Processering af en synonym-komponent i en flersproget ordbog

Halldóra Jónsdóttir & Þórdís Úlfarsdóttir

The ISLEX dictionary combines Icelandic as a source language (SL) and six Nordic languages as target languages (TLs) in a single database. This article describes the process of how Icelandic synonyms can be extracted and grouped together by using the TLs' equivalents. Each word class is processed independently. Selected material, a list of lemmas with their ID number plus their TLs, is exported from the database, going through the languages one at a time. This produces long lists of results which consist of the Icelandic headwords and their equivalents. The lists are then processed in certain ways. In cases where a lemma has more than one sense there is a danger that the boundaries between the senses become blurred so that the words in those groups tend to get mixed up. An example is the Icelandic noun *verslun* which means both 'shop' and 'trade', and is therefore divided into two senses in the dictionary. All the material comes from within the dictionary and no external data is used. It means that all the synonyms obtained in this way are necessarily also lemmas. Despite this we consider the method justifiable because it is quick and simple, and in many instances it gives very good results, i.e. groups of synonyms with varying degrees of exactness.

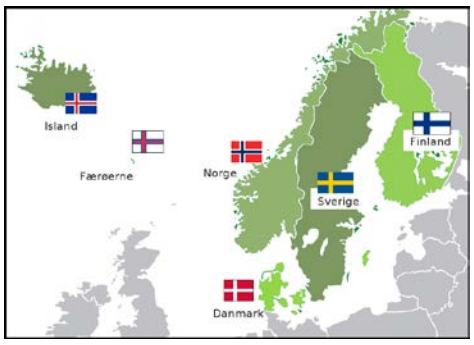
1. Baggrund

ISLEX er en tværnordisk, flersproget onlineordbog med islandsk som kildesprog og dansk, norsk bokmål, nynorsk, svensk, færøsk og finsk som målsprog. ISLEX er et samarbejdsprojekt mellem seks institutioner i Island, Danmark, Norge, Sverige, Finland og på Færøerne. Disse institutioner er *Stofnun Árna Magnússonar í íslenskum fræðum* (SÁM) i Reykjavík, *Det Danske Sprog- og Litteraturselskab* (DSL) i København, *Institutt for lingvistiske, litterære og estetiske studier* ved Universitetet i Bergen, *Institutionen för svenska språket* ved Universitetet i Göteborg, *Fróðskaparsetur Føroya* i Tórshavn og *Helsinki Universitet*. De to

sidstnævnte institutioner indgik dog i samarbejdet på et senere tidspunkt end de førstnævnte.

ISLEX-ordbogen blev udarbejdet på den måde at den islandske redaktion var ansvarlig for kildesproget samt udformning og udvikling af databasen. Redigeringen af målsprogene derimod blev varetaget af oversætterne i hvert af de øvrige lande. Arbejdet foregik i en webbaseret database specielt lavet til projektet, således at alle landenes redaktioner kunne arbejde samtidig.

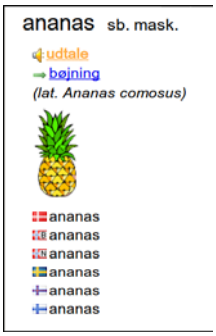
Den første version af ISLEX blev åbnet på webben i november 2011 (for målsprogene dansk, norsk og svensk), og den islandsk-færøske ordbog i marts 2015. Den islandsk-finske ordbog er stadig under bearbejdning og bliver åbnet senere.

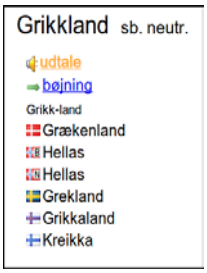


Figur 1: Landene som deltager i ISLEX-samarbejdet.

ISLEX-ordbogen er den første onlineordbog der omfatter og forbinder flere nordiske sprog. Ordbogen dækker moderne islandsk sprog, med særlig vægt på at repræsentere et mangfoldigt udvalg af kollokationer, idiomer og eksempler med tilhørende oversættelser til målsprogene. ISLEX udnytter de fordele som elektronisk formidling giver mulighed for så som illustrationer og lyd. Der vises bøjningsparadigmer til alle de opslagsord der kan bøjes, via et link til den morfologiske database over islandsk, *Beygingarlýsing íslensks nútímamáls*, som er udarbejdet hos SÁM.

På ordbogens hjemmeside er der forskellige typer af søgemuligheder. Man kan f.eks. vælge at foretage opslag i alle målsprogene samtidigt, for således at få et indblik i det islandske sprogs slægtskab med målsprogene og ligeledes et overblik over indbyrdes forskelle og ligheder mellem sprogene.





Figur 2: Ordbogen kan belyse ligheder og forskelle mellem sprogene. *Ananas* er ens på alle sprogene.

Figur 3: Variation i navnet *Grækenland*.

Fra starten ønskede man at ISLEX blev et bidrag til at styrke de kulturelle forbindelser i Norden og fremme den internordiske sprogforståelse, og det fremgår klart at efter at ordbogen blev åbnet, har den fået stor udbredelse og titusinder af brugere. Desuden gemmer et så omfattende værk på mange muligheder for al slags sprogforskning, og i denne artikel beskrives et nyt projekt: en synonym-komponent for kildesproget, islandsk.

2. Leksikografisk placering

Synonymordbøger i deres enkleste form får almindeligvis ikke særlig stor opmærksomhed indenfor leksikografien, og denne type ordbøger har hidtil ikke nydt stor respekt. Hovedårsagen er sandsynligvis de begrænsede relationer i selve ordforrådet, hvor alle ordene indtager den samme position i en flad struktur. Men selvom synonymordbøgerne måske ikke anses for at være udpræget videnskabelige værker, er de af stor nytte for mange brugere, f.eks. når de skal producere tekster eller løse et krydsord.

Tesaurusser og begrebsordbøger er nært beslægtede med synonymordbøgerne, men de er normalt mere indholdsrige, og indholdet inddeles gerne på en anden måde, oftest hierarkisk under overbegreber. I sådanne ordbøger er der plads til forskellige ord som ikke hører hjemme i synonymordbøger, f.eks. navne på måneder og træsorter. Tesaurus er en særlig ordbog eller

'ordliste hvor ordene er ordnet efter emne og betydning (og altså ikke alfabetisk)' (jf. forklaringen af *tesaurus* i DDO). Eftersom en begrebsordbogs indhold ikke ordnes alfabetisk, er det vigtigt at inkludere et indeks over emnet for at gøre det nemmere for brugerne at finde frem til det som de søger. Dette gælder først og fremmest trykte bøger, og eksempler på dette finder man i Jónsson 2005 og DDB 2014, som begge indeholder et kæmpemæssigt indeks. I de rendyrkede synonymordbøger ordnes ordene derimod alfabetisk hvilket gør det nemmere at finde frem til det man søger.

På islandsk findes der både en stor begrebsordbog (Jónsson 2005) og en synonymordbog (Sigmundsson 2012). Nærværende projekt er således ikke fremkaldt af et akut behov. Projektet er i højere grad en undersøgelse af hvorvidt det er formålstjenligt eller overhovedet muligt at generere synonymgrupper ud af de sprogdata som gemmer sig i ISLEX. På nuværende tidspunkt er det altså ikke meningen at udarbejde en fuldstændig digital synonymordbog, og ej heller en begrebsordbog, men redaktionen anser det som værende et spændende pilotprojekt som muligvis kan danne grundlag for noget større.

3. Materiale til synonymgrupperne

Der er ca. 50.000 oplagsord i ISLEX som dermed bliver kandidater til synonym-processeringen. Hele materialet stammer fra selve databasen og der hentes ingenting udefra, et faktum som selvfølgelig begrænser antallet af enkelte synonymer. På nuværende tidspunkt medtages ikke flerordsenheder som ofte er fyldestgørende synonymer til enkelte ord (jf. dansk *med det samme* og *straks*), og desuden har en del af ordforrådet ikke noget synonym.

Eftersom ISLEX-ordbogen har islandsk som kildesprog er den primære tilgang til ordforrådet gennem et islandsk opslagsord (Figur 4). Men det er også muligt at søge omvendt, dvs. ved at indtaste et ord fra et af målsprogene, hvor der efterfølgende fremstår en liste over de islandske ordbogsartikler som indeholder den ønskede ækvivalent, m.a.o. et udvalg af islandske opslagsord, som matcher søgningen. Figur 5 viser et eksempel på dette.

Figur 4: Et islandsk opslagsord, ’skrímsli’, med to danske ækvivalenter, ’uhyre’ og ’monster’.

Figur 5: Omvendt søgning, ’monster’, på dansk og svensk, med de islandske opslagsord efter flaget.

Det er dog ikke altid muligt at opnå et så omfattende søgeresultat som på Figur 5 når man søger et ord i målsprogene. I nogle tilfælde er der bare ét ord i målsproget imod ét andet ord i kildesproget (*historielærer*, *onsdag*). Derimod giver visse (ofte ukonkrete) ord i mange tilfælde udmærkede resultater, f.eks. det danske ord *adfærd* som afføder tolv gode islandske synonymer (Figur 6).

Figur 6: Søgeresultat for det danske ord ’adfærd’. Det matcher 12 islandske opslagsord.

Ordgrupperne foroven fremkommer når man søger i ordbogens bruger-grænseflade efter et ord på et af målsprogene. Dermed aktiveres en kom-

mando til databasen som besvarer søgningen og frembringer ækvivalent sammen med lemma og lemma-ID som knyttes dertil. Listen for neden viser outputtet fra en direkte SQL-forespørgsel om to bestemte danske ækvivalenter, der er foretaget udenom ordbogens hjemmeside. Rækkefølgen i nedenstående liste er: ækvivalent, ordklasse (no = navneord) og opslagsordet med dets lemma-ID. Dermed opstår der en liste med to danske ord, *adskillelse* og *adspredelse*, hvor de islandske match bliver potentielle synonymer:

adskillelse no	aðgreining: 2328
adskillelse no	aðskilnaður: 2397
adskillelse no	skilnaður: 36414
adskillelse no	sundurgreining: 55696
adskillelse no	viðskilnaður: 45572
adspredelse no	afþreying: 2817
adspredelse no	dægrastytting: 9027
adspredelse no	dund: 8824
adspredelse no	upplyfting: 43476
adspredelse no	yndisauki: 46608

Hermed får vi en bekræftelse af at denne metode er oplagt til at hente islandske synonymer frem med. Eftersom opslagsordene ikke er systematisk forbundet gennem selve kildesproget, er ækvivalenterne i mange tilfælde den stærkeste forbindelse mellem ordene[1]. Databasen indeholder flere tusinde ækvivalenter på hvert af ordbogens seks målsprog, og det er klart at det drejer sig om store mængder af sproglige data som kan anvendes til at generere synonymgrupper, som vises i Figur 5 og 6. Herunder beskrives nærmere den metode som anvendes for at trække synonymerne ud af databasen.

[1] Ordforrådet er i forvejen inddelt i semantiske felter, men inddelingen kan kun i begrænset omfang bruges til at gruppere ordene med.

4. Metode

På samme måde som beskrevet ovenfor gennemgås alle målsprogene, ét for ét, og således knyttes synonymgrupperne for islandsk sammen, med lemmaet som det centrale omdrejningspunkt. Som før nævnt er målsprogene i ISLEX flertallet af de sprog som tales i Norden (dog ikke grønlandsk og samisk). Finsk er det eneste målsprog i databasen som ikke er af samme sprogfamilie som de øvrige sprog. Til trods for dette er resultaterne af processeringen af finsk lige så gode som for de andre sprogs vedkommende, men for enkelhedens skyld fokuserer vi mest på dansk i denne artikel. Processen ved at søge potentielle synonymer sker trinvist, og her beskrives metoden i store træk.

1. Det første trin er en eksport fra databasen: *ækvivalent* + *ordklasse* + *opslagsord* + *lemma-id*. Dette gentages for hver ordklasse for sig og hvert målsprog for sig. Resultatet bliver en liste som den som vises herunder, i dette tilfælde drejer det sig om adjektiver på dansk med et tilsvarende islandsk lemma:

magtløs adj	magnvana: 51413
manglende adj	skertur: 36368
mangelfuld adj	ófullkominn: 31010
metamorfoseret adj	myndbreyttur: 52575
modig adj	skeleggur: 36272
morsom adj	fyndinn: 14317
morsom adj	gamansamur: 14794
morsom adj	skemmtilegur: 36328
motordrevet adj	vélknúinn: 45434
mættet adj	gegnsósa: 15104

2. Næste skridt er at sammenføje lister med resultater fra alle de seks målsprog og ordne dem alfabetisk. På denne måde blandes alle målsprogene sammen i en alfabetisk liste. Den ser sådan ud:

morsk adj	illúðlegur: 22448
morsk adj	yggldur: 46594
morsom adj	broslegur: 7016
morsom adj	fyndinn: 14317
morsom adj	hlægilegur: 20177
morsom adj	kátlegur: 23641
morsom adj	kímilegur: 23868
morsom adj	skemmtilegur: 36328
morsom adj	skondinn: 36792
morsom adj	skoplegur: 36799
morsom adj	sniðugur: 38129
morsom adj	spaugilegur: 38588
mosbevokset adj	mosagróinn: 57790

3. Nu *køres* to små perl-scripter *på resultatlisten. Det første script opstiller listens islandske ord (med ID) i samme linje med et komma imellem, efter den danske ækvivalent plus ordklasse*.

> morsom|adj|broslegur: 7016, fyndinn: 14317, gamansamur: 14794, hlægilegur: 20177, kátlegur: 23641, kímilegur: 23868, kostulegur: 24490, skemmtilegur: 36328, skondinn: 36792, skoplegur: 36799, sniðugur: 38129, spaugilegur: 38588

4. Dernæst bliver det sidste script *kørt på resultaterne.* Scriptet fjerner ækvivalenten (*morsom*) først i linjen og danner par med de ovenstående islandske ord:

broslegur: 7016	fyndinn: 14317
fyndinn: 14317	broslegur: 7016

broslegur: 7016	gamansamur: 14794
gamansamur: 14794	broslegur: 7016
broslegur: 7016	hlægilegur: 20177
hlægilegur: 20177	broslegur: 7016
broslegur: 7016	kátlegur: 23641
kátlegur: 23641	broslegur: 7016

Det er meningen at alle synonymerne fremstår i alle ordbogsartiklerne i listen, og processen skaber en fordoblingseffekt i parrene eftersom alle kombinationer skal parres. I de tilfælde hvor der kun findes to synonymer (*broslegur, fyndinn*) bliver linjerne kun to:

broslegur	fyndinn
fyndinn	broslegur

Hvis der derimod er tre synonymer (*broslegur, fyndinn, gamansamur*) bliver linjerne seks:

broslegur	fyndinn
broslegur	gamansamur
fyndinn	broslegur
fyndinn	gamansamur
gamansamur	broslegur
gamansamur	fyndinn

5. Til slut må man gennemgå hele materialet omhyggeligt fordi metoden indebærer at der altid vil være ord som ikke passer ind. Derefter er materialet klart til indlæsning i databasen eller til et andet formål.

5. Bearbejdning af resultaterne

Der er nu blevet genereret lange lister med materiale til synonymgrupper som er færdige til gennemlæsning. Den mest tidskrævende faktor er at fjerne de ord som ikke passer til synonymlisterne, og som udgør omkring 19% af alle ordene. Det ville faktisk være muligt at opnå en større procentdel af brugbart materiale ved ikke at anvende alle målsprogene i ISLEX som forbindelesled mellem de islandske ord, noget som også ville medføre at der fremstod færre synonymer end ellers. Det kræver en vis balancegang at finde frem til den bedste arbejdsmetode.

Under processeringen sker det tit at falske venner dukker op i listerne. I de fleste tilfælde drejer det sig om at ord af samme stamme i nærtbeslægtede sprog ikke har den samme betydning, f.eks. dansk og norsk *rolig* 'stille' og svensk *rolig* 'morsom'. Et andet forstyrrende moment er polysemien blandt ækvivalenterne, f.eks. betyder svensk *affär* både 'butik, forretning' og '(kærligheds)affære', og det danske ord *brud* betyder både 'afbrydelse' og 'en kvinde som skal giftes'. På dette sidste stadie spiller redaktøren hovedrollen for at sikre at kun de relevante ord indgår i listen.

For den ikke-islandske sprogbruger kan det være svært at vurdere hvor gode eller nære synonymerne er, og hvad angår deres valør, stilleje, frekvens osv. ville det uden tvivl være nyttigt at få oplysninger derom. Eftersom det drejer sig om automatisk generering, så er dette ikke så ligetil på nuværende stadie.

6. Konklusion

Ovennævnte projekt kan beskrives som et pilotprojekt, hvor man afprøver en metode ved generering af synonymer igennem målsprogenes ordforråd i en flersproget ordbog. Resultaterne viser at det er en effektiv og hurtig metode til dannelse af synonymer. Projektets materiale stammer alene fra indholdet i ISLEX og hvis man ønsker at udvide synonymkomponenten med flere ord, må disse hentes udefra. Desuden skal der gøres specielle tiltag for at det bliver muligt at inkorporere flerordsenheder i synonymgrupperne.

Projektet er i skrivende stund under forberedelse, men planen er at komponenten kommer som en tilføjelse til ISLEX-ordbogen og dets afledte

værker. Et af disse projekter er KATA, en ny islandsk webordbog som er baseret på det islandske materiale i ISLEX. Arbejdet med en ny islandsk-fransk ordbog med arbejdstitlen 'Lexia' er også påbegyndt som et spin-off-projekt af ISLEX. Alle disse værker vil nyde godt af en videre udvikling af ISLEX-ordbogens omfattende materiale.

Litteratur

Ordbøger

DDB (2014) = Nimb, Sanni (hovedred.): *Den Danske Begrebsordbog*. København: Det Danske Sprog- og Litteraturselskab.

DDO = Trap-Jensen, Lars et al. (red.): *Den danske ordbog.* København: Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (september 2015).

ISLEX = Úlfarsdóttir, Þórdís (red.): *ISLEX-ordbogen.* Reykjavík: Árni Magnússon-instiuttet for islandske studier. <islex.is>, <islex.dk> (september 2015).

Jónsson, Jón Hilmar. (2005): *Stóra orðabókin um íslenska málnotkun.* Reykjavík: Forlagið.

Sigmundsson, Svavar (2012): *Íslensk samheitaorðabók* (3. ed.). Reykjavík: Forlagið.

Halldóra Jónsdóttir
projektleder
Árni Magnússon-instituttet for islandske studier
Afdeling for leksikografi
Laugavegur 13
IS-101 Reykjavík
halldo@hi.is

Þórdís Úlfarsdóttir
hovedredaktør
Árni Magnússon-instituttet for islandske studier
Afdeling for leksikografi
Laugavegur 13
IS-101 Reykjavík
disa@hi.is