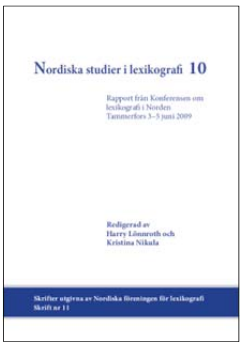
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Halvautomatisk ekserpering av anglisismer i norsk

Forfatter: Gisle Andersen

Kilde: Nordiska Studier i Lexikografi 10, 2010, s. 72-85
Rapport från Konferens om lexikografi i Norden, Tammerfors 3.-5. juni 2009

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Gisle Andersen

Halvautomatisk ekserpering av anglisismer i norsk

This paper reports on the status of ongoing corpus-based lexicographical work within the framework of the Norwegian Newspaper Corpus project (http://avis.uib.no/). Specifically it describes the work flow, tools and methods used in the identification and analysis of new anglicisms in Norwegian. The identification of recent English loan words serves a variety of purposes, including term extraction and the development of lexicography and terminology, and language political purposes such as surveying the amount and inventory of English loan words in various usage domains. While previous work in Norwegian lexicography has generally relied on manual methods for excerpting new words – and for identifying anglicisms among the new words, the current project is an effort to develop tools which automatises the process of identifying, segmenting and analysing new loan words from English. The article describes the overall workflow and focuses especially on alternative methods for identifying anglicisms (lexicon-based, n-gram-based, combinatory methods), as well as the relevance of these methods for lexicography.

Stikkord: korpus, aviskorpus, nyordsekserpering, anglisismer, importord

Det har de senere år blitt stadig vanligere å ta i bruk elektroniske korpus i forbindelse med leksikografiarbeid (Renouf 1987, 2007, Sinclair 1987, Church & Hanks 1989, Atkins 1993, Summers 1993, Stubbs 1995, Kilgarriff 1998, Munat 2007). Et tekstkorpus kan inneholde mye verdifull informasjon om nye ord i språket, og om bruksendringer og nye betydninger av gamle ord. En utfordring kan være å finne korpusmateriale som er oppdatert og tilstrekkelig omfattende. *Norsk aviskorpus* er et nettbasert og stadig voksende tekstkorpus som utgjør en verdifull kilde til informasjon om det norske språkets utvikling.[1] I prosjektet er det utviklet metoder for automatisk registrering og analyse av nyordsdanning. Denne artikkelen beskriver bruken av *Norsk aviskorpus* som grunnlag for

1 Prosjektet er et samarbeid mellom Uni Digital (tidl. Unifob AKSIS) tilknyttet Universitetet i Bergen og leksikografer ved Institutt for nordistikk og litteraturvitenskap ved Universitetet i Oslo (jf. http://avis.uib.no/). Se også Ruth Vatvedt Fjelds og Lars Nygaards artikkel i denne antologien.

leksikografiarbeid og fokuserer særlig på hvordan aviskorpuset kan brukes til å identifisere og analysere bruken av engelske ord i norsk.

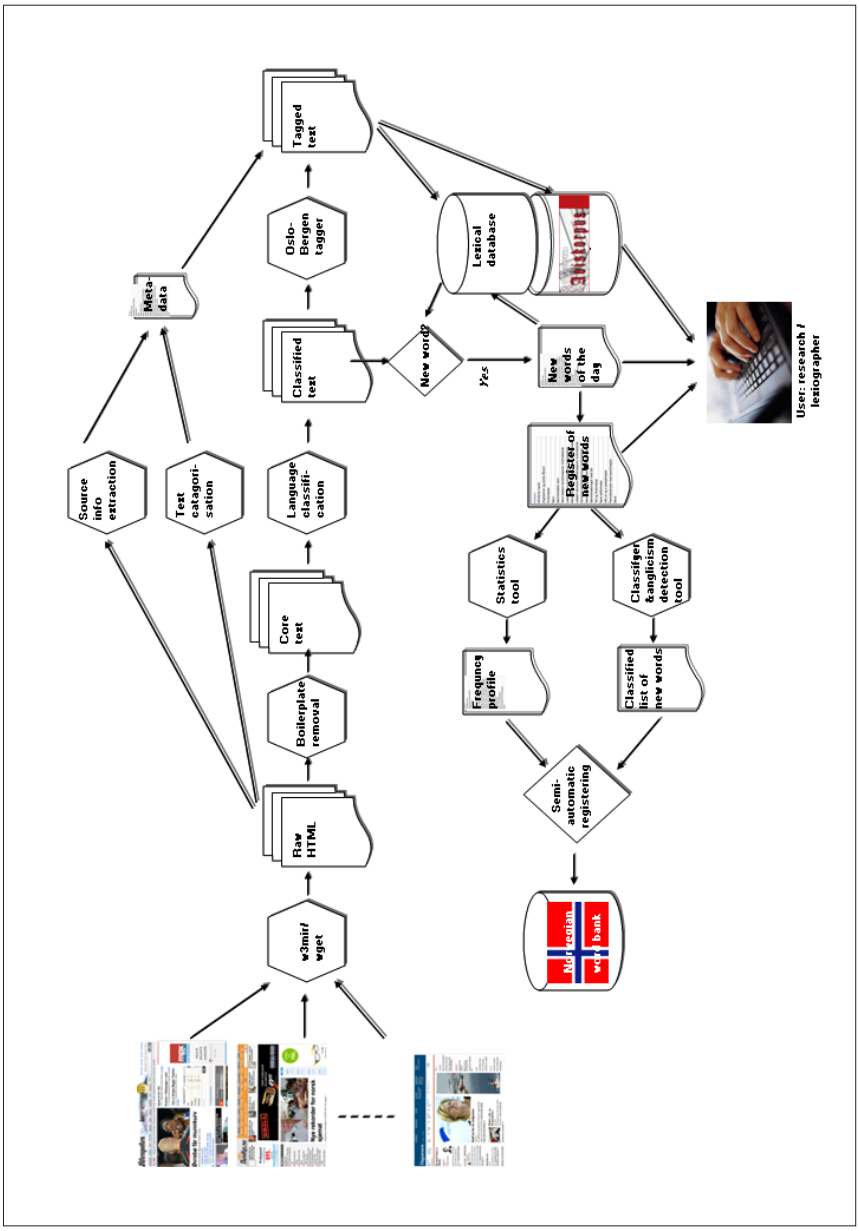
En utfordring ved bruk av maskinelle metoder for nyordsekserpering er å finne frem til relevante ord som er kandidater for ordlisteoppføring blant de svært mange nye ordformene som daglig fremkommer. Nedenfor beskriver jeg ulike moduler som inngår i korpusbasert nyordsarbeid. Dette omfatter et system for å finne frem til dagens nye ordformer, for å identifisere og analysere nye anglisismer som *crew* og *blogg*, og for å identifisere flerordsuttrykk som *due diligence* og *easy listening* blant anglisismene og blant nyordene mer generelt.

Norsk aviskorpus

Norsk aviskorpus er et nettbasert korpus som inneholder avistekst på bokmål og nynorsk. Tekstinnsamlingen har foregått siden 1998, og dette omfattende materialet består per desember 2009 av cirka 800 millioner løpende ord, og er dermed den desidert største søkbare og annoterte norske korpus. En serie dataprogrammer settes i gang automatisk hvert døgn, og materialet vokser hver eneste dag. *Norsk aviskorpus* er altså et monitorkorpus, på linje med for eksempel *Bank of English*, som har vært brukt som grunnlag for COBUILD-ordbøkene. Den daglige veksten er cirka 230 000 løpende ord, og blant dem finnes det mange ordformer som ikke har vært registrert tidligere, i snitt cirka 1 300 per dag. Samlet utgjør dette en verdifull kilde til informasjon om det norske språkets utvikling, nyordsdanning, bruken av lånord og språklige bruksmønstre mer generelt.

Systemets prosessering av data er vist som et flytdiagram i figur 1. De viktigste elementene i systemet er følgende:

1. Programmet *w3mir* laster ned den dagsaktuelle versjonen av et utvalg norske nettaviser.
2. Et program ekstraherer kjerneteksten, det vil si avisartiklenes overskrift, ingress, brødtekst og billedtekst. Programmet forkaster annonsetekst, navigeringsmenyer, metatekst, html-kommentarer og lignende.
3. Kjerneteksten blir automatisk klassifisert som bokmål eller nynorsk (eller engelsk, som forkastes).
4. Teksten blir analysert og merket med morfosyntaktisk informasjon ved hjelp av Oslo–Bergen-taggeren.
5. Programmet legger den merkete teksten inn i korpuset og gjør denne søkbar.



Figur 1. Systemflyt for korpusbasert nyordsekserpering

6. Alle ordformer i de nye tekstene blir sjekket mot allerede registrerte ordformer i korpuset.
7. Programmet lager en liste over ord som ikke var registrert fra før. Disse blir lagt inn i den totale ordlisten.
8. På grunnlag av bruksfrekvens blir et utvalg av de mest aktuelle ordene senere manuelt klassifisert og lagt inn i databasen *Norsk ordbank*.

Nyordsekserpering

Det er viktig å merke seg at et "ord" i dette systemet er et teknisk begrep som er maksimalt vidt definert som en sekvens av grafemer (tegn) mellom to mellomrom. Dessuten er et "nyord" definert som et ord som ikke fins i stor, akkumulert referanseordliste, som består av alle tidligere registrerte ord. Den akkumulerte listen er omfattende og består per desember 2009 av cirka 3,9 millioner ordformer, deriblant hele fullformsordlisten til *Bokmålsordboka*.[2]

Tabell 1. Eksempler på nyord fordelt på ulike kategorier

Kategori	Eksempel	#	%
Allmenne nyord/stavefeil	*tidsklemma*, *pingle*, *ektremistisk*	895 336	46,0
Anglisismekandidater	*whistleblower*, *blogg*, *subprime*	104 217	5,4
Forkortelser	*omg.*	2 838	0,1
Navn	*al-Duwasa*, *CanJet*, *Olsweek*	477 828	24,6
Sammensatte navn osv.	*Pan-Arctic*	104 960	5,4
Sammensetninger m. bindestrek	*e-meter-tester*, *blokk-bleik*	212 413	10,9
Sammensetninger m. annen markering	*Fabian/Wikimedia*	62 510	3,2
Tall og forkortelser m. tall	*KOMMUNE26*	16 802	0,0
Rene tall	*88,500*	358	0,9
URL-er og e-postadresser	*kickoff.com*	22 968	1,2
Skrot	*Rekdal.-*	44 790	2,3
TOTALT		1 945 020	100

2 I den versjon som ble benyttet i prosjektet SCARRIE (http://ling.uib.no/desmedt/scarrie/).

Tabell 1 viser eksempler på nyord etter denne definisjonen og fordelt på ulike kategorier. I den første kategorien finner vi ord som ikke inneholder noe bestemt ortografisk særtrekk, som versaler, tall, bindestrek eller lignende. Vi ser at blant ordformene som inngår her finnes det både reelle neologismer[3] – inklusiv nyformativer som *pingle*, og nye sammensetninger som *tidsklemma* – men også ikke tidligere registrerte stavefeil som *ektremistisk*. Denne gruppen inneholder i underkant av halvparten av nyordene.

En del ord blir maskinelt klassifisert som anglisismekandidater, og blant dem finner vi *whistleblower*, *blogg* og *subprime*. Om lag 5 prosent av ordene havner i denne kategorien. Hvordan denne klassifiseringen foregår, er beskrevet mer utførlig i avsnittet nedenfor.

En del nyordskategorier har et ortografisk særtrekk som brukes som klassifiseringsgrunnlag og som gjør dem mindre aktuelle for leksikografiformål. Om lag 10 prosent av ordene er produktive sammensetninger med bindestrek. Avisspråket inneholder forholdsvis mange navn, og anslagsvis 30 prosent av ordene er klassifisert slik på ortografisk grunnlag (inklusiv sammensatte navn). Dessuten forekommer forkortelser, rene tallformater, URL-er og e-postadresser blant nyordene, foruten en ubetydelig andel skrotord (2,3 %).

Klassifikasjon av anglisismer

Det er velkjent at mange engelske importord brukes i norske aviser og i norsk språk generelt. Følgende eksempler viser bruk av nye anglisismer i norsk avisspråk:

(1) Prøv den i en *smoothie*.
(2) De oppdaget en mann i hvit *cap*, mørk jakke og mørk bukse.
(3) Det finnes også en egen kategori for *podcast*.
(4) Avtalen er forutsatt av *due diligence*.
(5) En *übercool snowboard-dude* med franske foreldre fra Stavanger.
(6) Prins Charles i passiar med Englands *hotteste babe*, Catherine Zeta-Jones.
(7) Jeg vokste opp i en forstad, et *døllt* sted, derfor liker jeg aktivitet.

Av flere årsaker har vi valgt å se særskilt på disse ordene i prosjektet. En kartlegging av hvilke nyord som er engelske eller av annet fremmed opphav vil kunne

3 Se Ruth Vatvedt Fjelds og Lars Nygaards artikkel i denne antologien.

gi informasjon som er relevant for både allmenne, leksikografiske og språkpolitiske formål og vil kunne gi svar på spørsmål som

- Hvor omfattende er engelskens påvirking?
- Er den konstant eller varierende over tid? Har den økt de siste ti årene?
- Hvordan varierer engelsk ordtilfang i henhold til emnekategori? Er det for eksempel flest engelske ord innen tekster knyttet til sport eller informasjonsteknologi?
- Har andre variabler betydning for tilfanget av engelske ord. Er det forskjeller mellom ulike aviser, teksttypologisk kategorier eller forfattere?

Å fastslå om et nyord er av engelsk eller annen fremmed opprinnelse har også en betydning for leksikografi. Denne informasjonen vil kunne være nyttig for ordbøker med oppføring av etymologi eller ved utgivelser av anglisismeordbøker. Videre er det viktig å kartlegge tilfanget av anglisismer i fagspråklig arbeid, for eksempel for å vurdere innen hvilke fagområder hvor risikoen for domenetap er størst.[4]

Korpuset kan brukes som grunnlag for ikke bare å identifisere slike nyord, men også å studere deres morfosyntaktiske egenskaper, og dermed fremskaffe informasjon som også trengs når nyordene blir gjenstand for manuelt ordboksarbeid. For eksempel varierer et ord som *cap* ved at det enten følger vanlig norsk bøyningsmorfologi (*cap–capen–caper–capene*) eller at det kan ha den engelske flertallsformen som stamme (*caps–capsen–capser/caps–capsene*) (Graedler 1995). Et adjektiv som *trendy* kan ha endelsen *-e* i bestemt form, *trendye*, tross observasjonen til Jan Terje Faarlund m.fl. (1997: 375–376) om at slike adjektiv ikke blir bøyd i genus og tall. Et verb som *rule* viser tegn til to alternative bøyningsmønstre, enten slik som *kaste* (*å rule*, *ruler*, *rulet*, *har rulet*) (Graedler 1995) eller etter mønster av *lyse*, observert i verbformen *rulte*. Dette er eksempler på informasjon som lar seg trekke ut av et korpus og som vil være relevant for leksikografen som skal bearbeide disse ordene.

I prosjektet har vi utviklet et språkbehandlingsverktøy som identifiserer sannsynlige anglisismer (anglisismekandidater) blant nyordene. Denne modulen bruker en hybrid metode som omfatter n-gramstatistikk, regulære uttrykk og ordboksoppslag (Andersen 2005). Anglisismer som forekommer i norsk er i ulik grad preget av engelskhet, ortografisk sett. En del ord er lett gjenkjennelige anglisismekandidater fordi de har en fremmedartet ortografi, slik som *crew*, *quiz*, *comeback*, *chat* og *shotsene*. Slike ord vil bli klassifisert som anglisismer

4 Jf. språkpolitiske dokumenter som *Norsk i hundre!* og *Mål og meining*.

på grunn av de tegnsekvenser ordene inneholder. Ordet *crew* inneholder fem bigram:[5]

^c | cr | re | ew | w$

Fordi flere av ordets bigram er ikke-norske blir ordet klassifisert som en anglisismekandidat på statistisk grunnlag. Andre ord er forholdsvis lett klassifiserbare fordi de inneholder et avledningsmorfem som kun forekommer i engelsk, som *reality,* eller *temptation*. Slike ord klassifiseres som anglisismekandidater ved hjelp av regulære uttrykk.

Imidlertid finnes det mange anglisismer som ikke inneholder engelskspesifikk ortografi, og disse er mer kompliserte å klassifisere. Eksempler er verbene *date* og *rule*, som ikke inneholder noen ikke-norske tegnsekvenser; det er ikke noe særengelsk ved bigrammene ^d | da | at | te | e$. I slike tilfeller brukes oppslag i elektroniske ordlister som grunnlag for klassifikasjon. Den norske ordlisten inneholder fullformer fra *Bokmålsordboka*, mens den engelske er frekvensordlisten fra *British National Corpus* (BNC). Et ord som *date* klassifiseres som anglisismekandidat fordi det forekommer i den engelske ordlisten men ikke i den norske.

Anglisismer med norsk ortografi kan være oversettingslån, slik som *nedlaste* fra engelsk *download*, eller ord med norvagisert stavemåte, som *døll* (*dull*) og *sørvis* (*service*). Disse vil ikke bli maskinelt klassifisert som anglisismer, men vil bli fanget opp av modulen for nyordsklassifisering.

Engelskspråklig påvirkning kan også innebære at et eksisterende norsk ord får en ny betydning (neosemantisme) på grunn av et forekommende kognat i engelsk. Dette gjelder for eksempel verbet *adressere* i betydningen 'behandle, ta opp', som i *å adressere et problem*, eller den nye bruken av *karakter* i betydningen 'rollefigur'. Disse er atskillig mer problematiske å identifisere maskinelt. En annen utfordring er syntaktiske endringer, som verbet *disse*, som nå brukes som et transitivt verb, som i setningen *Ingen disser Anne B. Ragde!*. Denne bruken antas å være påvirket av engelsk *dis* (*disrespect*). Foreløpig er det ikke utviklet et opplegg for systematisk registrering av slike endringer, men det kan la seg observere på grunnlag av endringer i ordenes kollokasjonsmønstre (Sinclair 1991, Biber 2009) eller syntaks. Det er planlagt en videreutvikling av klassifikasjonsmodulen for å fange opp slike endringer, blant annet ved eksperimentering med alternative metoder som TIMBLE-basert maskinlæring og tilgjengelig programvare for språkgjenkjenning.

5 Symbolet ^ angir ordbegynnelse og $ angir ordslutt.

Totalt finner systemet 128 588 ordformer som antas å være anglisismer, av i alt 1 469 925 unike ordformer i korpuset, det vil si om lag 8,7 prosent. Tabell 2 viser toppen av korpusets frekvensordliste og de vanligste anglisismer i materialet.

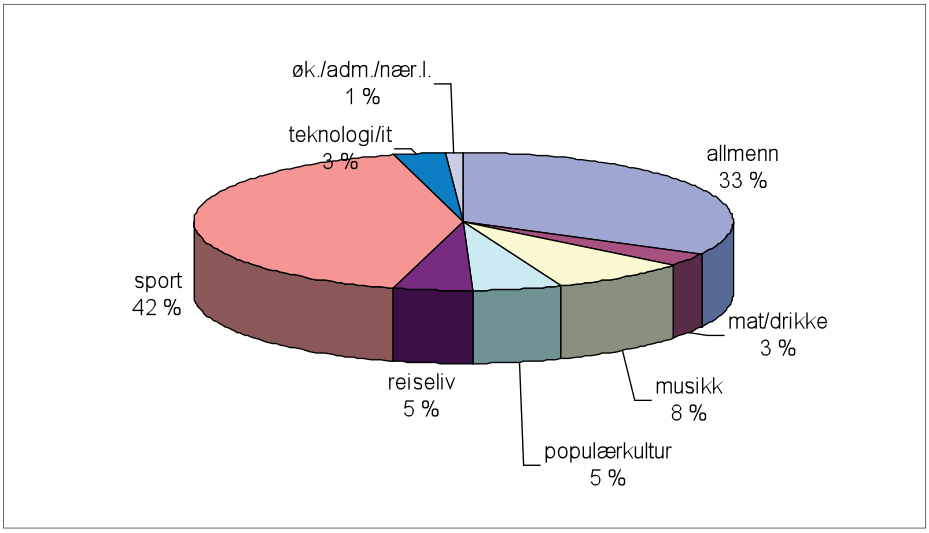
Tabell 2. Hyppigste ord og hyppigste anglisismer i Norsk aviskorpus

i	24 850 781	scoret	72 520
og	19 158 595	keeper	28 288
er	14 432 730	sex	20 711
til	11 937 335	manager	19 606
på	11 930 944	scoring	18 042
som	11 530 249	score	17 348
det	10 948 982	verdenscupen	15 564
å	10 475 491	scoringer	13 337
av	10 146 256	rock	12 199
en	10 040 356	ishockey	9 957
for	10 029 477	scorer	9 828
at	9 562 804	toppscorer	9 823
har	8 808 917	comeback	9 482
med	8 462 376	cupen	9 095
ikke	6 147 810	jazz	8 552
de	6 031 542	headet	8 296
om	5 095 203	mobbing	7 524
den	4 948 322	scoringen	6 862
et	4 437 130	corner	6 611
fra	4 217 612	keeperen	5 874
var	4 027 751	canadiske	5 289
han	3 608 843	cupfinalen	5 188
seg	3 420 451	matchvinner	5 175
ble	3 017 390	back	4 900
sier	2 959 084	that	4 880

En manuell gjennomgang av de 1 500 hyppigste anglisismene viser at anglisismer i korpuset særlig forekommer innenfor visse emneområder. I særdeleshet dreier dette seg om følgende emnekategorier:

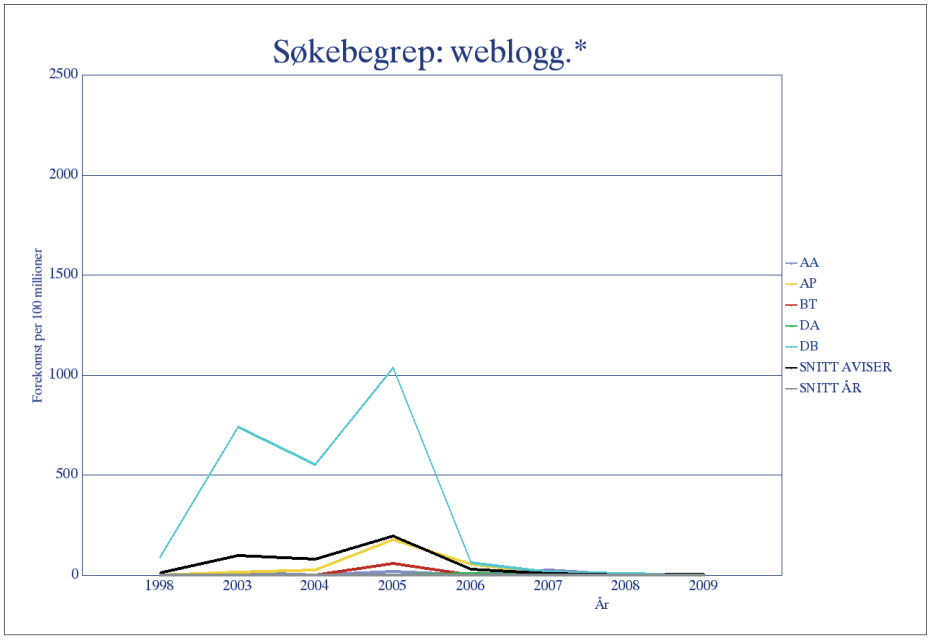
- Reiseliv: *campingplassen, campingvogner, sightseeing, charter, cruiseskip/-et, booket*
- Mat/drikke: *squash, cola, bacon, whisky, pizza*
- Teknologi/IKT: *blogg/-er/en, iPhone, mail*
- Øk./adm./næringsliv: *business, shipping, offshore*
- Sport: *score/-r/-t, scoring/-er, headet, cupen, corner, match, volley*
- Allmenn: *sexy, hint, servicen, must, audition, tagging, matching*
- Musikk: *musical, rocka, medley, rockeband, soul, country*
- Annen populærkultur: *science, fiction, trailer/-e/-en, action, thriller*

En foreløpig utregning viser fordelingen per emnekategori som vist i figur 2.

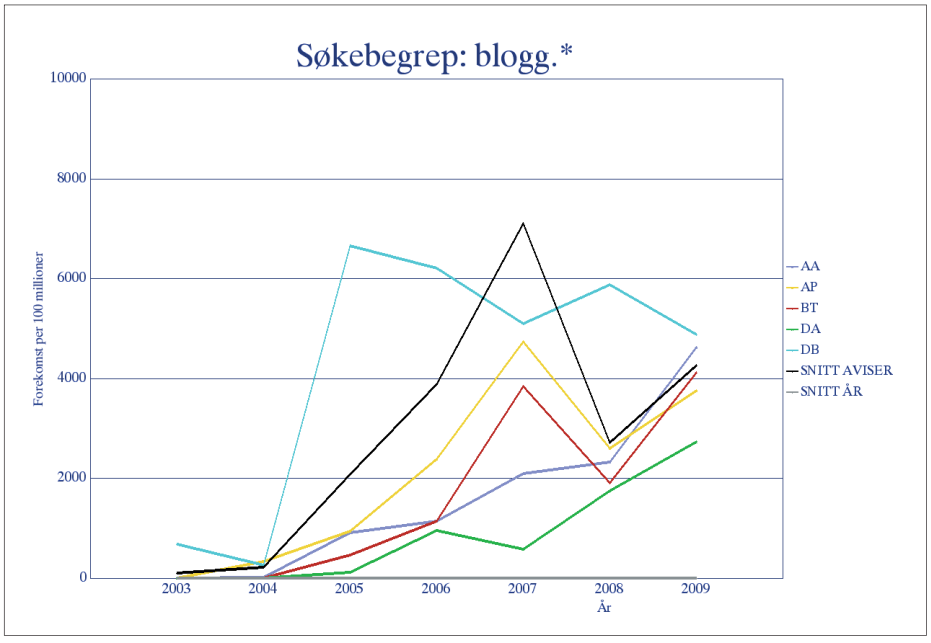


Figur 2. Fordeling av anglisismer på emnekategori

Ved hjelp av korpusets grensesnitt er det mulig å ekstrahere opplysninger om enkeltordenes bruk men også om deres frekvensutvikling over tid. Dette kan presenteres enten tabellarisk eller som grafer som genereres på direkten. Eksempler på dette er vist i figur 3–4. Figur 3 viser at anglisismen weblogg var forholds mye brukt i 2000-tallets første halvdel, mens figur 4 viser at det er etter hvert erstattet av det langt vanligere ordet blogg, som har etablert seg som den rådende termen for dette nye sosiale mediet. Slike frekvensopplysninger er avgjørende for hvorvidt nyord er aktuelle for oppføring i ordbøker (jf. Fjeld & Nygaard i denne publikasjon).



Figur 3. Frekvensfordeling over tid for ordet weblogg (høyretrunkert søk)



Figur 4. Frekvensfordeling over tid for ordet blogg (høyretrunkert søk)

Kollokasjonsanalyse

En del anglisismer består av flere ord, slik som *due diligence*, *easy listening*, *break even*, og *Get a life!*, som alle forekommer i aviskorpuset. Det er nødvendig med egne rutiner for å håndtere slike leksikaliserte flerordsuttrykk. I prosjektet er det utviklet et system for å identifisere kollokasjoner, det vil si sekvenser av ord som har en sterk tendens til å samforekomme. Å identifisere flerordsuttrykk er generelt av stor betydning for ulike formål innen leksikografi, terminologi og språkteknologi. Dette kan bidra til korrekt segmentering av fraseologiske enheter (*tilslørte bondepiker*, *guri malla*), ekstraksjon av fagterminologi (*ulcerøs kolitt*, *notarius publicus*), segmentering av anglisismer (*easy listening*) og automatisk prosessering av språk. Prosjektets medarbeidere har produsert n-gramstatistikk for hele korpuset og rangert disse ved hjelp av ulike statistiske assosiasjonsmål (*association measures*) for å identifisere sterke ordforbindelser (Andersen & Lyse 2009). Vi har også evaluert ulike assosiasjonsmål og hvorvidt de er egnet som grunnlag for analyse innen leksikografi og terminologi. Tabell 3 viser de toordssekvenser (bigram) som er høyest rangert når vi bruker ulike assosiasjonsmål.

Resultatene viser at det er store forskjeller på hvilke toordssekvenser som er høyest rangert, og på assosiasjonsmålenes egnethet til å finne leksikaliserte fraser. Assosiasjonsmålene *t-score*, *local-MI* og *likelihood ratio* gir høy rangering av ordsekvenser som ikke er leksikaliserte, slik som *det er*, *til å*, *for å*, *i en*, *å komme* og *millioner kroner*. Disse anses som lite relevante for leksikografi-/terminologiformål, men det bør bemerkes at leksikaliserte elementer som *i tillegg*, *i går* og *i fjor* er høyt rangert bruk av disse målene.

Andre assosiasjonsmål gir høy rangering av leksikaliserte ordsekvenser, hvor enhetens betydning kan ikke utledes fra enkeltkomponentene. Dette gjelder assosiasjonsmålene *chi square corrected*, *z-score*, *z-score-corrected*, *odds-ratio-discriminative* og *pointwise-MI*. Hvis vi studerer den videre listen ser vi at disse målene gir høy rangering til fagspråklige termer som *anaerobe terskelen*, *eneggede tvillingene*, *honorære konsuler* og *amyotrofisk lateralsklerose*, men også at det forekommer svært mange flerordsanglisismer blant de høyt rangerte bigrammene, eksempelvis *lucky loosers*, *corned beef*, *practical jokes*, *slow starters*, *jumpers knee*, *consumer confidence*, *honky tonk*, *splendid isolation*, *due diligence*, *extreme makeover* og *danish dynamite*. Samlet sett viser dette at identifisering av flerordsuttrykk er en viktig forutsetning for maskinell identifikasjon av anglisismer.

En manuell gjennomgang av de 500 høyest rangerte bigram i henhold til *odds ratio*-utregningen viste at cirka 18,6 prosent av de høyt rangerte bigram-

Tabell 3. Evaluering av ulike assosiasjonsmål

Chi-squared corr	**Likelihood ratio**	**Odds-ratio**	**Pointwise MI**	**Local MI**
tredje kvartal	det er	fiskeridirektoratets	vitello tonnato	det er
eiendomsmeglerbransjens	til å	kontrollverk	vilkårsett	til å
boligstatistikk	for å	jumpers knee	skattefritaking	for å
raskast veksande	millioner	terje pedersen	varannan damernas	millioner
stridande partar	kroner	consumer confidence	unilaterally destroyed	kroner
gammelnorske kyrra	har vært	corned beef	twam asi	at det
fremkall brekninger	blant annet	cage aux	suvas bohciidit	å få
fjerde kvartal	å få	practical jokes	skrimmi nimmi	er det
cottage cheesen	at det	garam masala	rondo capriccioso	har vært
charge refers	år siden	prymnesium parvum	rollon rolloff	blant annet
straffeprosesslovens	er det	tyrannosaurus rex	rødøret terrapin	at han
paragraf				

z-score	**z-score corrected**	**t-score**	**Dice**	**jaccard**
yom kippur	fiskeridirektoratets	det er	bestanden	appelldomstol
wishful thinking	kontrollverk	til å	betraktelig	avviste
whiter shade	jumpers knee	for å	forberedelsene	bytrikkens linjer
whistle blowers	terje pedersen	at det	pågått	effektive redskapet
vox populi	consumer confidence	er det	historiske trilogi	generell
vackraste visan	corned beef	å få	inkluderer	nedbemanning
utsletti	cage aux	det var	forsikring	gryende allianse
respateksbord	practical jokes	har vært	innvandrere forbigås	helsepersonell
tschoka tschoka	garam masala	at han	kraftige fallene	vasker
tribal peoples	prymnesium parvum	at de	kriminelle drapsmenn	hjemlige filmmiljø
totalavholdsselskaps	tyrannosaurus rex		legene tror	illojale fremfor
ungdomsforbund			sivile	lands utenlandsgjeld
			sikkerhetsfolk	røyking utendørs
			sterk ryggrad	

mene var anglisismer, mot 4,2 prosent av alle bigram som forekommer. En annen viktig observasjon er at svært mange høyt rangerte ikke-anglisismer er importord fra andre språk, hvorav svært mange fra latin (jf. frasemer som *commedia dell'arte*, *abortus provocatus*, *solar plexus*, *annus horribilis*, *notarius publicus*, *lingua franca*, *tabula gratulatoria*, *tabula rasa* og *mea culpa*). Blant høyt rangerte bigram finnes også vanlige norske faste ordforbindelser, slik som *navns nevnelse*, *flammenes rov*, *rangen stridig*, *tenners gnissel* og *bange anelser*.

Eksempler som *cage aux* og *erat demonstrandum* viser for øvrig at man bør analysere de lengste n-grammene først, for å finne enheter som henholdsvis *cage aux folles* og *qui erat demonstrandum* som trigram.

Oppsummering

I denne artikkelen har jeg pekt på fordeler ved bruk av korpusbaserte metoder i leksikografiarbeid. *Norsk aviskorpus* er et omfattende og selvekspanderende korpus som kan brukes til kontinuerlig observasjon av utvikling av nye ord og endringer i ordenes bruksmønstre. En rekke dataverktøy inngår i dette systemet. Beskrivelsen har vist at maskinelle metoder kan brukes til å ekserpere nyord, til å identifisere faste ordforbindelser og til å identifisere importord blant nyordene som forekommer. Samlet sett vil disse verktøyene kunne forenkle arbeidet med nyord og redusere behovet for manuelt arbeid.

Det må imidlertid understrekes at de fremgangsmåter som er beskrevet er et verdifullt supplement til tradisjonelt leksikografisk arbeid, men erstatter ikke manuelt arbeid. Korpuset og nyordslistene er et svært omfattende, og det er fortsatt behov for kvalitetssikring av komponentene som inngår.

Litteratur

Andersen, Gisle, 2005: Assessing algorithms for automatic extraction of anglicisms in Norwegian texts. I: Corpus Linguistics 2005. http://www.corpus.bham.ac.uk/pclc/

Andersen, Gisle & Lyse, Gunn Inger, 2009: Om vispet krem og ubeskyttet sex: flerordsuttrykk i Norsk aviskorpus. I: MONS 13. Trondheim. Otrykt.

Atkins, Sue, 1993: Tools for computer-aided corpus lexicography: the Hector project. I: Acta Linguistica Hungarica 41. S. 5–72.

Biber, Douglas, 2009: A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. I: International Journal of Corpus Linguistics 14. S. 275–311.

Church, Kenneth & Hanks, Patrick, 1989: Word association norms, mutual information and lexicography. I: ACL 27th Annual Meeting 76–83. Vancouver.

Faarlund, Jan Terje & Lie, Svein & Vannebo, Kjell Ivar, 1997: Norsk referansegrammatikk. Oslo.
Graedler, Anne-Line, 1995: Morphological, semantic and functional aspects of English lexical borrowings in Norwegian. Oslo.
Kilgarriff, Adam, 1998: The hard parts of lexicography. I: International Journal of Lexicography 11. S. 51–54.
Munat, Judith, 2007: Lexical creativity, texts and contexts. Amsterdam/Philadelphia.
Renouf, Antoinette, 1987: Corpus development. I: Looking up, red. John McH. Sinclair. London/Glasgow. S. 1–15.
Renouf, Antoinette, 2007: Corpus development 25 years on: from super-corpus to cyber-corpus. I: Corpus linguistics 25 years on, red. Roberta Facchinetti. Amsterdam/New York. S. 27–49.
Sinclair, John McH. (ed.), 1987: Looking up. London/Glasgow.
Sinclair, John McH., 1991: Corpus, concordance, collocation. Oxford.
Stubbs, Michael, 1995: Collocations and semantic profiles: On the cause of the trouble with quantitative studies. I: Functions of Language 2. S. 23–55.
Summers, Della, 1993: Longman/Lancaster English language corpus – criteria and design. I: International Journal of Lexicography 6. S. 181–208.