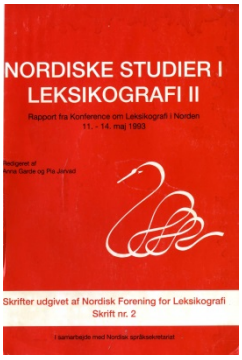
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Datamatstøttet vending af ordbøger

Forfatter: Anders Drejer Nygaard

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 237-242
Rapport fra Konference om leksikografi i Norden, 11.-14. maj 1993

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.

Datamatstøttet vending af ordbøger

Anders Drejer Nygaard

Denne artikel skildrer arbejdet med automatisk at producere en råudgave af en aktiv ordbog ud fra en tilsvarende passiv. En mærkning af den passive ordbogs indholdsstruktur efter den internationale standard SGML gør selve vendingen til en relativt simpel mekanisk proces. Et væsentligt delproblem er herefter at vælge hvilket opslagsord, udtrykkene i den aktive ordbog skal anbringes under, og en heuristik bliver beskrevet. Ideerne i denne artikel er blevet anvendt på Engelsk-færøsk ordbog, og Færøsk-engelsk ordbog er nu under produktion ud fra resultaterne.

Baggrund

Færøsk er, som bekendt, et lille sprog, så det store og bekostelige arbejde med at producere en ordbog fra grunden kan sjældent svare sig. Forlaget Stiðin besluttede derfor at købe manuskriptet til den røde engelsk-danske ordbog fra Gyldendal, og oversætte den danske del til færøsk, for herved at få en engelsk-færøsk ordbog. En tilsvarende fremgangsmåde lader sig ikke praktisere, når man vil producere en færøsk-engelsk ordbog. Imidlertid rådede man allerede over en maskinlæsbar udgave af en engelsk-færøsk ordbog, så hvorfor ikke prøve om det kunne lade sig gøre ad maskinel vej at 'vende' denne ordbog til den ønskede færøsk-engelske? Det var naturligvis fra starten klart, at der ikke ville kunne komme et færdigt produkt ud af anstrengelserne, men en stor del af indsamlingsarbejdet, og arbejdet med at finde ækvivalenter måtte kunne foretages på denne måde.

Fremgangsmåde

Centralt for arbejdet med at vende ordbogen er det begreb som Honselaar & Elstrodt [2] kalder en 'micro entry', men som jeg foretrækker at kalde en atomar artikel eller et atom. En atomar artikel indeholder præcis et par af (oversættelses)ækvivalente enkeltord eller udtryk, sammen med oplysning om emneområde, brugsoplysninger, eksempler, etc.

Ud fra denne synsvinkel er en ordbog en stor samling af atomare artikler, der er organiseret ved at man udnævner et af de to sprog til kildesprog (og dermed det andet til målsprog), sorterer indgangene efter kildesprogsord, og udfaktorerer information, der er fælles for flere atomare artikler. Denne information er så yderligere suppleret med

krydshenvisninger, kildesprogsforkortelser, begreber, der kun findes på kildesproget, etc.

Fremgangsmåden for at vende en ordbog er nu i princippet ganske enkel, og består af 6 trin:

1 Analyse
2 Udtynding
3 Opløsning
4 Vending
5 Sortering
6 Samling

Fase 1, Analyse: Ordbogen analyseres til struktureret form.
Dette er en velstuderet proces med en lang række kendte problemer (se fx Christensen [1] eller Jørgensen [3] for en nærmere beskrivelse). Resultatet af denne proces er en hierarkisk struktureret logisk opmærkning af hver artikel, således at der er klart afgrænset Hoved, OpslagsOrd, Krop, SubArt(ikel), Ækviv(alent), TydSpec(ifikation), etc.

Fase 2, Udtynding: Al supplerende information (henvisninger, parafraser, kildesprogsspecifikke forkortelser, etc.) bliver kasseret
En simpel proces, givet den logiske (modsat typografiske) opbygning af SGML-beskrivelsen.

Fase 3, Opløsning: Artiklerne opløses i atomer.
En simpel mekanisk proces, givet den hierarkiske opbygning af SGML-strukturen. Det eneste, der skal foretages er at kopiere hovedet og relevante TydSpecifikationer for hver Ækvivalent.

Fase 4, Vending: Alle atomer vendes (dvs. det tidligere målsprog udnævnes til nyt kildesprog).
Simpel, givet visse forudsætninger. I en ordbog forventer man at se ordklasseangivelser for opslagsordene, men under vending af Ensk-føroysk har der kun været adgang til ordklasseangivelser på de engelske ord. Dette er dog i praksis kun et lille problem, idet langt de fleste ækvivalentpar har samme ordklasse.

Fase 5, Sortering: Indgangene sorteres efter det ny kildesprog.
Skal for at være anvendelig også tage hensyn til alfabetiseringsreglerne for færøsk, samt kunne håndtere flere nøgler - fx har vi ønsket at den vendte ordbog havde alle udtryk og fraser samlet sidst i artiklerne, og anvendt dette som et af sorteringskriterierne.

Fase 6, Samling: Fælles information udfaktoreres.
En noget besværlig, men stadig mekanisk fjernelse af fælles information fra de atomare artikler. Resultatet af dette trin er en SGML-beskrivelse af den vendte ordbog, som kan redigeres videre.

I praksis er der dog adskillige komplikationer. Den principielle beskrivelse af en ordbog som en samling af atomare artikler er kun holdbar til en vis grænse - en ordbog er i reglen skrevet med henblik på fuld dækning af kildesproget, hvorimod dækningen af målsproget i høj grad vil være tilfældig, især når der findes flere synonyme eller nærsynonyme ord på målsproget. Dette, samt mangelen på begreber, forkortelser, etc., der er specifikke for målsproget, gør at resultatet (den vendte ordbog) i det mindste vil skulle suppleres med indholdet af en ordliste.

Primære ord i udtryk

Et centralt problem ved sorteringen af de atomare artikler er, hvordan man behandler kildesprogsudtryk. Leksikografens sædvanlige fremgangsmåde er at udnævne et af ordene i et udtryk til det primære, og anbringe oversættelsen sammen hermed. Denne udnævnelse vil imidlertid benytte sig af vage kriterier som 'det mest betydningsbærende ord' eller 'det mest karakteristiske ord' i udtrykket; noget end ikke menneskelige leksikografer kan blive enige om (som enhver der har prøvet at slå op i en ordbog vil vide!).

Nogle ting kan man dog sige om hvilket ord i et udtryk, der er det primære; det vil sjældent være en artikel, præposition, pronomen, hjælpeverbum eller konjunktion.

For en datamat er det således vanskeligt at finde det primære ord i et udtryk, men det er i nogen grad muligt at støtte sig til konteksten i den artikel, atomet stammer fra. Lad os tage et eksempel: I Ensk-Føroysk ser artiklen for substantivet 'keel' således ud:

I. keel *n* kjølur (eis *plfr*); kólpramur; *on an even* ~ á rættkjøl; (*flm*) róliga, javnt og samt; *lay (down) the* ~ *for* strekkja kjølin til.

Her har vi 6 atomare artikler, der alle har 'keel' som det primære ord på engelsk:

keel - kjølur
keel - kólpramur
on an even keel - á rættkjøl
on an even keel - róliga
on an even keel - javnt og samt
lay (down) the keel for - strekkja kjølin til

Tre af de færøske oversættelser består kun af et ord, der naturligvis er det primære. Udtrykket 'á rættkjøl' består af to ord, hvor det ene (á) er en præposition, så her må 'rættkjøl' være det primære. I udtrykket 'javnt og samt' kan vi kun udelukke konjunktionen 'og', og får en uopløselig tvetydighed. I udtrykket 'strekkja kjølin til' kan vi kun udelukke præpositionen 'til', men af konteksten fremgår det at 'kjølur' er en mulig oversættelse af det ord, 'keel', som forfatteren har valgt som det primære i det engelske udtryk, så derfor må den bøjede form 'kjølin' være et bedre valg end 'strekkja' som det primære ord.

Da færøsk er et sprog, hvor ordene bøjes flittigt (bøjningssystemet minder en del om det tyske: der er fire kasus for substantiver; verberne bøjes i person), vil ordene i et udtryk i reglen optræde i en bøjet form, så for at kunne genkende oversættelser af det pri-

mære engelske ord i de færøske udtryk må man kunne genkende bøjede former. En egentlig morfologisk analyse af ordene i de færøske udtryk falder imidlertid uden for projektets rammer, så jeg har måttet nøjes med en heuristik.

Det viser sig at hvis man betragter ord, hvor de første tre/fire bogstaver er ens, får man tilfredsstillende resultater. Denne heuristik kan naturligvis ikke tage højde for alle bøjningsfænomener (vokalforskydninger er et oplagt eksempel), og den giver også en del 'falske' bøjningsformer (typisk vil sammensætninger blive klassificeret som bøjede former), men disse fejlkilder er forsvindende i sammenligning med problemerne i Fase 1, Analysen.

Algoritme til udvælgelse af primære ord

Svarende til observationen ovenfor om hvilke ord, der bliver valgt som primære ord, benytter algoritmen sig af en stopliste, der indeholder ord, der kun med ringe sandsynlighed kan være det primære ord i et udtryk.

Stoplisten indeholder alle præpositioner (á, av, ...), artikler (alla, alt, ...), konjunktioner (at, ...), almindelige hjælpeverber (fáa, fara, ...), pronominer (eg, okkum, ...), diverse almindelige småord (ikki, ið, ...), samt en række almindelige adjektiver (gomul, litil, ...). Ialt er der 180 ordformer på stoplisten.

Enhver ækvivalent inden for en artikel bliver behandlet efter tur, idet man forsøger at finde primære ord. Hvis ækvivalenten kun indeholder ét ord, eller hvis den kun indeholder ét ord, der ikke optræder på stop-listen er det primære ord *entydigt*. Hvis der ikke er et entydigt primært ord, tages i prioriteret rækkefølge enten:

- alle tidligere forekommende entydige ord i denne artikel,
- alle ord, der er bøjningsvarianter af tidligere forekommende entydige ord i denne artikel,
- alle ord, der ikke optræder på stoplisten, eller
- alle ord.

Bemærk at fremgangsmåden i praksis garanterer, at der faktisk kommer artikler i den vendte ordbog, der omhandler ord fra stoplisten - de skal bare forekomme som eneste ord i en oversættelse, hvilket de typisk gør i netop de relevante artikler i den originale ordbog.

Vending af Ensk-føroysk orðabók

Filosofien bag produktionen af den endelige skitse har været, at det skal være så let som muligt for redaktørerne at fuldføre processen. Da det er meget lettere at slette i et forlæg, end at tilføje manglende informationer, har vi valgt at lade tvivlstilfælde falde ud til fordel for overgenerering.

I Ensk-Føroysk orðabók forekommer der følgende oplysningstyper, som behandles som angivet:

- Opslagsord
  Genanvendes som oversættelser
- Udtale
  Slettes, idet udtaleangivelser for målsproget ikke ønskes
- Ordklasse (engelsk)
  Genanvendes som færøsk ordklasse, i de tilfælde hvor ækvivalenten kun består af ét ord. Fejlprocenten ved at gøre dette er ganske lav (ca. 1-2), til gengæld for en stor besparelse i redaktionsarbejdet.
- Diskriminatorer
  Genanvendes som diskriminatorer. Dette vil ikke altid være tilstrækkeligt (end ikke korrekt), men anvendeligt som udgangspunkt.
- Oversættelser
  Genanvendes som opslagsord og udtryk.
- Parafraser
  Problematiske, idet de ikke altid er til at skelne fra egentlige oversættelser. I de tilfælde hvor en entydig identifikation som parafrase er mulig bliver oplysningen slettet, ellers bliver de behandlet som almindelige oversættelser.
- Eksempler (engelsk)
- Eksempler (færøsk)
- Forkortelser (engelsk)
  Slettes, hvis der ikke er en færøsk ækvivalent
- Forkortelser (færøsk)
- Henvisninger
  Slettes
- Rammebundne præpositioner
  Overføres, idet de bliver vendt
- Brugsoplysninger
- Latinske navne
  Bliver uddraget til en speciel fil, således at redaktøren kan behandle dem samlet.
- Typografisk hjælpeinformation: tegn, parenteser, nummereringer, etc.
  Anvendes kun til at styre analysefasen

Den mængde af atomer, der bliver produceret ved analysen af en ordbog som Ensk-føroysk, udgør en temmelig stor 'seddelsamling', som i fase 6 bliver organiseret til noget, der ligner en ordbog ved at bortfaktorere fælles information. Denne 'ordbog' skal, ud over indholdet i den endelige ordbog, også indeholde angivelser af, hvornår der har været flertydighed ved bestemmelsen af det primære ord - der er jo ingen grund til at have den samme oversættelse stående flere gange i samme ordbog. Dette har vi valgt at gøre ved at understrege ord i færøske udtryk, som er alternative placeringer for det pågældende udtryk.

Endvidere har vi ment, at det kan være praktisk for redaktøren at kunne se, hvorfra en given oversættelse stammer, og vi har derfor indført en '*' ved det engelske opslagsord i de tilfælde, hvor der kan opstå tvivl.

Konklusion

De 47.555 artikler i Ensk-føroysk orðabók er blevet vendt til 64.610 artikelskitser (her er bøjede former talt flere gange).

Der har været en enorm besparelse med hensyn til indsamling af oplysninger.

Der er en meget lille besparelse på redigeringen af de store artikler (se fx 'royna', figur 1), men mange af de helt små artikler er faktisk færdigredigerede.

Endelig har hele processen afsløret en del mangler, inkonsekvenser og egentlige fejl i forlægget, Ensk-føroysk orðabók - især alfabetiseringen af den vendte ordbog har vist sig at være nådesløs til at afsløre stavefejl i forlægget.

Man må således sige, at hele processen har medført en stor arbejdsbesparelse, samt givet et nyt og værdifuldt indblik i en eksisterende ordbog, altsammen for en relativt beskeden indsats.

Litteratur

1. Ole Norling Christensen: *Struktureret redigering af ordbøger*, i Nordiske studier i leksikografi, Nordisk forening for leksikografi, Oslo 1992.
2. Wim Honselaar og Marijke Elstrodt: *The electronic conversion of a dictionary: from Dutch-Russian to Russian-Dutch*, i proceedings from EURALEX '92, Department of Translation Studies, University of Tampere, Tampere, Finland 1992.
3. Claus Bo Jørgensen: *Parsing af ikke-strukturerede ordbogsdata*, i proceedings fra Konference om Leksikografi i Norden, 1993

royn: *kom og royn* come and have a *try; *royn aftur* have another *try; *royn eina ferð afturat* have another *try; *royn ikki at lumpa meg* don't *try anything on with me.

royna *hj* *out to *(td* make money); *#. n* put to *trial; **T** have a *smack at; **T** try-out; (ein *el* eitt) put to the *proof; (ein) put to the *test; (seg við) have a *go at; *#. s* *check up on; *go over *(td* let us *go over the last act); *try out; assay; attempt; endeavour; essay; experience; experiment; find; fish; offer *(td* he -ed to strike me); prove; prove *(td sannleikavirði á);* see; seek; taste; test; try *(td* his patience was tried); *(glm)* prove *(td* his worth); (um dýpd, *flm)* sound; (um skotvápn) range; (við snøri *el* tráð) angle

· *fara at royna eitt* **T** do it on *spec *(st* speculation) *(td* I don't know whether he is there, but I'll go there on *spec); *eg fari at royna at náa tokinum* I'll have a *shot for the train; *lat hann sleppa at royna seg (el at royna flogið)* give him a *trial; *royna aftur* repeat; *royna at fáa* *go for; *royna at* *study to; *royna eitt* have *(el* make) a *stab at sth; have a *whack at; *royna hann* put him through his *facings; *royna seg við* have *(el* take) a *fling at; *royna seg ímóti (el í)* **T** have a *bash at; *royna seg ímóti* *measure one's strength with; try *conclusions with; try a *fall with; *royna seg* try one's *luck; *(flm* um persón) expand; (ímóti) match *(td* *match your strength against his); *royna at fáa at vita hvat hann hevur í kvitta(num) (flm, eis)* *sound him out; *royna at fáa* *try for; *royna at vinna sær tíð* *play for time; *royna at* **S** have a *shot at sth; *royna seg við* try one's *hand at; *taka at royna eitt* **T** give it a *whirl.

roynandi *l* tentative.

roynast *s* *prove oneself (to be) *(td* he -d himself to be a true friend); deliver; make; prove *(td* the story -d false); **S** *deliver the goods; (við lýsingarorði, væl *el* illa) run; (væl *el* illa) *turn out; (væl) *prove itself *(td* this method has -d itself); (væl) *prove oneself *(td* this method has -d itself)

· *eg haldi meg vita hvussu hann fer at roynast* I have got him *taped; *hann fer at roynast* he *improves on acquaintance; *nógv bendir á at tað fer at roynast væl* show (great) *promise; *roynast sannur* *prove true; *roynast sum maður* *play the man; *tað at roynast verri enn evnini eru til* underachievement.