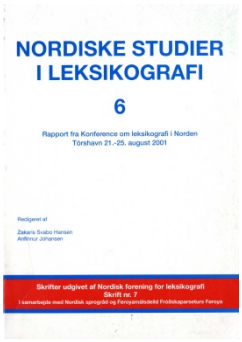
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Ordbogen og den daglige tale - Om den islandske talesprogsbank (ISTAL) og dens betydning i ordbogsredaktion
Forfatter:	Ásta Svavarsdóttir
Kilde:	Nordiske Studier i Leksikografi 6, 2003, s. 43-48 Rapport fra Konference om leksikografi i Norden, Tórshavn 21.-25. august 2001
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.

Ásta Svavarsdóttir

Ordbogen og den daglige tale

Om den islandske talesprogsbank (ISTAL) og dens betydning i ordbogsredaktion

The paper presents an Icelandic project, ISTAL, which has the purpose of compiling and analysing spontaneous speech, making the material accessible in a databank. The material consists of approximately 20 hours of personal conversations in Icelandic, taped in informal, natural settings. A preliminary study of the vocabulary in a part of the ISTAL-material is described in comparison with two samples of written texts. One consists of a number of diaries and memoirs, i.e. informal and personal writings (SKRIFT-1), and the other of newspaper articles representing a more formal and impersonal register (SKRIFT-2). A number of words in the spoken language material are not to be found in the comparative samples; in many cases the difference is obviously accidental, but some of these words are presumably more typical for the spoken language. Furthermore, a number of words and word forms are significantly more frequent in the conversations than in the comparative written texts and thus emerge as characteristic for the spoken language. It is no surprise that there is less difference between the spoken language material and the less formal, more personal register of SKRIFT-1 than between the conversations and the more formal SKRIFT-2. Finally, some of the words that this comparison shows as being characteristic for the spoken language are surveyed in the most common Icelandic dictionary. The conclusion is that access to data like ISTAL could improve the dictionary description.

Indledning

Ved redigering af islandske ordbøger har man hidtil ikke haft nogen direkte adgang til autentisk talesprog ud over redaktørernes eget sprog og det de hører i deres nærmeste omgivelser samt enkelte udsagn fra informanter, navnlig i tilknytning til dialektale ord og vendinger. Således har belæg for talesprogsord gerne været indirekte mens man har som oftest haft adgang til større eller mindre citatsamlinger som kilder for skriftsproget. Selvfølgelig er det først med ny teknik i det 20. århundrede som det er blevet muligt at arbejde med større mængder autentisk talesprog end enkelte ord eller setninger. Og selv om man nu til dags i teorien har direkte adgang til en masse talesprogsmateriale er det i praksis meget tidskrævende at arbejde med en sådan samling idet det forudsætter som regel en transskription af båndoptagelserne. Desuden er det ikke helt enkelt at indsamle visse typer autentisk tale, især fra mere uformelle talesprogssituationer, som fx. personlige samtaler.

I Island er der i de sidste år opstået øget interesse for talesproget og adskelligt talesprogsmateriale er blevet indsamlet og transskriberet i forbindelse med forskellige forskningsprojekter, især materiale fra radioen. Der har dog ikke været almen adgang til disse samlinger og de har ikke været udnyttet i ordbogssammenhæng. I 1999 blev det

såkaldte ISTAL-projekt sat i gang med det formål at indsamle og bearbejde materiale til en islandsk talesprogsbank. I det følgende gøres der rede for projektet og en præliminær undersøgelse af ordforrådet i en del af materialet sammenlignet med skriftsprog, bl.a. med hensyn til ordbogsarbejde.

Den islandske talesprogsbank ISTAL

I årene 1999-2000 fik en gruppe sprogforskere et større stipendium fra Islandsk forskningsråd for at indsamle og bearbejde materiale til en islandsk talesprogsbank. Initiativet er kommet fra fiórunn Blöndal, lektor ved Islands pædagogiske universitet, som også har været projektleder. De seks andre deltagere kommer fra Leksikografisk institut, Islands universitets humanistiske fakultet og det pædagogiske universitet.[1] Disse danner en styregruppe der sammen med projektlederen har ansvaret for projektets planlægning og organisation. Indsamling af materiale blev udført af studentermedhjælpere og en medarbejder har været fast ansat ved at transskribere båndoptagelserne; desuden har projektet haft en anden medarbejder et par måneders tid til videre bearbejdelse af materialet.

Formålet med ISTAL er at indsamle et udvalg af uformelt, islandsk talesprog og gøre materialet brugbart og tilgængeligt således at det kan danne grundlag for forskellige videnskabelige og praktiske projekter indenfor sprogforskning og sprogteknologi. Materialet består af spontane samtaler mellem to eller flere voksne mennesker, enten en blandet gruppe mænd og kvinder eller udelukkende kvinder eller mænd. For at gøre samtalesituationen så naturlig som muligt, blev samtalerne optaget i medhjælperenes egne hjem, blandt familie og venner, eller på deres arbejdsplads, oftest i en mad- eller kaffepause, og medhjælperene deltog selv i samtalerne. Deltagerenes alder, køn, bopæl og beskæftigelse registreredes nøjagtig samt forskellige oplysninger om samtalesituationen. Ialt blev der indsamlet 31 brugbare samtaler, omtrent 20 timer.[2] De fleste samtaler er 10-40 minutter lange, selv om nogle er kortere og andre længere, og der er som regel 2-4 deltagere i hver samtale.

Samtalerne blev optaget digitalt på en mini-disk og optagelsernes kvalitet er gennemgående meget tilfredsstillende. Transskriptionen er derfor gået forholdsvis godt, men den er alligevel den mest tidskrævende del af arbejdet. Samtalerne transskriberes ortografisk og i første omgang registreres desuden forskellige konversationstræk, som fx. turveksling, overlapning, tøven, pauser, baggrundslyd, m.m. Nu, i slutningen af 2001, er samtalerne stort set færdigtransskriberet og lemmatisering af materialet er godt i gang. ISTAL skulle være et 3-års-projekt men eftersom der ikke er bevilget midler for det tredje år bliver det ikke muligt at gennemføre en helt så detaljeret analyse som planlagt. Materialet, dvs. lydfilerne og transkriptionen sådan som den står, vil dog inden længe blive gjort tilgængeligt for sprogforskere og det er klart at selv i nuværende tilstand bliver det en vigtig kilde. Og forhåbentlig bliver ISTAL kun det første skridt i retningen af en større islandsk talesprogsbank som vil i fremtiden udvides med flere slags talesprogsmateriale.

Ordforrådet i ISTAL sammenlignet med skriftsprog

Materiale og instrumenter

I denne afsnit præsenteres der en præliminær analyse af ordforrådet i en del af ISTAL-materialet, dvs. de første 23 samtaler der er blevet færdigtransskriberet. De udgør ca. 10 timers og 40 minutters optagelse, dvs. omtrent halvdelen af materialet.Teksten er blevet analyseret ved hjælp af det engelske tekstanalyseprogram *WordSmith Tools* (Scott 1996, Oaks

1998:193-4) som gør det bl.a. mugligt at lave ordlister og konkordanser og at sammenligne ordforrådet i to eller flere tekster eller tekstsamlinger.

For sammenligning med ISTAL-materialet valgtes der to forskellige skriftsprogsprøver, her kaldt SKRIFT-1 og SKRIFT-2. Hver prøve er en samling af mindre tekstafsnit som tilsammen svarer nogenlunde til størrelsen af ISTAL-prøven. Den første består af personligt og forholdsvis uformelt skriftsprog. Der er to slags tekster i denne prøve: på den ene side, en samling personlige dagbogsoptegnelser fra 1998 som opbevares på Nationalmuseets afdeling for folkloristik, og, på den anden side, eftermæle og fødselsdagsminder fra *Morgunblaðið*, den største avis i Island. De sidstnævnte tilhører en ganske særpræget islandsk tekstgenre. Der er tale om korte avisartikler der skrives af navngivne privatpersoner på en vens eller slægtnings fødsels- eller begravelsesdag. Som oftest har disse artikler et ganske personligt præg lige som dagbogsoptegnelserne. Den anden skriftsprogsprøve består af mere formelt og upersonligt skriftsprog. Alle teksterne er korte artikler fra *Morgunblaðið* som i avisens database klassifiseres som indenlandsk materiale, dvs. nyheder, meddelelser og den slags. De er skrevet af anonyme journalister om emner som ikke er af direkte personlig interesse for skriveren. Avismaterialet i begge skriftsprogsprøver blev hentet fra Leksikografisk instituts tekstsamling.

Resultat

Et overblik over de tre tekstprøver vises i tabel 1.

	ISTAL	**SKRIFT-1**	**SKRIFT-2**
Bytes	432.482	466.710	439.454
Løbende ord	52.012	77.633	64.659
Ordformer	5.846	14.470	13.173
Lemmaer	3.571	8.752	7.505

Tabel 1: Kvantitativ overblik over ISTAL-materialet sammenlignet med de to skriftsprogsprøver. SKRIFT-1 består af uformelt, personligt skriftsprog og SKRIFT-2 af mere formelt og upersonligt skriftsprog.

ISTAL-teksterne og den sidste skriftsprogsprøve er nogenlunde lige store med hensyn til antal bytes mens SKRIFT-1 er lidt større men regnet i antal løbende ord er talesprogsprøven noget mindre en de to andre. Der er alligevel væsentlig større forskel mellem skrift- og talesprog i antallet af forskellige ord og ordformer, fx. er ordforrådet i ISTAL kun omtrent halvdelen af det som findes i sammenligningsmaterialet efter en grov lemmatisering af ordlisterne. Tallene peger altså på at talesproget har et betydelig mindre ordforråd end skriftsproget.

I hele materialet er der omkring 15.000 forskellige lemmaer hvoraf omtrent 3.600 findes i ISTAL. Lidt over en tredjedel af disse, dvs. 37%, findes i alle tre tekstprøver. Langt de fleste af dem er temmelig frekvente og kun en mindre del af dem forekommer mindre end 10 ganger. Ord der er fælles med talesprogsmaterialet og den ene af skriftsprogsprøverne er henholdsvis 17% og 7% af ISTALs ordforråd, og ikke overraskende har talesproget flere ord tilfælles med det personlige, mere uformelle skriftsprog end med den anden prøve. Til gengæld forekommer næsten 40% af ordforrådet i ISTAL-samtalerne slet ikke i skriftsprogsteksterne og i det følgende vil vi se lidt nærmere på de pågældende ord.

Det skyldes nok en ren tilfældighed at mange af de pågældende ord udelukkende findes i samtalerne, for det er helt klart at ord som *áhyggjufullur* 'bekymret', *regnhlíf* 'paraply', *símaskrá* 'telefonbog', *tengdafaðir* 'svigerfar' også bruges i skriftsprog og ville sikkert forekomme i et større udvalg af skriftsprogstekster. Største delen af disse er sammensatte ord men der er også tale om simpleksord, fx. *brú* 'bro', *draugur* 'spøgelse', *kurteis* 'høflig', *florna* 'tørre', samt ganske mange person- og stednavne. Kun et mindretal af ordene som ikke optræder i skriftsprogstekssterne er mere typiske talesprogsord. Blandt dem finder man især fire typer eller kategorier:

1. Forskellige forstærkende adjektiver og adverbier.
   i. ord med intensiverende præfiks eller forled: *alleiðinlegasti* 'allerkedeligst', *dauðsyfjaður* 'dødsenssøvnig', *langreiðastur* 'allervredest', *skíthræddur* 'skidebange' osv.
   ii. ord med svækket og/eller overført betydning: *geðveikur*, *geggjaður* 'sindssyg, dvs. meget god', *glataður* '(for)tabt, dvs. meget dårlig', *ógeðslegur* 'modbydelig'.
2. Interjektioner som *almáttugur* 'herregud', *andskoti* 'pokkers' osv.
3. Slangudtryk, fx. *gaur* 'fyr', *gella* 'pige', *glætan!* 'hold kæft' osv.
4. Fremmedord
   iii. forholdsvis nye ord, gerne mindre tilpassede og/eller slangagtige: *aktífur* 'aktiv, virksom', *breinstorma* (verbum, jf. eng. *brainstorm*), *díler* 'dealer, pusher', *expert* 'ekspert' , *frílans* 'freelance', *streit* (jf. eng. *straight*), osv.
   iv. mere tilpassede og udbredte ord: *batterí* 'batteri', *partí* 'fest', *flass* 'blitzlys', *gratín* 'gratin', *pizza* 'pizza', *sería* 'serie' og *servíetta* 'serviet'.

I ISTAL-teksterne er der kun ét eller meget få eksemplarer af de omtalte ord og derfor er det en ganske spekulativ påstand at de er typiske for talesproget. Med *WordSmith*-programmet kan man sammenligne to tekster eller tekstsamlinger og finde ud hvilke ordformer karakteriserer den ene i forhold til den anden. Når man på den måde sammenligner ordlisterne fra ISTAL med skriftsprogsprøverne finder man ikke kun frem til ord og ordformer som findes i den ene og ikke den anden, men først og fremmest til ord som er meget mere frekvente i samtalerne end i de pågældende skriftsprogstekster. Også her viser det sig, at der er større forskel mellem ISTAL og det mere formelle og upersonlige skriftsprogsmateriale, dvs. SKRIFT-2, end mellem samtalerne og SKRIFT-1, selv om det er stort set de samme ordformer som i begge tilfælder viser sig som karakteristiske talesprogsudtryk. Der iblandt er der fx. forskellige former af personlige pronominer, især 1. og 2. person, såvel som tilsvarende former af nogle meget frekvente verber, som fx. *vera* 'være' og *vita* 'vide'. Derimod finder man næsten ingen substantiver eller adjektiver på disse lister. De mest interessante udtryk som er typiske for talesproget ifølge denne sammenligning kan deles i tre kategorier:

5. Ordene *já* 'ja' og *nei* 'nej' og varianter af dem (*jájá*, *neinei* osv.) samt *jæja* 'javel', det sidstnævnte dog kun karakteristisk i forhold til SKRIFT-2
6. Mange forskellige adverbier, f.eks. *bara* 'kun', *hérna* 'her', *flarna* 'der'; *svoleiðis*, *flannig* 'således, den slags'; *kannski* 'måske', *einmitt*, *akkúrat* 'netop'; *náttúrulega* 'naturligvis', *eiginlega* 'egentlig', *rosalega* 'forfærdelig (meget)'; *ókei* 'O.K'
7. Et par verber og verbale former der spiller en særlig rolle i talesproget: *heyrðu* (imperativ af heyra 'høre'), *ætli* 'mon' (en form af verbet *ætla* 'ville, skulle'; kun i forhold til SKRIFT-2), *meina* 'mene', *sko* 'se; sgu' (jf. verbet *skoða* 'betragte')

Tabel 2 viser frekvensen af nogle udvalgte ord og ordformer i de tre tekstprøver og tal fra den islandske frekvensordbog (IFO; Pind et al. 1991) følger også for at få sammenligning med en større skriftsprogssamling. Frekvensordbogen bygger på en tekstsamling på omtrent en halv million løbende ord fra 100 forskellige skriftsprogstekster.[3]

	ISTAL	SKRIFT-1	SKRIFT-2	IFO
jæja	37	20	0	86
bara	757	100	7	606
svoleiðis	65	6	0	23
kannski	128	57	7	405
einmitt	79	18	2	94
náttúr(u)lega	134	6	0	28
rosalega	44	2	0	5
ókei	16	0	0	1
heyrðu	65	0	0	20
meina	80	1	1	34
ætli	27	9	2	53
sko	670	5	0	52

Tabel 3: Frekvensen af nogle typiske talesprogsords i ISTAL sammenlignet med SKRIFT-1 og SKRIFT-2, som består af henholdsvis uformelt/personligt og formelt/upersonligt skriftsprog, og den islandske frekvensordbog (IFO). Ordformerne *jæja* og *ætli* anses ikke som karakteristiske for talesproget sammenlignet med SKRIFT-1men kun med SKRIFT-2.

Med hensyn til frekvensen af alle de udvalgte ord er der en klar forskel mellem talesprogsmaterialet på den ene side og de to skriftsprogsprøver på den anden, selv om forskellen ikke er signifikant mellem ISTAL og SKRIFT-1 med hensyn til to af ordene, *jæja* og *ætli*. Det er ikke mindre interessant at en sammenligning med en langt større skriftsprogssamling, dvs. frekvensordbogen, stort set ikke ændrer på billedet. Kun to af de udvalgte ord, der ifølge resultaterne kan anses som typiske talesprogsord, dvs. *bara* og *kannski*, har en rimelig stor udbredelse og frekvens ifølge den islandske frekvensordbog. I den pågældende skriftsprogssamling forekommer de henholdsvis godt 600 og 400 ganger, mens frekvensen af alle de andre ord og ordformer er under 100. Disse to ord findes i henholdsvis 71 og 75 af de 100 tekster som samling består af, mens ingen af de andre forekommer i mere end halvdelen af teksterne og de fleste i langt færre tekster.

ISTAL som redskab i ordbogsredaktion

Hensigten med ISTAL-projektet er at materialet skal være anvendelig til mange forskellige formål og vi skal slutte med en kortfattet vurdering af materialets betydning i ordbogssammenhæng. Nogle udvalgte ord og ordformer, som ifølge de foregående resultater er blandt de karakteristiske talesprogsord, er blevet slået op i *Íslensk orðabók*, den eneste almene ensproglige ordbog over islandsk, både i den nye, reviderede elektroniske udgave (2000) og i den sidste trykte udgave (1983), for at se efter om de fandtes og hvordan de er blevet behandlet med hensyn til ordenes brug og betydning i talesprogsmaterialet.

Først blev der set på nogle få af fremmedordene som forekommer i ISTAL og ikke i

skriftsprogsprøverne. De udvalgte ord er antagelig ganske hyppige og udbredte. Nogle af dem findes slet ikke i ordbogen, som fx. adjektivet *aktífur* 'aktiv', verberne *brillera* 'brillere' og *grilla* 'grille' (men derimod substantivet *grill*), og ligeledes adverbiet *akkúrat*, et af de karakteriserende talesprogsord ifølge vores resultater. Ordene *erótískur* 'erotisk' og *gratín* 'gratin' var heller ikke i 1983-udgaven af ordbogen men er blevet tilføjet i den nye. Til gengæld har faktisk største delen af fremmedordene som blev slået op sin plads i ordbogen.

De forstærkende adjektiver som kendetegner talesproget findes fleste i ordbogen men beskrivelsen er ofte ufuldstændig med hensyn til deres brug i samtalerne. Der gøres fx. ikke altid rede for den overførte betydning der er så karakteristisk for nogle af dem, jf. ordet *geðveikur* 'sindssyg'. Dette gælder ikke mindst adverbier som ofte er underordnede de tilsvarende adjektiver. I nogle tilfælder bliver adverbiet endog slet ikke nævnt, fx. *rosalega* og *ofboðslega*, begge to hyppigt brugt forstærkende med adjektiver. Jón Hilmar Jónsson (2000) har for nylig diskuteret adverbiernes status i islandske ordbøger og peget på nødvendigheden af at behandle dem på en anden og bedre måde. Denne præliminære undersøgelse af ISTAL-materialet peger på at det kunne være et værdifuldt indlæg i den sammenhæng og i det hele taget er det klart at talesprogsmateriale som ISTAL ville være en vigtig tilføjelse til ordbogskilder.

[1]Deltagere er Ásta Svavarsdóttir og Kristín Bjarnadóttir fra Leksikografisk institut (Orðabók Háskólans), Eiríkur Rögnvaldsson og fióra Björk Hjartardóttir fra det humanistiske fakultet (Heimspekideild Háskóla Íslands) og Hrafnhildur Ragnarsdóttir og Sigurður Konráðsson samt fiórunn Blöndal fra det pædagogiske universitet (Kennaraháskóli Íslands). Medhjælpere er Kolbrún Eggertsdóttir og Halldóra Björt Ewen.

[2]Selve optagelserne omfattede 36 samtaler på ca. 21 timer men 5 korte samtaler blev udeladt på grund af dårlig kvalitet.

[3]For at give en klarere idé om forskellen mellem omfanget af den nuværende undersøgelse og frekvensordbogen kan det nævnes at et af de mest hyppige ord, *að*, forekommer ca. 2.600 ganger i ISTAL, ca. 3.200 ganger i den ene af vores skriftsprogsprøver og 20.500 ganger i frekvensordbogen.

Bibliografi

Íslensk orðabók handa skólum og almenningi. 1983. Red. Árni Böðvarsson. [2. udg.] Reykjavík: Bókaútgáfa Menningarsjóðs.

Íslensk orðabók. Tölvuútgáfa. 2000. Red. Mörður Árnason. [Elektronisk udgave; 3. udg.]. Reykjavík: Edda h/f.

Jónsson, Jón Hilmar. 2000. Bráðum. I: *Orðhagi*, afmæliskveðja til Jóns Aðalsteins Jónssonar 12. október 2000, s. 72-79. Reykjavík.

Oakes, Michel P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pind, Jörgen [red.], Friðrik Magnússon og Stefán Briem. 1991. *Íslensk orðtíðnibók*. Reykjavík: Orðabók Háskólans.

Scott, Mike. 1996. *WordSmith Tools Manual*. Oxford: Oxford University Press.