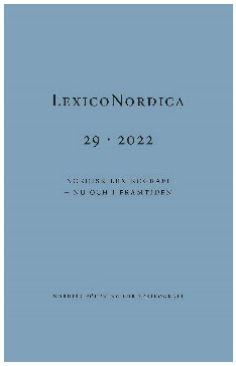
LexicoNordica

Titel:	Dannelsen af en tosproglig ordbog med hjælp af sprogteknologiske metoder
Forfatter:	Þórdís Úlfarsdóttir & Steinþór Steingrímsson
Kilde:	LexicoNordica 29, 2022, s. 153-173
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Dannelsen af en tosproglig ordbog med hjælp af sprogteknologiske metoder

Þórdís Úlfarsdóttir & Steinþór Steingrímsson

The article discusses a new approach in Icelandic bilingual lexicography where English as a target language (TL) is produced by using technological means. In order to generate the English equivalents, various TLs from online dictionaries are used as pivot languages, among other methods. Icelandic phrases and examples are translated by translation tools. Thereafter, lexicographers do necessary post-processing of the TL, monitor the quality of the work and finalise each article.

1. Indledning

Árni Magnússon-instituttet for islandske studier (AMI) har i årenes løb udgivet ordbøger eftersom leksikografi falder ind under instituttets faste opgaver. Gennem instituttets webside er der adgang til adskillige digitale ordbøger: den monolingvale *Íslensk nútímamálsorðabók* (Ordbog over moderne islandsk); de tosprogede ordbøger ISLEX (Úlfarsdóttir 2013, 2014), hvor kildesproget er islandsk og målsprogene er dansk, norsk, svensk, færøsk og finsk, som er samlet i én og samme database; samt ordbogen LEXIA mellem islandsk og henholdsvis fransk og tysk. Til trods for denne omfattende udgivelse af digitale ordbøger er der dog en åbenlys mangel, da der ikke findes en ordbog mellem islandsk og engelsk blandt instituttets udgivelser. På det islandske ordbogsmarked findes der selvfølgelig ældre ordbøger mellem islandsk og engelsk, men de imødekommer ikke længere nutidens krav om tilgængelighed, størrelse og løbende opdateringer. Man besluttede derfor at starte et nyt projekt, en islandsk-engelsk ordbog, og samtidig gøre et forsøg med at anvende nye metoder som man ikke har brugt tidligere på AMI.

I de seneste år og årtier har man i Island, ligesom i andre lande, oplevet en hastig udvikling inden for sprogteknologien, og AMI tager aktivt del deri. Forudsætningerne er at der foreligger gode sproglige ressourcer, og ordbogsmateriale er i vidt omfang blevet anvendt i sprogteknologi inden for forskellige sprog. Der er stadig større sammenfald mellem sprogteknologi og leksikografi, og i den forbindelse kan man bl.a. pege på den europæiske platform ELEXIS (European Lexicographic Infrastructure).

I takt med dette opstod den idé at anvende sprogteknologi ved udarbejdelsen af en ny islandsk-engelsk ordbog og hente ressourcerne blandt de sproglige data som AMI har opbygget i årenes løb, nemlig en ordbogsbase med 54.000 opslagsord som danner grundlag for de ovenfornævnte ordbøger. I denne base forefindes kildesproget, islandsk, med fuld ordbogsbeskrivelse, dvs. ordforklaringer, sprogbrugseksempler, faste udtryk osv. Projektet var desuden et interessant leksikografisk eksperiment, da de traditionelle metoder, som bekendt, er meget tidskrævende.

I denne artikel beskrives det islandsk-engelske eksperiment, både den tekniske side og selve ordbogsarbejdet. Artiklens første del omhandler de tekniske forudsætninger der ligger bag projektet, som for det første går ud på at danne ordpar mellem islandsk og engelsk (generering af engelske ækvivalenter), og for det andet at oversætte faste udtryk og sprogbrugseksempler med oversættelsesmaskiner. Herefter importerer hovedredaktøren de engelske ækvivalenter i en færdig islandsk kildesprogsdatabase. I sidste halvdel af artiklen drøftes den efterfølgende proces hvor redaktørerne tager over og foretager den endelige udvælgelse af engelske ækvivalenter og færdigredigerer artiklerne.

Artiklen er inddelt i følgende kapitler: I kapitel 2 gøres der rede for dannelsen af en ordliste mellem islandsk og engelsk, og kapitel 3 handler om maskinoversættelser af sprogbrugseksempler og faste udtryk. Kapitel 4 behandler dels det redaktionelle ordbogsarbejde efter at materialet med ordlisterne foreligger, dels projektets

vigtigste leksikografiske udfordringer. Kapitel 5 er en evaluering af dannelsen af en tosproget ordbog efter denne metode, og kapitel 6 indeholder konklusion.

2. Dannelsen af ordlister: metoder og data

Der blev udarbejdet en stor islandsk-engelsk ordliste som en del af Handlingsplan for sprogteknologi for islandsk (Nikulásdóttir, Guðnason & Steingrímsson 2017). Handlingsplanen strækker sig over årene 2018-2022. Handlingsplanen fremhæver behovet for at opbygge en infrastruktur for islandsk sprogteknologi, bl.a. ved dannelsen af forskellige sproglige databaser. Herunder falder den islandsk-engelske ordliste som giver muligheder for at blive brugt i projekter tilknyttet maskinoversættelser og ved opbygning af gode databaser til træning af oversættelsesmaskiner, f.eks. for at forbedre udvælgelsen af de sætningspar der bruges til træningen. Ordlisten kommer endvidere til gavn i software der bruges til at hente oplysninger i flersprogede databaser (Cross-language information retrieval) eller til sentimentanalyser. Den anvendte islandsk-engelske ordliste blev udarbejdet på AMI i 2021. Man brugte automatiske metoder til at generere en liste med forslag til ordpar, som siden blev gennemgået manuelt af en medarbejder for at godkende ordparrene som skulle opfylde én bestemt betingelse: at ordene kan anses som ækvivalenter i en bestemt kontekst.

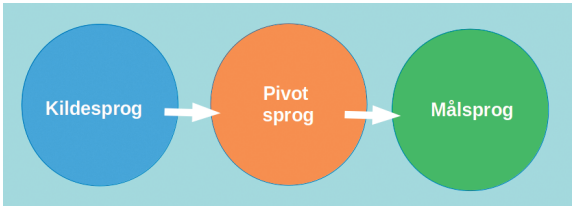
Denne ordliste blev grundlaget for de engelske ækvivalenter i den islandsk-engelske ordbog. Ved generering af ordlisten blev der anvendt fire forskellige metoder som bliver beskrevet i de efterfølgende afsnit: brug af pivot-sprog, parallelkorpusser og to slags maskinoversættelser.

2.1. Pivot-sprog

Den første metode til dannelsen af en islandsk-engelsk ordliste bestod i brugen af såkaldte pivot-sprog, som indebærer at de engelske ækvivalenter hentes igennem et tredje sprog som mellemled. Pivotsproget danner dermed en sproglig bro. Wikipedias definition af *pivot language* er følgende:

> A pivot language, sometimes also called a bridge language, is an artificial or natural language used as an intermediary language for translation between many different languages – to translate between any pair of languages A and B, one translates A to the pivot language P, then from P to B.

Da der var tale om engelsk som målsprog, var det nemmere end ellers, eftersom engelsk er et verdensomfattende sprog, og der foreligger en stor mængde sproglige data som knytter engelsk til andre sprog. AMI har adgang til sproglige data for forskellige målsprog som er blevet til gennem arbejdet med de tosprogede ordbøger. Målsprogene er de seks målsprog i ISLEX foruden de to målsprog i LEXIA. De fleste af disse sprog kunne anvendes ved genereringen af det engelske ordforråd.



Figur 1: Processen fra kildesprog (islandsk) via pivotsprog (nordiske sprog m.m.) til målsprog (engelsk).

At bruge et pivot-sprog er ikke en ny metode, men den er tilsyneladende ikke blevet anvendt inden for leksikografien i udstrakt grad. En af grundene til dette er selvfølgelig at metoden ikke var realiserbar indtil ca. år 2000 hvor ordbøger for alvor blev digitale. Verdens mest anvendte sprog, engelsk, rådede allerede over et udvalg af bilingvale ordbøger, og behovet var derfor evt. ikke så presserende. For sprog som ikke råder over så mange sproglige data, har pivot-metoden derimod været gavnlig, se bl.a. Varga & Yokoyama (2009) som brugte metoden mellem japansk og ungarsk med engelsk som pivot-sprog, og Aker et al. (2014) mellem engelsk og tysk, med fransk som pivot-sprog; se desuden Gamallo & Campos (2010).

I dette projekt anvendte man de bilingvale ordbøger med islandsk som kildesprog, og målsproget blev brugt som pivot-sprog. Siden blev ordbøger hvor pivot-sproget var kildesprog og engelsk var målsprog, parret sammen med de islandske opslagsord. På den måde blev det muligt at generere ækvivalenter til en stor del af kildesprogets opslagsord (islandsk). Her har man dog det problem at når det islandske ord, eller ordet på pivot-sproget, har mere end én betydning, får man forkerte ækvivalenter sammen med de korrekte. Denne metode alene bliver derfor aldrig særlig præcis. De brugte ordbøger var de islandsk-skandinaviske ordbøger i ISLEX samt de islandsk-fransk/tyske ordbøger i LEXIA. De ordbøger der blev brugt til at oversætte fra pivot-sprogene til engelsk, var *Apertium* (norsk/svensk/finsk/fransk/engelsk) og *dict.cc* (norsk/svensk/finsk/fransk/tysk/engelsk). *Apertium* er en open-source platform for maskinoversættelse. Den indeholder sproglige data, herunder ordbøger for et stort antal sprog. *Dict.cc* er en onlineordbog der er opbygget ved frivillig indsats. Den indeholder et stort antal sprogpar. Disse ordbøger er ikke blevet udarbejdet af professionelle leksikografer.

	Apertium		dict.cc	
	præcision	antal par	præcision	antal par
norsk	53 %	15.261	74 %	31.213
svensk	64 %	34.915	76 %	26.622
finsk	43 %	214.659	75 %	19.304
fransk	63 %	20.865	64 %	39.590
tysk			54 %	137.970

Tabel 1: Tabellen viser størrelsen og vurderet præcision af de lister med forslag til ordpar med engelsk der blev genereret igennem hvert sprog og hver ordbog.

Tabel 1 viser det antal forslag som hver metode gav og præcisionen, evalueret efter 500 tilfældigt valgte ordpar fra hver liste. Da det er tidskrævende at gennemgå listerne, ønsker man at projektets medarbejdere præsenteres for et materiale der allerede har en høj præcision. Ved kun at vælge ordpar der blev genereret med mange forskellige gennemgange af flere forskellige ordbogsresurser, var det muligt at opnå større præcision, men samtidig blev forslagene selvfølgelig færre. Ved f.eks. kun at vælge forslag som blev opnået ved at gennemgå svensk og fransk, fik man en liste med 11.274 forslag som blev evalueret til at være 97 % egnet. Andre kombinationer gav mindre præcision, men der blev dog fundet nogle kombinationer med over 90 % præcision. Alle forslag der blev genereret, blev gennemgået manuelt og accepteret eller afvist.

2.2. Oversættelsesmaskiner

Den anden metode til at fremkalde engelske ækvivalenter var at anvende oversættelsesmaskiner direkte på lemmaerne i ISLEX og oversætte ordene til engelsk. Alle de islandske opslagsord blev samlet til en liste som blev oversat til engelsk ved hjælp af to oversættelsesmaskiner som er tilgængelige på nettet, Google Translate

og Microsoft Translator. En gennemgang af 500 tilfældigt valgte ordpar fra hver oversættelsesmaskine gav 59 % præcision for Google Translate og 60 % præcision for Microsoft Translator.

2.3. Ord på pivot-sprog oversat til engelsk i oversættelsesmaskiner

Den tredje metode indebærer at man laver en liste med samtlige ækvivalenter af de islandske opslagsord i ISLEX og LEXIA-ordbøgerne, som beskrevet i afsnit 2.1, men i stedet for at slå pivotordene op i andre ordbøger oversætter vi dem med oversættelsesmaskiner. Her brugte vi fire forskellige oversættelsesmaskiner, Google Translate og Microsoft Translator (MS) ligesom før, men desuden en oversættelselsmodel fra OPUS-MT (Tiedeman & Thottingal 2020) samt M2M modellen (Fan et al. 2021). Microsoft Translator gav de bedste resultater, over 60 % præcision for alle pivotsprog, M2M gav derimod de ringeste resultater. Se tabel 2.

	Opus	**M2M**	**Google**	**MS**	**Samlet antal**
dansk	52 %		59 %	63 %	80.074
norsk			59 %	61 %	66.129
svensk	56 %	32 %	65 %	65 %	69.884
finsk	53 %	27 %	66 %	62 %	62.876
fransk	56 %	35 %	67 %	71 %	45.533

Tabel 2: Proportion af brugbare ordpar i 500 tilfældigt valgte par for hver oversættelsesmaskine og pivot-sprog.

2.4. Parallelkorpusser

Den fjerde metode til dannelsen af en islandsk-engelsk ordliste var at fange modsvarende ord i sætningspar i parallelkorpusser. Parallelkorpusser er tekster på to sprog, hvor den ene tekst er en

oversættelse af den anden. Teksterne er delt op i sætninger, og modsvarende sætninger sidestilles. Ved at undersøge et stort antal sætningspar er det muligt at finde modsvarende ord i sætningerne. På samme måde er det muligt at anvende tekster hvor der ikke er tale om oversættelser, men om tekster der forholdsvis specifikt handler om samme emne (comparable corpora), men her må man bruge andre metoder for at parre sætningerne, og der behøves mere materiale for at få et acceptabelt resultat (se f.eks. Steingrímsson et al. 2021).

For at opnå størst mulig præcision ved at finde modsvarende ord i sætningsparrene kører vi data igennem fem forskellige automatiske ordaligneringsværktøjer og vælger derefter de ordpar som de fleste af værktøjerne er enige om er de rigtige, og smider de resterende ud. Når man på den måde har dannet ordpar i alle sætninger i hver database, regner man points ud for hvert ordpar. Pointene bliver udregnet ud fra hvor tit ordparringsværktøjet danner ordpar proportionalt med hvor tit ordene forekommer i samtlige sætninger. Parrene bliver så accepteret, eller de bliver forkastet på grundlag af om scoren opnår et vist minimum som man finder frem til ved at gennemgå en lille del af forslagene og undersøge kvaliteten af ordparrene med visse mellemrum.

Der bliver anvendt seks tekstkorpusser, af tre forskellige slags: et parallelkorpus, ParIce (Barkarson & Steingrímsson 2019), tre sammenlignelige sprogkorpusser (comparable corpora) som allerede er blevet inddelt i sætningspar, WikiMatrix (Schwenk et al. 2021) og Paracrawl 7,1 og Paracrawl 8 (Bañón et al. 2020). Til slut bruges der to syntetiske korpusser (synthetic corpora) der bliver dannet via en oversættelsesmaskine der oversatte forskellige nyhedstekster henholdsvis fra islandsk til engelsk og fra engelsk til islandsk (Símonarson et al. 2020). Denne proces resulterede i sætningspar fra alle seks korpusser, hvor sætninger forfattet af personer blev oversat af en maskine, jf. tabel 3.

		Tillidsscore hvor over 50% af parrene er brugbare		
Korpus	**Samlet antal par**	**Godkendelse i %**	**Antal par**	**Forventet antal brugbare**
ParIce	346.723	51,6 %	45.646	25.553
Paracrawl 7.1	107.959	59,6 %	70.281	41.887
Paracrawl 8	342.444	62,6 %	93.850	58.750
WikiMatrix	15.781	77,2 %	6.944	5.360
Syntetisk data is-en	191.934	67,2 %	13.215	8.880
Syntetisk data en-is	229.661	60,2 %	132.381	79.693

Tabel 3: Antal ordpar, forhold mellem brugbare par og anslået antal brugbare par i forskellige korpusser.

2.5. Analyse af resultaterne fra de automatiske metoder

Der blev taget stikprøver fra resultaterne af samtlige automatiske metoder for at vurdere præcisionen af dem, som det fremgår af tabellerne foroven. For at reducere arbejdsindsatsen ved gennemgang af oversættelsesforslagene anvendtes kun de lister hvor formodet præcision var meget stor. Tilbage stod der et stort antal ordpar, og for at udnytte materialet mest muligt delte vi listen med forslag i to kategorier, på den ene side de forslag som blev genereret i arbejdet med korpusserne, og de forslag som blev genereret med maskinoversættelse eller ordbogsopslag gennem pivot-sprog på den anden side. Der blev siden udarbejdet en ny liste som kun indeholdt forslag som forekom i begge kategorier. På den måde fik man en ny liste med knap 30.000 par som blev anslået til 93,2 % præcision, se tabel 4.

		Også opnået med ordbøger igennem pivot-sprog eller maskinoversættelse		
Korpus	**Samlet antal par**	**Godkendelse i %**	**Antal par**	**Forventet antal brugbare**
ParIce	45.646	90,4 %	3.713	3.356
Paracrawl 7.1	70.281	95,8 %	18.836	18.045
Paracrawl 8	93.850	96,2 %	16.522	15.894
WikiMatrix	6.944	97,4 %	3.343	3.256
Syntetisk data is-en	13.215	97,3 %	4.986	4.851
Syntetisk data en-is	132.381	94,4 %	19.423	18.335

Tabel 4: Antal ordpar, forholdet mellem brugbare par og anslået antal brugbare par i forskellige korpusser, som også blev genereret med andre metoder.

3. Maskinoversættelse af sprogbrugseksempler og faste udtryk

I skrivende stund er systematisk arbejde med oversættelse af sprogbrugseksempler og definitioner ikke begyndt. For at fremskynde dette arbejde har vi oversat alle sprogbrugseksempler og definitioner med fire forskellige oversættelsesmaskiner, både direkte fra islandsk og igennem pivot-sprogene i ISLEX og LEXIA. Det første skridt bliver at gennemgå 1000 tilfældigt valgte islandske sætninger og undersøge om oversættelsesmaskinerne har leveret brugbare oversættelser for nogle af sætningerne. De brugbare sætninger bliver udvalgt, og det undersøges hvilke oversættelsesmaskiner og metoder der giver det bedste udfald. Resultaterne bruges siden til at inddele listerne med forslag i prioritetsorden således at de bedste oversættelser står øverst. På den måde kan redaktøren udvælge brugbare sætninger og/eller ændre de fremkomne forslag. Dette bliver gjort i håb om at det i stor udstrækning vil fremskynde det leksikografiske arbejde.

4. Leksikografisk arbejde

Efter den maskinelle bearbejdning som tidligere er beskrevet, er ordbogens redaktører i besiddelse af lange lister i alfabetisk orden som indeholder ordpar mellem islandsk og engelsk (sorteret efter det islandske ord). For at vælge de ord som skal med i ordbogen, er det næste skridt at gennemgå listerne og tynde ud i dem. I denne arbejdsgang bliver ca. 40 % af de engelske ord smidt ud, og tilbage står ca. 60 % af de engelske ækvivalentkandidater. Dette materiale bliver gjort klart til at blive indlæst i ordbogsbasen hvor de engelske ord indgår i ækvivalentfeltet.

handgerður adj		
1	HLUTAR	HAND-GERÐUR
2	BEYGING	→BEYGING
3	SKÝRING	búinn til eða unninn í höndunum
4	EN-jafn	handmade
5	DA-jafn	håndlavet

Figur 2: Fra den engelske ordbogsbase: lemmaet *handgerður* 'håndlavet' samt en engelsk ækvivalent som stammer fra ordlisten. Det danske ord under det engelske stammer fra ISLEX.

4.1. De islandsk-engelske ordpar

I bedste fald optræder de engelske ækvivalenter umiddelbart, oftest 1-4 ækvivalenter for hver betydning i en ordbogsartikel. I nogle tilfælde resulterer ordlisten dog i et stort antal ordpar. Det gælder bl.a. i de tilfælde hvor et ord har mange betydninger (polysemi, f.eks. *sterkur* 'stærk; robust' og *jörð* 'jord; landejendom').

Den oprindelige islandsk-engelske ordliste havde en mere omfattende rolle end den at danne det engelske målsprog i en tosproget ordbog. Til dette formål var listerne med ord alt for lange, og af den grund måtte de sorteres manuelt. Dette var enkelt når det drejede sig om rene fejloversættelser (som forekom nogle gange),

men i andre tilfælde var resultatet en vifte af ordpar hvor mange af kandidaterne var gode, jf. tabel 5.

ægilegur	
islandsk	*engelsk*
ægilegur	formidable fearsome atrocious gruesome terrible abominable horrific terrific horrid awful horrible appalling horrendous frightful dreadful

endrum og eins	
islandsk	*engelsk*
endrum og eins	at times every so often from time to time now and then occasionally once in a while sometimes

Tabel 5: To lemmaer med for mange engelske ækvivalenter som kræver en nøjagtig leksikografisk analyse og bearbejdning.

Som det fremgår af tabel 5, kan der blandt ordparrene opstå flere engelske ækvivalenter til det samme islandske ord, og disse ordlister kræver omhyggelig leksikografisk analyse. Ækvivalenterne er gerne (nær)synonymer i engelsk, og hvis de er flere end hvad der anses som passende i ordbogen, må der skæres ned på antallet. I andre tilfælde er der tale om islandske homonymer (f.eks. *kanna* subst. 'kande' og vb. 'undersøge', *ferja* subst. 'færge' og vb. 'færge') eller polysemer (f.eks. *slanga* 'slange; haveslange'), og så er det nødvendigt at indsætte ækvivalenterne i den rigtige ordbogsartikel eller i det rigtige nummererede afsnit i ordbogsartiklen.

En positiv ting var at når det drejede sig om tekniske begreber (termer), indeholdt listen tit et almindeligt engelsk ord samt ter-

men. Som eksempel kan nævnes de medicinske termer *rauðkorn* som fik ækvivalenterne *red blood cell* og *erythrocyte*; *þíamín* der fik ækvivalenterne *vitamin B1* og *thiamine*; *lærbein* der fik ækvivalenterne *thigh bone* og *femur*. Det redaktionelle arbejde går ud på at sætte en etiket på de tekniske begreber. Desuden skal der indsættes stilistiske markører på de engelske ækvivalenter hvor dette er nødvendigt, f.eks. *informal, vulgar, literary* og *dated.*

4.2. Britisk og amerikansk engelsk

Allerede i begyndelsen blev det besluttet at ordbogen skulle omfatte både britisk og amerikansk engelsk. Andre varianter af engelsk (f.eks. i Canada, Sydafrika og Australien) blev ikke taget i betragtning. Der er tit en betydelig forskel på britisk og amerikansk engelsk, f.eks. forskellig ortografi eller ordbrug. Eksempler på ord der har forskellig stavemåde er *colour* (Br), *color* (Am); *licence* (Br), *license* (Am); *jewellery* (Br), *jewelry* (Am). Eksempler på at der bruges forskellige ord er *pavement* (Br), *sidewalk* (Am); *anticlockwise* (Br), *counterclockwise* (Am); *(car)boot* (Br), *trunk* (Am).

I ordbogen bruges britisk som udgangspunkt mens amerikanske ord markeres specielt af redaktørerne.

4.3. Store verber

De store verber kan volde problemer eftersom et islandsk verbum tit får et stort antal engelske ækvivalenter. Det kommer ikke som en overraskelse da store verber foruden at være polyseme også forekommer i mange faste ordforbindelser i meget forskellige betydninger.[1]

1 Dette er sket i mange pivot-projekter selv om det drejer sig om ubeslægtede sprog, jf. Varga & Yokoyama 2009:869 (projekt som sammenkobler japansk og ungarsk).

Islandsk lemma	Engelsk ækvivalent
gefa	give yield grant give in donate accord confer allow quit sign up impart give up

Tabel 6: Verbet *gefa* 'give' samt nogle engelske ækvivalenter der blev genereret.

Tabel 6 viser det islandske verbum *gefa* 'give' og nogle ækvivalenter som er knyttet til det i listen med ordpar, men de engelske ord blev en del flere end vist her. Det er klart at de lange lister med forslag til ækvivalenter ikke er velegnede når man redigerer de store verber da brugen af dem i høj grad indgår i forskellige faste udtryk (f.eks. *gefa eftir, gefa upp, gefa út* osv.), noget som listerne med ækvivalenter ikke fanger så godt. For disse dele af ordbogen er det sandsynligvis bedre at anvende maskinoversættelser. Det må dog fremhæves at hvad angår "almindelige verber", dvs. langt de fleste verber på nær de ca. 60 største, er dette ikke noget problem. Det er især når verbet indgår i mange fraser at metoden med listerne ikke virker optimalt.

4.4. Ingen kandidater til ækvivalenter

Ved den manuelle redigering af ordbogsartiklerne må redaktørerne udvælge ækvivalenter fra de ordlister der er blevet indlæst i ordbogsbasen, og i nogle tilfælde må der også tilføjes ækvivalenter, da der somme tider behøves flere ord end dem som indlæses auto-

matisk. Desuden får nogle ord ingen ækvivalenter fra listerne, og de må derfor bearbejdes manuelt.

Ordbogen indeholder 54.000 opslagsord og 82 % af ordene fik én eller flere engelske ækvivalenter i den tekniske bearbejdning af de islandsk-engelske ordpar. Derimod fik 9.700 ord (18 %) ingen ækvivalenter. Dette kan skyldes flere ting:

1. Lemmaet er f.eks. en variant og indeholder kun en henvisning/et internt link til en anden ordbogsartikel (ca. 1.000 ord eller 2 % af ordforrådet).
2. Ordet anvendes kun i faste udtryk (ordforbindelser) og har derfor ikke en egentlig ækvivalent (ca. 900 ord eller 1,6 % af ordforrådet). Eksempel: *byssubrenndur* forekommer kun i det faste udtryk *hlaupa eins og byssubrenndur* 'løbe alt hvad remmer og tøj kan holde'.
3. Der eksisterer ikke en engelsk ækvivalent f.eks. for ordene *mannaferðir* 'menneskelig færden', *mófugl* 'mindre fugl af flere arter (fx hjejle) hvis biotop er græsbevoksede heder, moseområder e.l.'.
4. Ordet har en engelsk ækvivalent, men til trods for dette kom den ikke med på listen efter de anvendte metoder, formodentlig fordi de brugte ressourcer var mangelfulde.

Når et lemma ikke har en engelsk ækvivalent, må dets betydning forklares manuelt. Hvis det drejer sig om et lemma med kun et fast udtryk, anvendes maskinoversættelse for udtrykket, men i skrivende stund er man ikke kommet så langt. Faktum er at det stadig ikke er klart hvor mange af de 9.700 ord der skal bearbejdes manuelt, noget som vil vise sig på et senere stadie.

5. Evaluering af resultaterne

Selv om ordparrene alle er blevet evalueret som acceptable på første trin, er det ikke ensbetydende med at de alle kan anvendes i en ordbog. Nogle engelske ækvivalenter er meget specifikke, andre tilhører ældre sprog, er forældede eller sjældne. Desuden plejer man ikke at give mange ækvivalenter til samme betydning i en ordbog, og derfor må man kun vælge dem som passer bedst. Listerne med engelske ækvivalenter bearbejdes i to omgange. I første omgang gennemgår ordbogens redaktør listerne og skærer ned. På næste trin gennemgår medarbejderne forslagene til ækvivalenter og foretager det endelige valg af de engelske enheder der medtages i ordbogen.

Vi har undersøgt resultaterne af den første udvælgelse for tre bogstaver som er færdige, dvs. ord som begynder på bogstaverne L, M og N. Listen med forslag fra den oprindelige ordliste indeholdt 20.817 par for disse tre bogstaver. Af dem kom godt en tredjedel, dvs. 6.445 ord, ind på listen ved brug af én metode alene, mens de øvrige par kom med på listen ved brug af mere end én metode. Langt de fleste, eller over 85 % af ordparrene, blev genereret ved at bruge ordbøger via pivot-sprog, enten ved hjælp af denne metode alene eller kombineret med andre metoder.

		Maskin-oversættelse	Maskin-oversættelse via pivot-sprog	Ordbog via pivot-sprog	Korpus	**I alt**
Oprindelig liste	**Én kilde**	68	1.437	4.070	230	**6.445**
	Flere kilder	3.246	8.415	13.072	10.093	**14.372**
	I alt	3.314	9.852	17.782	10.323	**20.817**

Bearbejdet liste	**Én kilde**	68	670	2.277	80	**3.095**
	Flere kilder	2.931	6.644	8.009	5.770	**9.202**
	I alt	2.999	7.314	10.286	5.850	**12.297**
Beholdt ord i procenter	**Én kilde**	100 %	46,6 %	55,9 %	34,8 %	**48,0 %**
	Flere kilder	90,3 %	78,9 %	61,3 %	57,2 %	**64,0 %**
	I alt	90.5 %	74,2 %	57,8 %	56,7 %	**59,1 %**

Tabel 7: Antal par i den oprindelige liste med forslag og efter redaktionel gennemgang af bogstaverne L, M og N. Tabellen viser forslagenes antal og procentdel som opstår med kun én af de fire metoder, og siden med flere end én metode. Således er et forslag, som opstår både i oversættelsesmaskine og med hjælp af et korpus, talt på begge steder. Det samlede tal i den sidste kolonne er derfor ikke nødvendigvis summen af tallene i kolonnerne til venstre.

Efter en redaktionel gennemgang står 12.297 ordpar tilbage, eller knap 60 % af listen med forslag. Som forventet blev der skåret forholdsvis mere ned i de ordpar som blev genereret ved brug af én metode alene. På den anden side blev der skåret mindst ned i det ordforråd som blev genereret ved hjælp af maskinoversættelse eller maskinoversættelse mellem pivot-sprog. Forklaringen kan være at oversættelsesmaskinerne kun leverer én oversættelse af hvert ord, og når oversættelsen er korrekt, er det oftest den mest almindelige oversættelse. Derfor er sådanne ordpar også forholdsvis mere oplagte til at blive valgt i den type leksikografisk arbejde som beskrives her.

Under redaktørens gennemgang skulle der fjernes en del "støj". Hermed menes forkert stavede ord, ord med stort bogstav (foruden den korrekte skrivemåde), og forkerte ord efter redaktørens vurdering. Disse ord blev sigtet fra i første gennemgang. Desuden

må redaktørerne nogle gange tilføje de engelske ækvivalenter som de synes mangler i ordbogsartiklerne.

6. Konklusion

Projektet inddeles i nogle trin. Det første trin var at udarbejde lister med ordpar mellem islandsk og engelsk, som er en af de definerede opgaver i Handlingsplan for sprogteknologi for islandsk. For at producere listerne blev der anvendt fire forskellige metoder: brugen af pivot-sprog mellem islandsk og engelsk, to slags anvendelse af oversættelsesmaskiner, samt samkørsel af korpusser. Listerne med ordpar indeholdt langt flere islandske ord end der er opslagsord i ordbogen, og derfor blev de overflødige ord sigtet fra inden ordbogsredaktøren modtog ordlisterne. I alt blev der genereret engelske ækvivalenter for 82 % af ordforrådet.

Projektets anden fase er at filtrere listerne med ordpar for senere at bruge dem som stammen i ordbogen. Der skal skæres i ordlisterne med engelske ord indtil ca. 60 % står tilbage, som siden bliver indlæst i ordbogens database. Her kommer målsprogsredaktørerne ind i billedet, og de foretager den egentlige redaktion af ordbogsartiklerne. Dette arbejde foregår i skrivende stund.

En speciel fase i projektet er at anvende maskinoversættelse ved overførslen af sprogbrugseksempler og faste udtryk fra islandsk til engelsk. Denne fase er under udvikling, og de første forsøg giver forhåbninger om gode resultater.

Man kan gøre sig overvejelser over hvorvidt dette er en praktisk metode ved udarbejdelsen af en bilingval ordbog, og om metoden er tidsbesparende. Vi kan måle den tid der bliver anvendt og sammenligne med den tid det tog at udarbejde de bilingvale ordbøger i ISLEX, hvor målsprogsarbejdet udelukkende skete manuelt. Resultaterne viser os at denne nye metode er betragteligt hurtigere for målsprogsredaktørerne – men her tages de forudgå-

ende to gennemgange af de oprindelige islandsk-engelske ordpar dog ikke i betragtning. Projektets formål var imidlertid ikke alene at lave en islandsk-engelsk ordbog, men også at foretage en almen undersøgelse af maskinoversættelser mellem islandsk og engelsk.

Litteratur

Ordbøger

Apertium. A free, open-source machine translation platform. <apertium.org/> (marts 2022).

Dict.cc. <dict.cc/> (marts 2022).

ISLEX. Þórdís Úlfarsdóttir (hovedred.). Reykjavík: Árni Magnússon instituttet for islandske studier. <islex.dk/> (marts 2022).

Íslensk nútímamálsorðabók. Halldóra Jónsdóttir & Þórdís Úlfarsdóttir (red.). Reykjavík: Árni Magnússon instituttet for islandske studier. <islenskordabok.arnastofnun.is/> (marts 2022).

LEXIA. Þórdís Úlfarsdóttir (hovedred.). Reykjavík: Árni Magnússon instituttet for islandske studier. <lexia.arnastofnun.is/> (marts 2022).

Anden litteratur

Aker, Ahmet, Monica Paramita, Mārcis Pinnis & Robert Gaizauskus (2014): Bilingual dictionaries for all EU languages. I: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavík, Island, 483-489. <lrec-conf.org/proceedings/lrec2014/index.html>.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías,

Marek Strelec, Brian Thompson, William Waites, Dion Wiggins & Jaume Zaragoza (2020): ParaCrawl: Web-Scale Acquisition of Parallel Corpora. I: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4555-4567. <aclanthology.org/2020.acl-main.417>.

Barkarson, Starkaður & Steinþór Steingrímsson (2019): Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. I: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland, 140-145. <aclanthology.org/W19-6115>.

ELEXIS. European Lexicographic Infrastructure. <elex.is/> (marts 2022).

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli & Armand Joulin (2021): Beyond English-Centric Multilingual Machine Translation. I: *Journal of Machine Learning Research* 22(107), 1-48.

Gamallo, Pablo & José Campos (2010): Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora. I: *Computational Linguistics and Intelligent Text Processing, 11th International Conference*. Iasi, Romania, 473-483.

Nikulásdóttir, Anna Björk, Jón Guðnason & Steinþór Steingrímsson (2017): *Language Technology for Icelandic 2018–2022*. Project Plan. Reykjavík: Icelandic Ministry of Education, Science and Culture.

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong & Francisco Guzmán (2021): WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. I: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1351-1361. <aclanthology.org/2021.eacl-main.115>.

Símonarson, Haukur Barri, Vésteinn Snæbjarnarson & Vilhjálmur Þorsteinsson (2020): *En-Is Synthetic Parallel Corpus (20.09)*. CLARIN-IS <hdl.handle.net/ 20.500.12537/70>.

Steingrímsson, Steinþór, Pintu Lohar, Hrafn Loftsson & Andy Way (2021): Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. I: *Proceedings of the 14th Workshop on Building and Using Comparable Corpora*. <aclanthology.org/2021.bucc-1.3/>.

Tiedemann, Jörg & Santhosh Thottingal (2020): OPUS-MT – Building open translation services for the World. I: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal, 479-480. <aclanthology.org/2020.eamt-1.61>.

Úlfarsdóttir, Þórdís (2013): ISLEX – norræn margmála orðabók. I: *Orð og tunga* 15, 41-71.

Úlfarsdóttir, Þórdís (2014). ISLEX – A Multilingual Web Dictionary. I: *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavík, 2820–2825. <aclanthology.org/L14-1>.

Varga, István & Shoichi Yokoyama (2009): Bilingual dictionary generation for low-resourced language pairs. I: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 862-870. <aclanthology.org/D09-1090>.

Þórdís Úlfarsdóttir
hovedredaktør
Árni Magnússon-instituttet for islandske studier
Laugavegur 13
IS-101 Reykjavík
thordis.ulfarsdottir@arnastofnun.is

Steinþór Steingrímsson
sprogteknolog
Árni Magnússon-instituttet for islandske studier
Laugavegur 13
IS-101 Reykjavík
steinthor.steingrimsson@arnastofnun.is