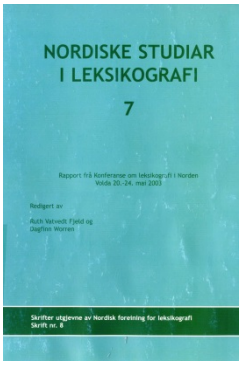
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Nynorskkorpuset vid Norsk Ordbok 2014: Integrering med redaktionsarbete
Forfatter:	Daniel Ridings
Kilde:	Nordiske Studiar i Leksikografi 7, 2005, s. 315-325 Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. maj 2003
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

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.

Daniel Ridings

Nynorskkorpuset vid Norsk Ordbok 2014: Integrering med redaktionsarbete

This paper describes the collection, construction and integration of the language corpus, *Nynorskkorpuset ved Norsk Ordbok 2014*, with the editorial work of the lexicographers at the same project. It touches on legacy data that already existed and explains how new data was collected and prepared for a corpus access application and for integration with the editing software used by the lexicographers. The techniques include an implementation of the TEI system for tagging text to corpus access through a relational database. It also exemplifies some of the methods being developed for making access to large data collections easier, in particular statistical methods for identifying collocations and semantic groups.

Introduktion

Denna presentation kommer att redovisa tre aspekter av korpusarbetet inom ramen för Norsk Ordbok 2014:

- Korpusuppbyggnad
- Applikationer som används för att ge tillgång till korpusen
  - Integration med Metaordboken
- Integration med Metaordboken
- Utvidgningar med metoder från datalingvistik

Det handlar om ett samarbete mellan *Enhet for digital dokumentasjon* (EDD) och projektet *Norsk Ordbok 2014* (NO 2014) vid Universitetet i Oslo. EDD ansvarar för att systematisera och implementera de moderna arbetsmetoder som används inom NO 2014 med syftet att korta ned produktionstiden och bevara arbetsinsatserna på ett sådant sätt att de kan återanvändas i andra sammanhang.

Korpusuppbyggnad

En korpus för ett språk skapas för bestämda syften och det finns olika typer av korpusar beroende på vilka syften användaren har. Allmänlingvister har helt andra behov än översättare. De förra vill bemöda sig om en allmän- eller referenskorpus. De senare har oftast större glädje av en textsamling som är mer domänspecifik. Allmänlingvister vill kunna undersöka ett brett spektrum av fenomen medan en översättare oftast är mer intresserad av terminologi för ett specifikt ämne. Den senare blir mindre besviken på vad man kan hitta i en korpus än den förra eftersom ingen korpus kan beskriva ett språk i dess

helhet. Samtidigt finns det få korpusar, oberoende av hur små de är, som inte kan bidra med något man inte kände till förut (Fillmore 1992).

Lexikografer utgör inte en homogen samling av yrkesmänniskor. Somliga arbetar med tvåspråkiga ordböcker medan andra arbetar med ett enda språk i ett modernt samhälle. En korpus för dessa måste motsvara olika krav. Den som arbetar med tvåspråkiga ordböcker vill vara noga med att inte för många översättningar kommer från samma översättare. Den som arbetar med moderna språk vill inte blanda i för många romaner från tidigare århundraden. En korpus ska motsvara den språkformen som lexikografen arbetar med.

Men det finns flera typer av lexikografer, sådana vars arbete representeras av *Svenska Akademiens Ordbok* (SAOB), *Woordeboek van die Afrikaanse Taal* (WAT), *Het Woordenboek der Nederlandsche Taal* (WNT) och Norsk Ordbok. Dessa ordböcker gör anspråk på att beskriva sitt respektive språk under en lång period. Alla ord som har använts under denna period, även om de är inaktuella idag, ska bearbetas. En korpus för dessa arbeten ska i princip bestå av allt som någonsin har skrivits eller talats under den aktuella perioden. Ordböckerna är allomfattande och ingenting är av «enbart historiskt intresse.»

Norsk Ordbok 2014

Norsk Ordbok ska dokumentera nynorska. Den ska beskriva nynorska idag, nynorska i historien och dessutom ska den dokumentera dialekterna. I detta arbete finns en bestämd plats för en korpus.

Vid Institutt for nordistikk og litteraturvitskap, Universitetet i Oslo, finns stora samlingar med nynorska och dialektala belägg från runt om i landet. Många av de viktigaste skriftkällorna har genomarbetats av frivilliga under många års tid. Dessa finns som traditionella «slips», papperskort, och är tillgängliga i institutionens omfattande kartotek. Det finns flera miljoner sådana slips och de täcker i synnerhet det historiska arvet. Däremot finns inte samma tillgång till modernt bruk. En förutsättning för den nynorska korpusen var att den skulle fylla denna lucka vad gällde nyare material och inte duplicera redan befintliga resurser. Denna var den grundläggande principen för att bygga upp en nynorsk korpus.

Tillgängligt material

Under nästan hela 1990-talet pågick ett omfattande arbete med att sätta moderna databehandlingsmetoder i bruk på en rad viktiga materialsamlingar inom språk och kultur i Norge. Arbetet gick under namnet «Dokumentasjonsprosjektet» (http://www.dokpro.uio.no). Lexikografi var ett prioriterat område för projektet av flera skäl, men inte minst på grund av de stora kartoteksamlingarna och insikten om att ett så viktigt arbete som att skriva ordboksmanus borde utnytta modern teknik för att kunna återanvändas på ett rationellt sätt. Ordböcker behöver nämligen ständigt uppdateras och den stora satsningen som ett ordboksarbete innebär är en del av det nationella kulturarvet som inte får försvinna.

Som ett led i Dokumentationsprojektets arbete för lexikografi började man samla texter på nynorska. Detta resulterade i lite mer än 2,5 miljoner löpande ord fördelade på ca 9.500 sidor text. Det mesta av materialet var äldre klassiska verk för vilka upphovsrätten hade löpt ut.

Ungefär samtidigt ledde Ruth V. Fjeld (2002) ett forskningsprojekt för att bygga upp en korpus över bokmål. Även om det inte var huvudsyftet skapade projektet en bra början på även en nynorsk korpus. Utöver en bokmålskorpus, samlade projektet ca 1,5 miljoner löpande ord fördelade på drygt 3.700 sidor. Dessa texter representerade ett modernare material från samtida publikationer som *Syn og Segn*, moderna romaner och faktaböcker och texter från tidningar (jfr. Runde 2000:24 och Almenningen 2001:11).

Textformat

Korpusar är dyrbara resurser. Det tar mycket tid att ta fram en modern korpus över ett språk. Därför är det viktigt att en korpus struktureras efter kända kriterier så att nästa forskningslag kan återanvända materialet med minsta möjliga problem. Det är bättre att material återanvänds än att hjulet återuppfinns.

En standard för att uppmärka texter är en förutsättning för att kunna enkelt återanvända resurser. LE2-4017-10379 PAROLE (Preparatory Action for Resources Organization for Language Engineering) var ett EU-projekt från 1996 till 1998 som tog fram korpusar på 14 språk och lexika på 12. Alla språk höll sig till en gemensam standard vad gäller inkodning av både korpusarna och lexikonen. Korpusstandarden (Ridings 1996) var en utvidgning av *Text Encoding and Interchange* (TEI) med ett öga på *Corpus Encoding Standard* (CES) från *Expert Advisory Group on Language Engineering Standards* (EAGLES) (Sperberg-McQueen and Burnard 1994; Ide and Veronis: 1996).

En del av det befintliga materialet för den nynorska korpusen var inkodad med HTML, en del med SGML-inspirerad uppmärkning och en del hade ingen uppmärkning alls annat än en CES textheader och sidbrytningar. Det bestämdes att korpusen skulle uppmärkas och parsas med en TEI/PAROLE DTD och allt befintligt material skulle anpassas till den standarden. Det gör att alla som har rutiner som kan fungera med TEI kan arbeta med korpusen.

Nyanskaffning

Det var en målsättning att den nynorska korpusen skulle bestå av minst 30 miljoner ord fördelat enligt de objektiva kriterier som användes inom PAROLE: böcker, tidskrifter och tidningar. En korpusgrupp bildades för att planera och genomföra detta arbete. Utöver denna grupp, som fungerar som en styrgrupp, bildades en mindre grupp för att hantera moment som inte kunde göras automatiskt.

Diverse kontakter hade tagits tidigare med förlag och tidningar. Dessa kontakter återupplivades och man träffade avtal med veckotidningen *Dag og Tid*, dagstidningen *Firda* och *Norsk Barneblad* för att få kontinuerliga leverenser av moderna nynorska texter av tidnings- och tidskiftstexttyper. Ett avtal har också upprättats med bokförlaget Det Norske Samlaget i Oslo. Alla nya böcker skickas till NO 2014 så snart Samlaget har fått godkännande från författarna. Den här processen, att få en författares godkännande, är nu en del av Samlagets rutiner. Det innebär att de flesta böckerna kommer till NO 2014 samtidigt som de går till sista instansen i Samlagets produktionskedja. Det handlar om såväl romaner som läroböcker och facklitteratur. Utöver böckerna kommer alla nummer av *Syn og Segn*.

På våren 2003 uppgår den nynorska korpusen till drygt 13 miljoner ord som är i produktion och 20 miljoner som väntar på att införlivas i korpussystemet.

Korpustillgänglighet

Att ha en stor korpus är inte mycket värt om man inte kan komma åt den. En beprövad metod, ända sedan medeltiden, för att studera texter systematiskt är konkordanser. Lexikografer har gjort det i alla år och de spelade en nyckelroll i COBUILD (Sinclair 1987).

Applikationer

Det finns två typer av applikationer som lexikograferna kan använda för att undersöka korpusen: en som erbjöds via webb och en annan som är integrerad med redigeringsprogrammet som används för att skriva artiklar i ordboken. Den förra kom till först därför att redigeringsprogrammet byggs upp parallellt med korpussystemet. Den webbaserade versionen kom också till för att pröva ut funktionaliteten och ge användarna ett tillfälle att påverka hur rutinerna skulle se ut i det färdiga systemet för artikelförfattande.

I princip vill man göra fyra saker när man som lexikograf söker i en korpus: man vill söka på ett ord, fras eller delar därav, man vill granska en resulterande konkordans, man vill ibland få lite mer kontext än vad konkordansen ger och man vill välja ut belägg från korpusen för att användas som exempel i en ordboksartikel.

Systemet finns tillgängligt från NO2014:s hemsida, http://no2014.uio.no. Där kan man skriva ett ord, en fras eller delar av ett ord. I *figur 1* har delar av ett ord angetts, nämligen alla ord som börjar med prefixet 'sam'. Procentteckent, %, är ett «wild-card», dvs, det kan ersättas med vad som helst så länge ordet börjar med 'sam'.

Nackdelen med att söka på delsträngar såsom 'sam' är att man riskerar att få en oöverskådlig mängd tillbaka som resultat. Därför när man söker på en delsträng öppnas ett litet fönster med en lista över alla ord som passar sökkriteriet. Detta ser man i *figur* 1. Man kan använda det fönstret för att komma vidare till en konkordans. Man gör det genom att bläddra i listan och klicka på det ordet man är intresserad av.

Det lilla fönstret finns kvar och på en större skärm kan det flyttas ur vägen men ändå finnas tillgängligt för att välja flera ord ur listan. Hänvisningen till vänster av varje konkordansrad i *figur* 2 är förkortningen som används i *Norsk Ordbok* för referensen.

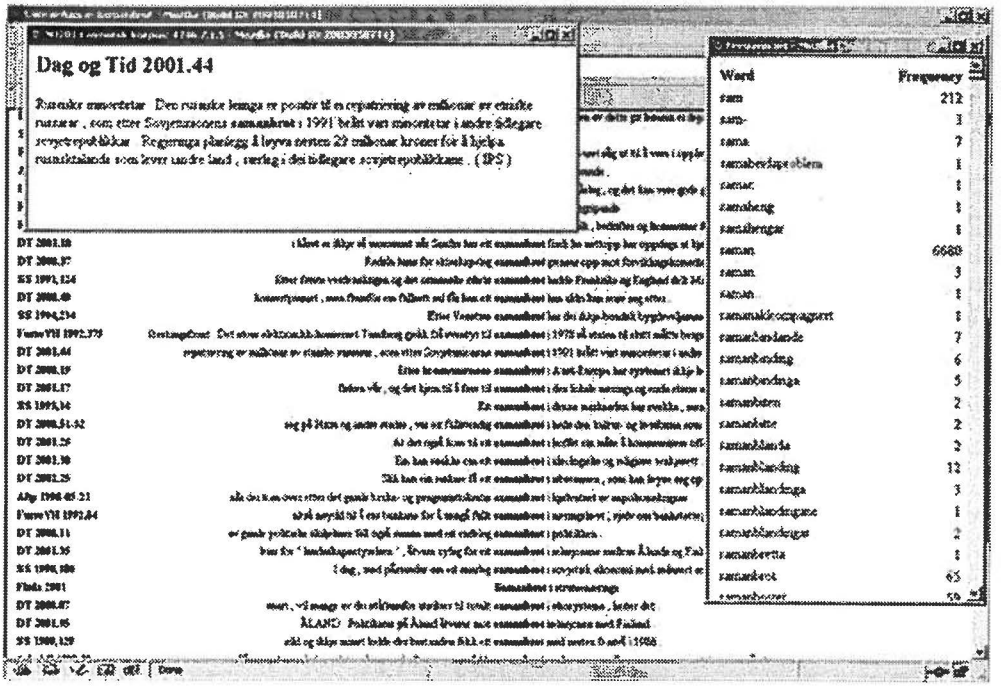
En kreativ användning av 'wild-cards' kan avhjäpa det faktum att korpusen ännu inte är lemmatiserad. Detta arbete pågår och beräknas vara färdigt i 2005. Under tiden kan man använda tecken som '%' (vad som helst, hur mycket som helst) och '_' (ett tecken, vilket som helst) för att söka på böjningsstammar, prefix och suffix.

På samma sätt som alla rader är aktiva i det lilla fönstret med enstaka ord och kan användas för att visa fram en konkordans, är alla enstaka rader i en konkordans också aktiva, och om man klickar på en rad, får man upp en större kontext. Detta illustreras i *figur* 3.

Det finns möjlighet att få ännu mer kontext om det inte räcker. Systemet bygger på SGML:s dokumentträd och går högre och högre upp i hierarkin: från en mening till ett stycke, vidare till avsnitt och därifrån till kapitel och slutligen till hela dokumentet men då är det stopp. I princip skulle man kunna fortsätta upp ett steg till, men då får man hela korp-

Figur 1: När man söker med ett «wild-card», i detta fall «%», får man upp en lista med passande ord istället för en potentiell enorm konkordans.

Figur 2: Genom att klicka på 'samanbrot' i det lilla fönstret visas en motsvarande konkordans i det stora fönstret.



Figur 3: Genom att klicka på en enskild konkordansrad får man ett fönster med mer kontext.

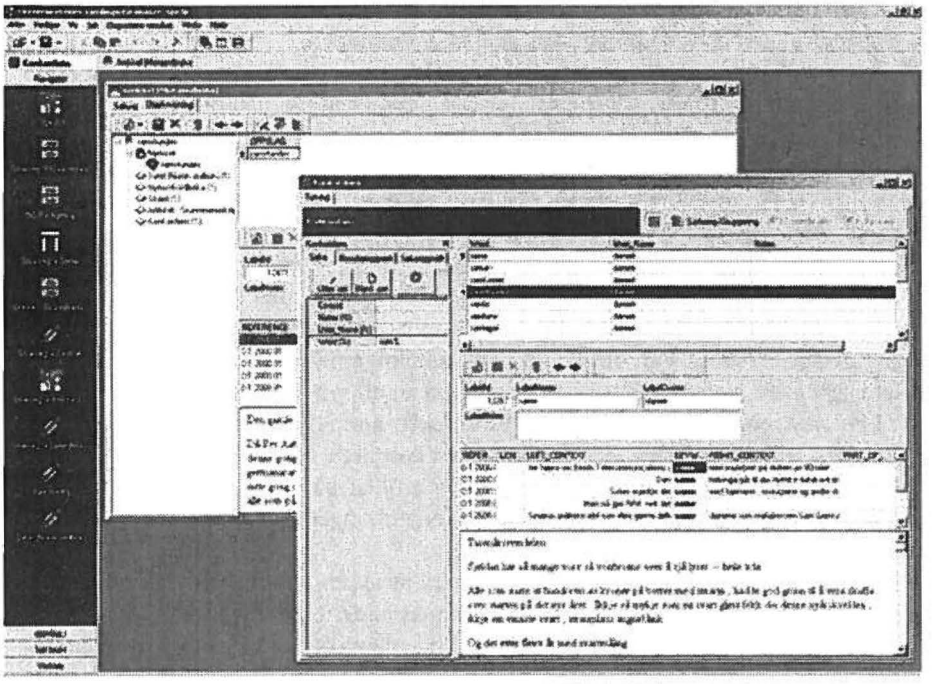
usen och det är lite väl mycket. Denna möjlighet är begränsad i den offentliga versionen för att undvika en situation där någon illasinnad kopierar en hel text.

Så långt den webbaserade versionen. Den uppfyller inte det sista kravet att integrera arbetet med redaktionsarbete så att en lexikograf kan flytta över ett belägg från korpusen till ett exempel i en ordboksartikel. För att bättre förstå hur detta har hanterats behöver man känna till hur *Enhet for digital dokumentasjon* (EDD) vid Universitetet i Oslo arbetar.

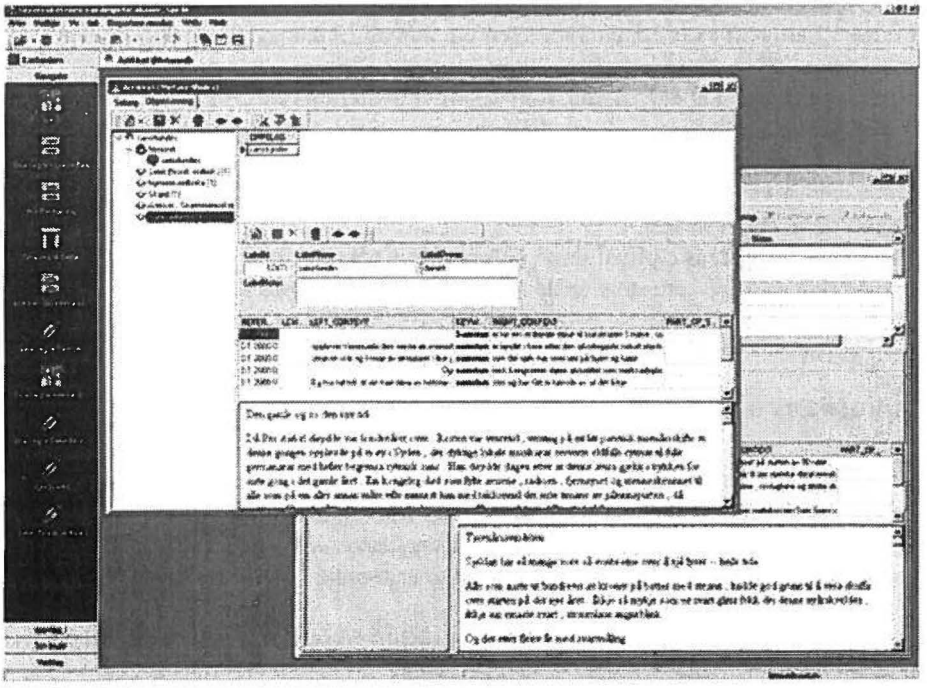
EDD har lång erfarenhet av att systematisera stora mängder material från allt från riksmuseer med arkeologi som sin specialitet till stora fotodatabaser. Allt lagras i en relationsdatabas varefter applikationer för att ge olika användare tillgång till sitt, och andras, material. Dessa applikationer är anpassade till Microsoft Windowsmiljön.

Ars longa vita brevis

EDD är involverad i flera projekt på riksnivå och alla är stora. Det finns en begränsad tillgång till utvecklingsresurser, dvs människor, och om man ska kunna leverera kvalitet, måste insatserna samordnas. Allt EDD sysslar med är information. Det kan vara ett föremål på ett arkeologiskt museum, zoologiskt museum, fotomuseum, texter eller ordböcker. Informationen modelleras och stuvas om till ett databassystem. Att göra fristående applikationer till varje projekt skulle snabbt nå gränsen för vad som låter sig göras med befintlig bemanning. Därför har man skapat en struktur med en enda bas, en METABASE, som är navet som håller ihop de olika ämnesområdena.



Figur 4: Här ser man en konkordans och större kontext för en rad.



Figur 5: Här ser man hur en konkordans har sparats undan i en artikel i metaordboken tillsammans med andra belägg för samma huvudord.

EDD:s system liknar ett bibliotek. Man söker bland böcker i det egna fackområdet men har samtidigt tillgång till lexika och upplagsverk från andra områden.

Till den överordnande metabasen har man utvecklat en motsvarande applikation, *felles-applikasjon*. Genom den får man tillgång till egna databaser samtidigt som man kan, med de rätta privilegierna, få tillgång till alla andra samlingar som EDD förvaltar.

I detta sammanhang är NO 2014:s lexikaliska databas och korpus två bland många samlingar. Situationen är lite mer komplicerad. Om man tillämpar samma synsätt på de lexikala samlingarna som man gör på EDD:s alla övriga samlingar finner man också en hierarki. Lexikografer arbetar med många materialkällor. De omfattande kartoteken har redan nämnts. Dessa har digitaliserats och inordnats i databassystemet. Lexikografen kan gå till det fysiska kartoteket eller man kan söka fram ett faksimil av samma kort från databasen.

I analogin med en *metabas* som binder ihop alla samlingar inom EDD:s verksamhetsområde, har det skapats en *metaordbok* som knyter samman all typ av information som ingår i en ordbok: kort med belägg från det fysiska kartoteket, normaliseringar, andra ordböcker, andra elektroniska beläggsamlingar och den nynorska korpusen, för att nämna några stycken (Grønvik 2000; Ore 2000). Lexikografen som skriver artiklar har tillgång till allt detta samtidigt som det arbetet man utför, en ny artikel till exempel, ingår i samlingarna och blir tillgängligt för andra.

I det här systemet ska korpusbelägg som har valts ut av en medarbetare ingå under ett huvudord i likhet med andra typer av belägg. Allt samlas under huvudordet och blir tillgängligt för artikelförfattaren, som inte behöver vara samma människa som har samlat in beläggen. Det man vill göra, i första hand, är att placera en konkordans över ett ord under dess huvudord så att artikelförfattaren kan använda det i sitt arbete med att dela upp språkexempel i betydelser (semantisk sortering). När artikeln är färdig finns all dokumentation för hur man kom fram till artikeln kvar i databasen även om inte allt fick utrymme i den färdiga artikeln.

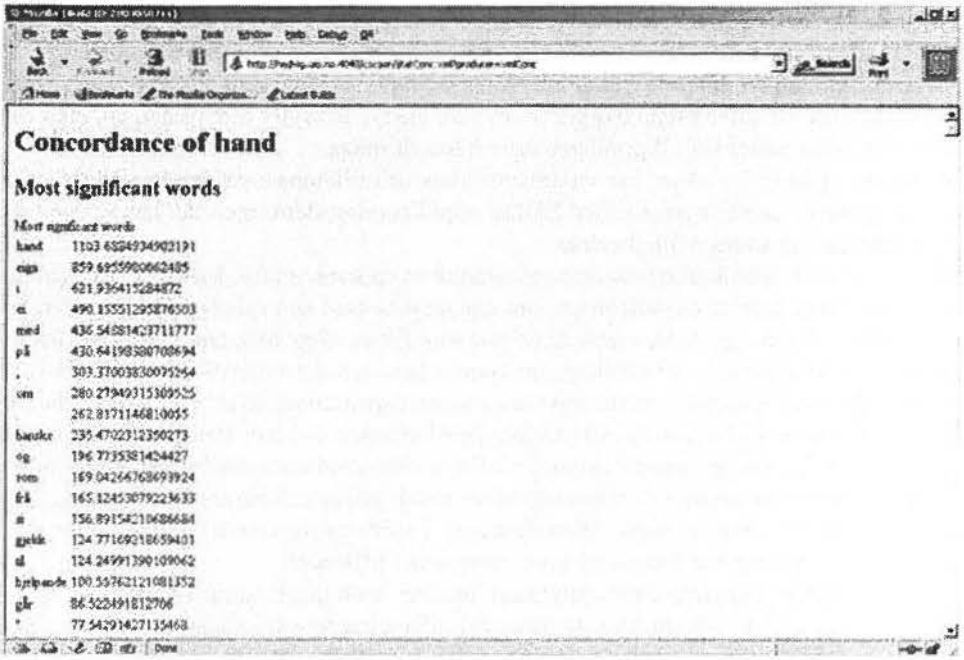
Bakom den webbaserade applikationen som beskrevs ovan finns en struktur som är samma för EDD:s applikationer. Samma funktionalitet som ovan finns. Man kan skapa konkordanser, granska dem och få mer kontext för utvalda konkordansrader. Det som tillkommer är att man kan spara undan en konkordans under ett huvudord för att sedan användas i en artikel. Detta ser man i *figurerna* 4 och 5.

I *figur* 4 är konkonkordansapplikationen i förgrunden och metaordboken är bakom. Till vänster ser man ikoner som representerar de andra datasamlingar som kan öppnas på samma sätt som dessa två är öppna.

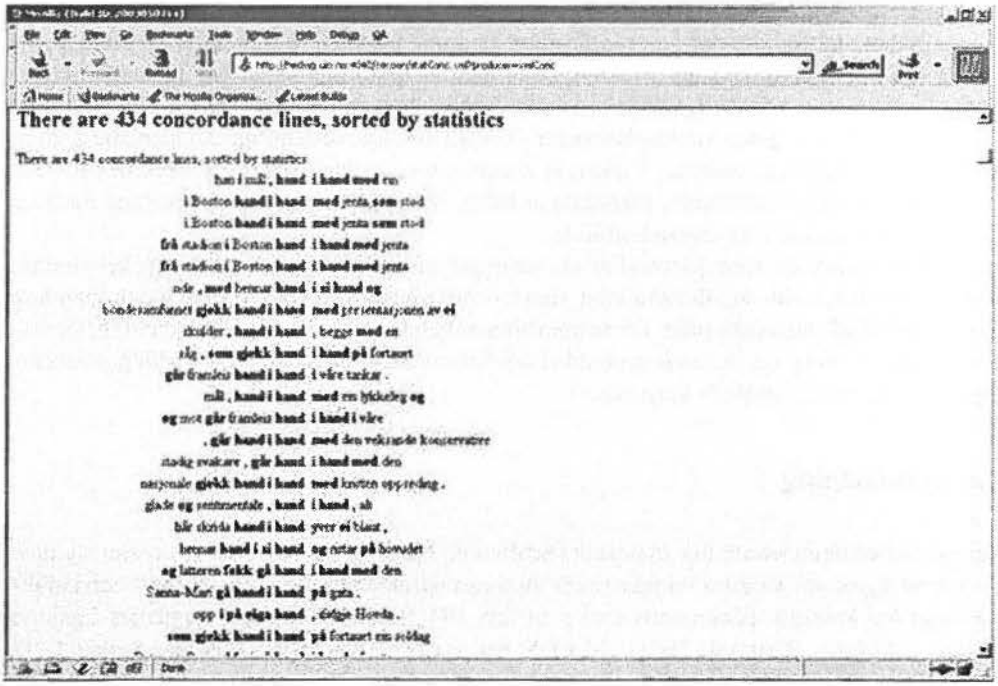
Från konkordans och vidare

Ju större en korpus blir desto mer tidskrävande blir det att bearbeta de stora resultat som en sökning kan ge. Man kommer långt med att alfabetisera konkordansrader på höger- och vänsterkontexter, men när konkordansen består av flera tusen rader räcker det inte. En lexikograf behöver stora korpusar. Ju större desto mer sannolikt att sällsynta fenomen blir belagda.

Konkordansens styrka ligger i att visa mönster kring utvalda ord, men dessa mönster kan vara nedgrävda i tusentals rader och vara svåra att komma åt. Språkteknologer (Church et al. 1991) har länge arbetat med detta problem och ett första försök har gjorts för att integrera



Figur 6: Kollikon försöker att isolera ord som är signifikanta i en kontext med utgångsordet.



Figur 7: Dessa konkordansrader har sorterats i fallande ordning efter det samlade värdet på signifikanta ord.

dem med den aktuella uppgiften, att underlätta för lexikografer. Johansson (2001) arbetade med detta, att integrera datalingvistik med verktyg för lexikografer, i sin magisteruppsats. Den styrande principen är att ett ords betydelse bestäms av det sammanhang ordet förekommer i. Många språkliga grepp bygger på att man medvetet bryter mot principen, men det normala är att man håller sig till principen efter bästa förmåga.

Johanssons system, Kollikon, har vidareutvecklats och tillämpats på den nynorska korpusen. Det har ännu inte integrerats med EDD:s applikationssystem, men det har kommit så långt att man kan illustrera möjligheterna.

I *figur* 6 ser vi att Kollikon arbetar lite annorlunda än en konkordans. En konkordans är *de facto* en subkorpus. Den är en subkorpus som har skapats med en avsiktlig vinkling. Om en korpus är, i någon mening, balanserad, är en konkordans en skev delmängd. Man använder ett ord som ett sökkriterium för att skapa en konkordans och det ordet förekommer på varenda konkordansrad. Sökordets relativfrekvens i delmängden kommer att vara mycket högre än dess relativfrekvens i korpusen. Delmängden, subkorpusen, avviker från det normala med hänsyn till sökordet. Om ett ord förekommer tillsammans med andra ord på grund av principer som inte är slumpmässiga, så följer det att även ord som är signifikanta i en kontekt med sökordet kommer att visa en högre relativfrekvens i subkorpusen, konkordansen. Detta är bara ett annat sätt att explicit formulera konkordansens förtjänster.

Kollikon försöker att automatisera processen med att hitta dessa signifikanta ord och det är det som visas i *figur* 6. Sökordet var *hand* och Kollikon har föreslagit *hand, eiga, i, ei, med* osv i fallande ordning som ord som är signifikanta i en kontext med *hand.* Vid första anblick kan detta kännas fel. Varför *hand*? Det var ju sökordet. Förklaring finns i *figur* 7.

I *figur* 6 ser vi att varje ord har tilldelats ett mått på hur signifikant ordet är i en kontext med *hand.* Ju högre tal desto mer signifikant. Om man går igenom varje konkordansrad, summera detta tal för alla ord i en rad får man ett annat mått, ett samlat mått över alla signifikanta ord i en konkordansrad. Ju högre detta mått är desto mer sannolikt att man kommer att hitta en intressant fras eller kollokation med sökordet som en beståndsdel. Det Kollikon gör är att presentera dessa konkordansrader efter att sorterat dem enligt det samlade måttet, istället för en alfabetisk ordning. Tanken är att de intressantaste konkordansraderna kommer att presenteras först, sedan andra i fallande ordning. Syftet är, återigen, att underlätta för lexikografen i arbetet med allt större korpusar.

Man kan justera ett antal parametrar till varje sökning. Man kan välja hur mycket kontext man vill ta hänsyn till, om det ska vara vänster- eller högerkontext eller båda och man kan välja bland olika statistiska mått. De måtten som används är ett urval ur Church *et al.* (1991) och Dunning (1993). En liknande metod har använts av författaren för att identifiera översättningsekvivalenter i parallella korpusar.

Sammanfattning

Presentationen är ett resultat av många års arbete i ett flertal projekt. Några har redan nämnts. Det bör tilläggas att det allra viktigaste för att dessa spridda ansatser i lexikografi och språkteknologi har kommit tillsammans under ett tak, NO 2014, är African Languages Lexicon Project – ALLEX (Grønvik 2001). ALLEX har varit ett fruktbart samarbete sedan 1992 mellan universiteten i Harare, Zimbabwe, Oslo, Norge och Göteborg, Sverige. Inom ramen

för ALLEX har man kunnat testa och implementera idéer i en kreativ miljö som syftade till två saker: att producera ordböcker och att överföra kompetens.

Litteratur

Almenningen, O. 2001: 'Seksjon for leksikografi og målføregranskning'. I: *Ord om Ord* 7, Årskrift for leksikografi, Oslo, 6-14.

Calzolari, N., Baker, M. & Kruyt, J.G. (red.) 1996: *Towards a Network of European Reference Corpora, Report of the NERC Consortium Feasibility Study*, Giardini Editori e Stampori in Pisa, Pisa.

Church, K., Gale, W., Hanks, P., Hindle, D. 1991: 'Using statistics in lexical analysis.' I: Uri Zernik (red) *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115-163. Hillsdale, N.J.

Dunning, T. 1993: 'Accurate methods for the statistics of surprise and coincidence'. I: *Computational Linguistics* 19:1.

Fillmore, C.J. 1992: '«Corpus linguistics» or «Computer-aided armchair linguistics»'. I: Jan Svartvik (red) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, Mouton de Gruyter, Berlin and New York, 35-60.

Fjeld, R. V. 2002: 'Oppbygging av leksikografisk bokmålskorpus'. I: *Ord om Ord* 8, Årsskrift for leksikografi 2002, Oslo, 23-26.

Grønvik, O. 2000: 'Metaordboka – bruk, problem og løysingar'. I: *Ord om Ord* 6, Årsskrift for leksikografi 2000, Oslo, 33-37.

Grønvik, O. 2001: 'ALLEX-prosjektet – ti års samarbeid over språkgrenser', *Ord om Ord* 7, Årsskrift for leksikografi 2001, Oslo, 42-51.

Ide, N. & Veronis, J. 1996: *Corpus Encoding Standard*, EAGLES document EAG-CWG/CES.

Johansson, S. 2001: *Kollikon – frasidentifikation och -extrahering*, Masters Thesis in Computational Linguistics, Göteborg. (http://folk.uio.no/danielr/Kollikon.pdf)

Ore, C.E.S. 2000: 'Metaordboka'. I: *Ord om Ord* 6, Årsskrift for leksikografi 2000, Oslo, 30-32.

Ridings, D. 1996: *Text representation in PAROLE*. Unpublished PAROLE report. Göteborg.

Runde, Å. 2000: 'Korpusoppbygging ved Seksjon for leksikografi og målføregranskning'. I: *Ord om Ord* 6, Oslo, 23-29.

Sinclar, J.M. (ed.) 1987: *Looking up: An account of the COBUILD Project in lexical computing*, Collins ELT, London and Glasgow.

Sperberg-McQueen, C.M. & Burnard, L. (red.) 1994: *Guidelines for Electronic Text Encoding and Interchange*, ACH, ACL, ALLC, Chicago.