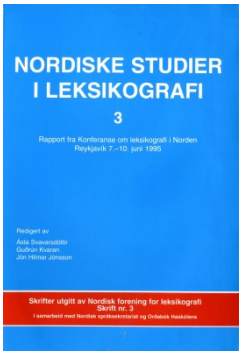
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Mer enn nok er for mye

Forfatter: Rik Schutz

Kilde: Nordiske Studier i Leksikografi 3, 1995, s. 365-373
Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995

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.

Rik Schutz

Mer enn nok er for mye

This paper describes a research project at the Dutch publishing house Van Dale during 1994 and 1995. The main goal was to develop a practical working method for corpus lexicography. The second aim was to explore the borderline between general language and technical language. The title "More than enough is too much" hints at the idea that — for the sake of both lexicographer and budget — a corpus should be limited in size. During the project we built a small, but very rich **micro corpus** with medical texts. Selecting adequate texts, according to guidelines dictated by the profile of the dictionary, will become an important task for lexicographers.

Mitt foredrag "Mer enn nok er for mye" består av fem deler:

1. Van Dale og korpusleksikografi
2. Opplegg for undersøkelsen
3. Programvaren
4. Pilotundersøkelsen som er utført for det medisinske område
5. Konklusjoner

1 Introduksjon VDL & korpusleksikografi

Nesten alle eksisterende Van Dale-ordbøker er laget uten at en elektronisk tilgjengelig samling med tekster er ekserpert systematisk. Altså uten korpusleksikografi.

Våre ordbøker har et godt navn blant publikum, og selv er vi heller ikke misfornøyde med dem. Det er altså ikke absolutt nødvendig å endre den metoden som er brukt til nå. Men i dag *har* man muligheten til å utføre korpusleksikografi, og vi ville gjerne bli bedre kjent med metodens fordeler. Utgangspunktet måtte være å oppdatere og forbedre våre eksisterende ordbøker. (Fra nå av vil jeg snakke om "ordboken", og jeg henviser da til alle ordbøker hvor nederlandsk står på venstre siden. Dette er både moderne og diakroniske ordbøker, og både forklarende ordbøker og oversettelsesordbøker.)

Formålet var altså ikke å lage en ny leksikal beskrivelse av nederlandsk "from scratch", slik som COBUILD har gjort det for engelsk. Vi ønsket å bearbeide ordboken vår med utgangspunkt i noen av de fem nøkkelord som prof. Verkuyl (1994) nevnte i sitt foredrag under Euralex 1994 in Amsterdam: komplett, konsistent, korrekt, aktuell, sitater.

Vi vet at det er visse svakheter i behandlingen av enkelte fagområder i eksisterende ordbøker. Terminologien fra sjakk er ikke **komplett** hvis *springer* og *konge* er tatt med, men *bonde* mangler. Behandlingen er ikke **konsistent** hvis det i definisjonen for *løper* fortelles hvordan brikken beveges, og denne informasjonen mangler for *tårnet*. Informasjonen er ikke **korrekt** hvis det står at et *sjakkbrett* har 100 felt. Ordboken er ikke **aktuell** hvis det

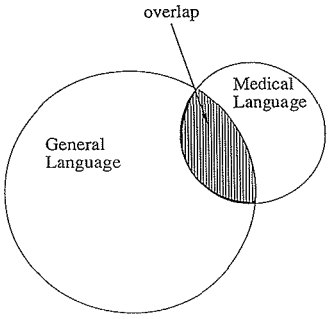
Nordiske studier i leksikografi III, 365–373

kun er tatt med den gamle termen for en brikke, *kasteel*, istedenfor *toren*. Eksemplene er fiktive, men de viser hva fire av de fem nøkkelordene innebærer.

Hensikten med prosjektet som omtales i dette foredraget, var å undersøke hvordan bruken av tekstkorpus kunne føre til bedre kvalitet på ordbøker og større effektivitet under det dyre manuelle leksikografiske arbeidet.

2 Opplegg for undersøkelsen

Helt fra begynnelsen av hadde vi to målsetninger med undersøkelsen. Den ene var å utvikle en metode for innsamling av termer ved å bruke tekstkorpuser. Den andre var å avgrense fagområder i den generelle ordboken. I figur 1 skisseres hvordan fagspråkene overlapper det generelle språket – eller det utvalget som hører hjemme i en generell ordbok. Hvordan fastslås denne grensen? Hvilke termer fra fagspråkområdet hører hjemme i en vanlig ordbok?



Figur 1.

Vi har dannet en arbeidsgruppe av sakkyndige, med blant annet en terminolog, redaktøren for en enspråklig medisinsk ordbok og forfatteren av en tospråklig juridisk ordbok. Altså mennesker med kunnskap om både fagområde og leksikografi. Denne arbeidsgruppen har gitt oss råd om opplegget for undersøkelsen.

Arbeidsgruppen hadde følgende kommentarer:

1. Det finnes to typer korpusbruk. Den ene kaller vi **verifisering,** den andre **oppsporing**. Verifisering innebærer at man vet hvilket ord man vil vite mer om. For verifisering gjelder at et korpus nesten ikke kan være stort nok, og at det kan oppnås gode resultater med relativt enkel programvare. I vår undersøkelse ligger hovedvekten på **oppsporing**.

2. Det er vanskelig å formulere søkeoppdrag for ukjente termer. Så snart man ved at ordet X finnes i teksten, kan man undersøke den leksikale omgivelsen til X med ethvert computerprogram som har en søkefunksjon. Men først må man vite hvilke ord man ønsker å vite mer om. Hvilke ord er interessante? Hvilke ord er de ukjente X-ene? (Jeg kommer tilbake til svaret på disse spørsmålene.)

3. I forbindelse med **oppsporing** støter man ganske snart på fenomenet om begrenset ekstra utbytte. Jo mer tekst som er tilgjengelig, desto større blir arbeidet med å

finne noe interessant. Flaskehalsen i den leksikografiske prosessen er leksikografens arbeidstid. Gjør korpuset minst mulig, og sørg for en elektronisk "toolkit" som muliggjør hurtig søking i korpuset.

4. Gjør utvalget så bredt og produkt-uavhengig som mulig. Senere kan det foretas egne delutvalg for forskjellige ordbøker. Til en forklarende ordbok vil det bli valgt andre termer enn for en oversettelsesordbok. Et generelt kjennetegn for samlingen er: ikke-spesialistisk.

5. Velg tekstene slik at spørsmålet "skal denne termen tas med i ordboken?" allerede er besvart så snart termen finnes i en tekst. For det medisinske området er en populær legespalte i et ukeblad mer relevant enn et legevitenskapelig tidsskrift.

6. Ta hensyn til den relative, samfunnsmessige viktigheten av fenomener. Et "medisinsk" eksempel: Det faktum at det snakkes og skrives mye om en bestemt sykdom er viktigere enn antall personer som lider av den sykdommen.

7. Sørg for (transkribert) talespråk i korpusmaterialet.

Nå vil jeg så vidt gå tilbake til den andre kommentaren: For å kunne oppspore ukjente termer, har vi forsøkt å omskrive hva "interessant" innebærer innenfor rammen av vårt prosjekt. For hver kategori nevner jeg eksempler fra sjakkterminologien. Interessant er:

- et uvanlig ord som ikke hører til basisordforrådet, men som finnes i teksten (*matt*)
- et ord som finnes mye hyppigere i teksten enn i et korpus av generelt, ikke-spesialistisk, gjennomsnittlig språk (*hest, hvit*)
- faste kombinasjoner av ord som finnes utelukkende, eller oftere enn gjennomsnittlig i teksten (*fri bonde, slå en passant*)

3 Verktøy

En første betingelse for å kunne utføre korpusleksikografi er en **taggerlemmatiser**, et program som tilføyer ordtype og lemmaform til ordet. Det fantes ikke et slikt program for nederlandsk, så Van Dale har satt i gang arbeid med å utvikle det. Avdelingen for leksikologi ved Vrije Universiteit i Amsterdam har laget **D-tale** (Dutch-tagger-lemmatiser) for oss. På figur 2 kan man se hvordan programmet virker.[1] For hvert ord bestemmer det lemmaform, med et **kommersielt à-tegn** (@) foran. Etter **nummertegn** (#) kommer en tall/bokstavkode for ordtype og bøyningsform.

En annen forutsetning er en **toolkit**, som både kan brukes til verifisering og oppsporing. Hos Van Dale har vi utviklet programmet **Coco** til dette formålet.[2]

[1] Eksemplet betyr: Er det kaldt i Reykjavik?

[2] Det eiendommelige navnet angir at programmet først og fremst hjelper til med søking av kollokasjoner (**co**llocations) i et korpus (**co**rpus).

D-tale (Dutch tagger/lemmatizer).

before D-tale:

Is het koud in Reykjavik?

after D-tale:

Is@zijn#3E **het**@het#7 **koud**@koud#2
in@in#6 **Reykjavik**@Reykjavik#1N
?@?#|

Figur 2.

Jeg begrenser meg til å nevne noen viktige funksjoner:

KWIC. D-tale og Coco arbeider med hele setninger. Ut-data avviker derfor fra det vanlige KWIC-formatet. På figur 3 kan man se hva søk etter ordet *operatie* (operasjon) i kombinasjon med et verb fører til.

Lagring av resultatet. Resultatet av et søkeoppdrag lagres automatisk. Materialet får den strukturen som brukes til lagring av leksikal informasjon hos Van Dale.

Frekvens. Frekvensen kan beregnes både for ordform og lemma. Spørsmålet om frekvensen av ordet *operatie* (operasjon) gir resultatet man kan se på figur 4.

Z-score. Den viser hvilke ord gjerne står i nærheten av hverandre. Eller mer vitenskapelig formulert: den relative grupperingsstyrke mellom søkeordet og dets leksikale omgivelse. Figur 5 viser resultatet av spørsmålet om preposisjoner fremfor ordet *operatie*.

in KWIC format

In een latere fase wordt de definitieve
[> operatie <] [> verricht <].

in Lemma format

<COMPO.> operatie
<GRAMT.> 1
<COMPO.> verrichten
<GRAMT.> 3
<STVRM.> een operatie verrichten
<CITAAT> In een latere fase wordt de definitieve operatie verricht .
<BRONC.> Nigsys teksten
<BEWER.> rik

Figur 3.

Coco query result

Frequency of the lemma operatie

1	**Operatie**@operatie#1E
3	**Operaties**@operatie#1M
163	**operatie**@operatie#1E
25	**operaties**@operatie#1M

Total 192

Figur 4.

Coco query result

Z-score for **operatie**, preceded by a preposition

Expression A: operatie
Scope : 3
Surr : only before A
Collocator : (B)
Minimum : 2
Expr.Coll. : preposition
Totaal #Wrd : 444111

Score	F(B)	F(A+B)	preposition (B)
29.2802	587	30	na [after]
5.3387	265	4	tijdens [during]
4.2142	4146	18	bij [at]
2.0803	1943	7	door [result of]
1.8926	958	4	uit [from]

Figur 5.

Kompleks grupperingssøker. Denne kalkulerer de sterkeste grupperinger i en tekst uten at søketermen må angis. Eksempler fra et korpus om sjakk kunne være: *svart står sjakk*; *kort rokade*.

For å begrense regnetiden mens man bruker programmet, og for å ha en rettesnor mens man utfører valget, er fremstilling av hjelpelister en del av standardprosedyren ved forberedelsen av en tekst. Listene oppfyller det tidligere nevnte kriteriet "interessant".

Uvanlige ord. Ved å sammenligne ordet fra et korpus med to kilder med utelukkende "vanlige" ord, valgte vi det uvanlige ordet fra korpuset. Som kilde for vanlige ord brukte vi (1) Basisordbok fra Van Dale, en skoleordbok med 24 000 oppslagsord og (2) et normkorpus, basert på avistekster.

Siden det er valgt tekster med enkel språkbruk, er ordene i denne listen nøyaktig de ordene som er forbundet med tekstens tema, altså med det undersøkte fagområdet. Interferens oppstår særlig av sammensetninger, navn, akronymer, tall og skrivefeil.

Relativt høy frekvens. I denne listen sammenligner vi frekvensen til et ord i korpuset med den forventede frekvens på grunnlag av tellinger i normkorpuset. På denne måten håper vi å finne en spesiell betydning eller konnotasjon for det undersøkte fagområdet. Programmet virker slik at jo større avviket er fra det forventede, desto høyere score. Scoren er altså ikke direkte avhengig av ordets absolutte frekvens.

4 Pilotundersøkelse for det medisinske området

Vi har valgt det medisinske området for å utføre en pilotundersøkelse. Området er stort, det er i stadig utvikling og den eksisterende ordboken er ikke ajourført. Undersøkelsen består i å:

- fastslå hvilken type tekst innenfor området brukeren må kunne dekke/bearbeide med ordboken (tekster skrevet for ikke-medisinere, som når et stort publikum)
- velge tekster som oppfyller typen (legespalter i ukeblader, opplysningsbrosjyrer fra sykehus og pasientforeninger, samtaler mellom lege og pasient, en fjernsynsserie fra et sykehus)
- samle tekster og forarbeide dem elektronisk (be forlag eller forfattere om disketter; fjerne ikke relevante koder i tekstene; D-tale)
- ekserpere tekstene systematisk (redaksjonsarbeid)
- sammenligne utvalget med det som står i den eksisterende ordboken
- få en sakkyndig på området til å bedømme utvalget
- trekke konklusjoner. Er spørsmålene besvart, og er målene nådd?

De data som nevnes her, gjelder de medisinske *n*-ord, det vil si ord som begynner på en *n* og sammensatte uttrykk hvor ett av ordene begynner på en *n*.

Det finnes 199 *n*-oppslagsord med en <medisinsk> etikett i ordboken (det kan også være anatomi, patologi eller noe annet). Av de sammensatte forbindelsene som finnes med disse oppslagsordene, regner jeg 25 til det medisinske området. Totalt blir det 224. Det er for øvrig vanskelig å fastslå grensen. Under *nese* tar jeg også med *en tett nese* og *puste gjennom nesen* til det medisinske området, men ikke *pille nese*.

Det er valgt 443 *n*-termer fra mikrokorpusene (300 enkle som begynner på en *n*, 143 sammensatte hvor ett av ordene begynner på en *n*). Mange av disse termene står i ordboken uten <medisinsk> etikett. Det forklarer de forskjellige tallene i feltet "+/+" på figur 6.

in dictionary	In μ-corpus of relevant domainrelated texts + 443 N	-
+ 224 N <med.> without label	+/+ 64 (30%) 148 (33%)	+/- 160 (70%)
-	-/+ 295 (67%)	-/-

Figur 6.

Forskjellene kan gjengis i følgende skjema:

+/+ Gir en bekreftelse på at utvalget allerede er gjort tidligere av en leksikograf i forbindelse med ordboken.

+/− I korpuset er det ikke funnet noen bekreftelse på ord som er tatt med i ordboken. Bedømmelse av en sakkyndig gir svar på om den eksisterende ordboken inneholder noe som ikke hører hjemme i den, eller om utvalget av tekstene eller av termene fra de tekstene ikke var det riktige.

$-/+$ Ordboken inneholder åpne felt. Også i dette tilfellet må en sakkyndig bedømme om det dreier seg om nyttige tillegg, eller om det skyldes et mindre omhyggelig valg av tekster, eller av termer fra de samme tekstene.

$-/-$ Alle ikke-eksisterende ord kommer i denne kategorien. Men vi forblir nysgjerrige etter de interessante termene som vi åpenbart ikke har valgt de riktige tekstene til.

4.1 Bedømmelse av kategori $+/-$ (finnes i ordboken, ikke i μ-korpus)

Som forventet er mange termer modne for en kritisk bedømmelse. Flere vil bli sløyfet, eller blir merket som <gammel>. De har for lett blitt tatt med fra tidligere utgaver, selv om termen eller fenomenet som blir angitt med termen, er gått ut av bruk. Enkelte burde kanskje aldri ha vært tatt med, fordi de var altfor spesielle for en generell ordbok, f.eks. *net* (fettrik hinne mellom buk og innvoller), *netelkoorts* (feber som forårsakes av irritasjon av huden med brennesle ol.), *nijdnagel* (der hvor huden er revnet langs en negl), *nootgewricht* (kuleledd, ledd som kan beveges i alle retninger), *natriëmie* (å forebygge natrium i blodet), *negenponder* (baby som veier ni pund ved fødselen).

I andre tilfeller dreier det seg om sammensetninger hvor det er tvilsomt om de er relevante for ordboken. Betydningen av ordene er ganske åpenbar, de forekommer ikke hyppig og de utgjør ikke grunnlaget for lengre sammensetninger, f.eks. *neusklier* (nesekjertel), *neusheelkunde* (nesemedisin), *nierader* (nyreåre), *nierkanker* (nyrekreft). Det er ikke nødvendig å fjerne dem fra databanken, men det er ikke særlig sannsynlig at de tas med i det utvalg som ethvert leksikografisk produkt er.

En annen, mindre del av ordene i denne kategorien hører imidlertid absolutt hjemme i ordboken. Det at de ikke finnes igjen i μ-korpusen, skyldes som regel den vage begrensningen av området. De "medisinske" etiketter i ordboken er mer omfattende enn de kriteria som tekstene for korpusene ble valgt etter. Anatomiske termer er et eksempel på dette, særlig når de i større grad gjelder dyr enn mennesker (*nekhaar* (nakkehår), *nekvel* (nakkeskinn)).

4.2 Bedømmelse av kategori $-/+$ (finnes i μ-korpus, ikke i ordboken)

Siden ekserperingsarbeidet er utført med en instruks om at termer skal velges i tilfelle tvil, er samlingen stor, større enn det som får plass i en gjennomsnittlig generell ordbok av normalt omfang. Det er selvsagt ingen ulempe, men det er viktig å være klar over det ved bedømmelsen av materialet.

Antall tillegg i forhold til ordboken er imponerende. Helt vanlige ord, som også mange lekfolk på det medisinske området umiddelbart vil gjenkjenne, viste seg til vår forundring å mangle i ordboken: *nepmiddel* (placebo), *neurotransmitter* (transmitter; definisjon i *Bokmålsordboka*: "stoff i kroppen som overfører impuls fra nerve til muskel"), *natriumbeperkt dieet* (diett med nedsatt natriuminnhold), *natuurlijke bevalling* (naturlig fødsel), *bijkomen uit de narcose* (våkne/komme seg etter narkosen).

Det er også noe interferens i materialet.

1. I denne gruppen finnes det dubletter som kun avviker i stavemåte og et ganske stort antall semi-dubletter av typen *negatieve uitslag* (negativt resultat), *negatieve testuitslag* (negativt prøveresultat), *negatieve aids-test* (negativ aids-test).

2. Det finnes et temmelig stort antall forbindelser og sammensetninger hvor ordet (orddelen) som begynner på bokstaven *n*, ikke tilhører det medisinske området, slik som *ta* (pulver, medisin, tablett, paracetamol) og *ikke* (ikke gravid, ikke arvelig, ikke medfødt).

Til nå har jeg utelukkende snakket om innsamling av termer og ikke om hvordan disse termene må behandles leksikografisk. Noen av de valgte tekstene er av opplysende art. Det innebærer at mange termer blir forklart i teksten. Coco lagrer en valgt term automatisk, sammen med hele setningen som termen brukes i. Slik oppstår det i tillegg til ordene en samling med sitater som kan spille en viktig rolle ved definisjonen av termene. Noen eksempler:

nekrose: lokal død i vev (for eksempel ved koldbrann eller forbrenning).

neseendoskopi: innvendig undersøkelse av nesen med endoskop.

5 Konklusjoner

5.1 Pilotprosjektet (medisinske termer)

Selv et svært begrenset korpus gir mye brukbart materiale for oppgradering og supplering av den eksisterende ordboken.

- Mange av termene som er tatt med i ordboken, kan fjernes eller revurderes ved hjelp av materialet i μ-korpusene.

- Materialet gir en sammenhengende, relativt komplett samling av termer, hvorav mange utgjør et viktig supplement til den eksisterende ordboken.

5.2 Den valgte fremgangsmåten

- Omfattende ekserpering av et ganske lite korpus kan føre til en avbalansert samling av termer, forutsatt at korpuset er fremstilt svært nøye. Kvaliteten på utvalget er helt avhengig av valget av tekstene. Den nøyaktige begrensningen av området er av avgjørende betydning.

- Hjelp fra område-ekspertise — altså en lege i vårt eksempel — kan begrenses til rådgivning om valget av tekster og bedømmelse av resultatene. En stor del av arbeidet kan utføres av computere og ikke-eksperter. Det reduserer utgiftene.

- En videre automasjon av prosessen, særlig ved indre kontroll av ordboken, ser ut til å være mulig og vil gjøre prosessen enda mindre arbeidskrevende.

- Den undersøkte metoden begrenser seg ikke nødvendigvis til fagspråk. Også til oppsporing av neologismer eller ungdomsspråk kan det arbeides med et lite korpus av nøye utvalgte tekster.

Litteratur

Johansson, Stig 1991: Times change, and so do corpora. I: Karin Aijmer/Bengt Altenberg (eds.): *English corpus linguistics*. London: Longman.

Meijs, Willem 1994: Computerized lexicons and theoretical models. I: Nelleke Oostdijk (ed.): *Corpus-based research into language*. Amsterdam: Rodopi.

Summers, Della 1993: Longman/Lancaster English Language Corpus — Criteria and Design. I: *International Journal of Lexicography* 6, 3, 181–207.

Verkuyl, Henk 1994: *Knowledge Representation in Dictionaries*. Keynote lecture delivered at the 6th Euralex International Congress 1994. Amsterdam.