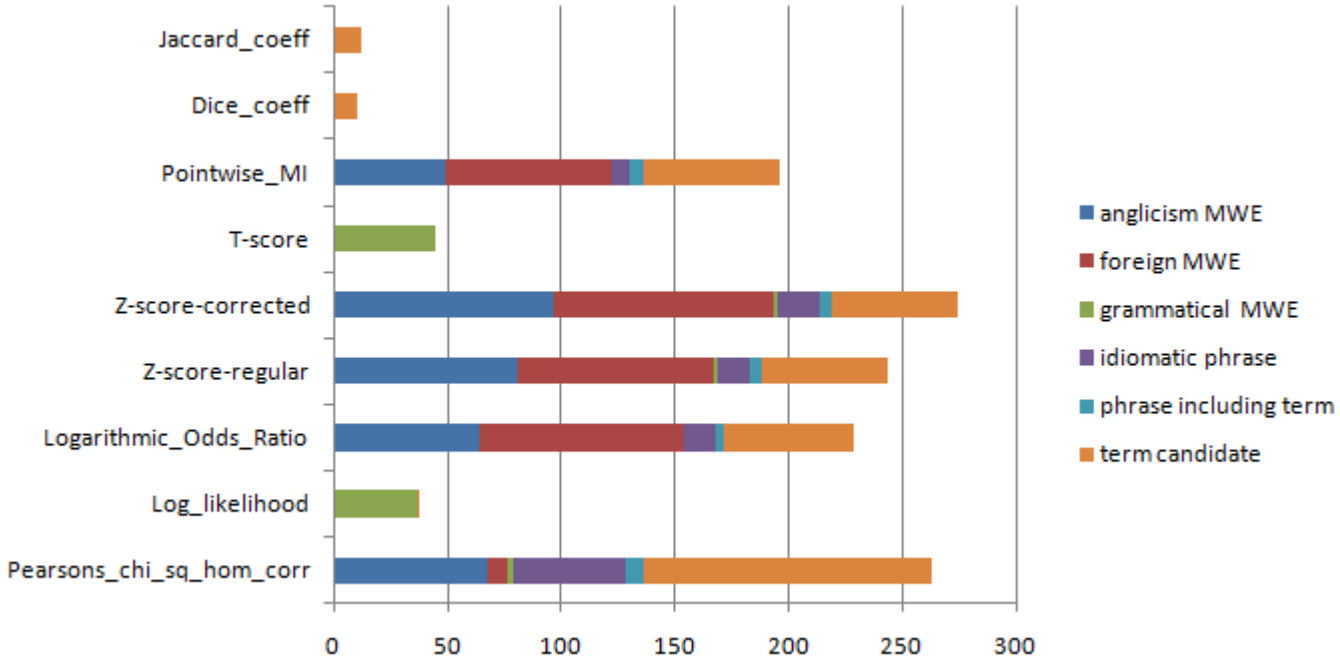
Concept structure appositional phrase

19481.045	stresshormonet kortisol	phrase incl term
17400.227	reststoffet hypoxantin	phrase incl term
15357.731	nervegassen sarin	phrase incl term
14988.746	blodsykdommen thalassemi	phrase incl term
14931.618	muggsoppgiften aflatoksin	phrase incl term

Survey of Association Measures 1

Association measure	anglicism MWE	foreign MWE	gramm'l MWE	idiomatic phrase	phrase incl term	term cand.	SUM	PCT relevant
Pearsons_chi_sq_hom_corr	68	8	3	49	8	127	263	52,6 %
Log_likelihood	0	0	37	0	0	1	38	7,6 %
Logarithmic_Odds_Ratio	64	90	0	14	3	58	229	45,8 %
Z-score-regular	81	86	2	14	5	56	244	48,8 %
Z-score-corrected	97	96	2	19	5	55	274	54,8 %
T-score	0	0	45	0	0	0	45	9,0 %
Pointwise_MI	49	73	0	8	6	60	196	39,2 %
Dice_coeff	0	0	0	0	0	10	10	2,0 %
Jaccard_coeff	0	0	0	0	0	12	12	2,4 %

Survey of Association Measures 2



Trigram analysis

Association measure / Cat.	anglicism MWE	foreign MWE	gram. MWE	idiomatic phrase	appos. term phr	term cand.	SUM	% rel.
Log likelihood	0	0	1	0	0	0	1	0,2 %
Poisson-Stirling	0	0	0	62	0	0	62	12,4 %
Pointwise MI	59	26	0	8	17	17	127	25,4 %
True MI	0	0	0	1	0	0	1	0,2 %

MWEs in Forskning.no

Preliminary analysis

- Much higher term-density, very little noise
  - for some AMs, up to 80-100 per cent of highly ranked n-grams are lexicalised phrases/term candidates
- AMs well suited for identifying term candidates
  - z-score regular, z-score corrected, pointwise MI, odds-ratio, (jaccard, dice)
- AMs unfitted for identifying term candidates
  - t-score, poisson-sterling, Pearson, log likelihood,

Concluding remarks

- A large general newspaper corpus seems to be a valuable repository for terminology including multiword terms
- Wide coverage of topics, useful reference for specialised translation
- Web as corpus methodology is also potentially useful for LSP purposes
- Extracting collocates is an efficient method for retrieving term candidates

- Major differences between different association measures
- Choice of AM dependent on purpose
- It is not known how much of these findings have a language-specific application or whether they apply across languages

References

Atkins, Sue, and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford and New York: Oxford University Press.

Baroni, Marco, and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. Paper read at LREC 2004.

Evert, Stefan. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations, IMS, University of Stuttgart.

Granger, Sylviane, and Fanny Meunier, eds. 2008. *Phraseology: an interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins.

Hundt, Marianne, Carolin Biewer, and Nadja Nesselhauf. 2007. *Corpus linguistics and the web*, *Language and computers*. Amsterdam: Rodopi.

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

Kilgarriff, Adam, and David Tugwell. 2002. Sketching words. In *Lexicography and Natural Language Processing - A Festschrift in Honour of B.T.S. Atkins*, edited by M.-H. Corréard. Gothenburg: EURALEX.

Moon, Rosamund ed. 1998. *Fixed expressions and idioms in English: a corpus-based approach* Oxford: Clarendon Press.

Pulcini, Virginia. 2008. Corpora and lexicography: the case of a dictionary of Anglicisms. In: *Investigating English with corpora : studies in honour of Maria Teresa Prat*, ed. by A. Martelli & V. Pulcini, 189-203. Monza: Polimetrica.

Renouf, Antoinette. 1996. The ACRONYM Project: Discovering the textual thesaurus. In *Synchronic corpus linguistics*, edited by C. E. Percy, Charles F. Meyer and Ian Lancashire. Amsterdam/Atlanta: Rodopi.

Renouf, Antoinette. 2007. Corpus development 25 years on: from super-corpus to cyber-corpus. In *Corpus linguistics 25 years on*, edited by R. Facchinetti. Amsterdam/New York: Rodopi.

Renouf, Antoinette, Andrew Kehoe, and Jayeeta Banerjee. 2007. WebCorp: an integrated system for web text search. In *Corpus linguistics and the web*, edited by M. Hundt, N. Nesselhauf and C. Biewer. Amsterdam/New York: John Benjamins.

Renouf, Antoinette, and John McH. Sinclair. 1991. Collocational frameworks in English. In *English Corpus Linguistics - Studies in Honour of Jan Svartvik*, edited by K. Aijmer and B. Altenberg. London/New York: Longman.

Sinclair, John McH. 1991. *Corpus, concordance, collocation*, *Describing English language*. Oxford: Oxford University Press.

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