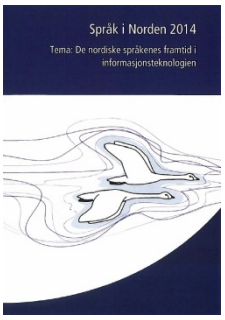
Sprog i Norden

Titel: Det islandske ordklasseopmærkede korpus (MÍM)

Forfatter: Sigrún Helgadóttir

Kilde: Sprog i Norden, 2014, s. 83-94

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) 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.

Det islandske ordklasseopmærkede korpus (MÍM)

Sigrún Helgadóttir

I artiklen gives der en kort redegørelse for projektet "Det islandske ordklasseopmærkede korpus" som kaldes "MÍM" på islandsk. Der beskrives kort hvorfor projektet blev påbegyndt. Endvidere beskrives hvad et ordklasseopmærket korpus er for noget og hvordan det kan bruges. Desuden følger der en redegørelse for hvordan korpusset blev oprettet og dets tilgængelighed. Til sidst beskrives to andre korpusser som er tilgængelige på webpladsen http://mim.arnastofnun.is/.

Indledning

I 1998 nedsatte undervisnings-, forsknings-, og kulturminister Björn Bjarnason et udvalg for at undersøge statussen af sprogteknologi i Island og fremsætte forslag om hvordan sprogteknologien kunne styrkes (Rögnvaldur Ólafsson m.fl. 1999). Udvalget afleverede sin rapport i april 1999, og i 2000 blev undervisningsministeriets sprogteknologiske projekt lanceret. En af de sprogresurser som udvalget mente manglede var et ordklasseopmærket tekstkorpus. *MÍM*-projektet var et af de seneste projekter, som blev finansieret af det sprogteknologiske projekt, og det skulle aflevere et sådant korpus. Projektet blev startet i 2004 på *Orðabók Háskólans* (Leksikografisk institut ved Islands universitet) og færdiggjort på *Stofnun Árna Magnússonar í íslenskum fræðum* (Árni Magnússon-instituttet for islandske studier) i 2013. Artiklen giver en kort redegørelse for projektet. Ordklasseopmærket korpus beskrives, og derefter følger en redegørelse for *MÍM*-korpusset, hvordan tekster blev samlet, og hvordan man behandlede ophavsret. Endvidere beskrives hvordan teksten blev renset og opmærket med oplysninger om ordklasse, bøjning og lemmaer. Tilgængelighed og brug beskrives og to andre korpusser som også er tilgængelige til søgning på samme måde som *MÍM*-korpusset. Til sidst gives et eksempel på søgning í *MÍM*.

Ordklasseopmærket korpus

Et ordklasseopmærket korpus (e. *tagged corpus*) er en samling af elektroniske eller digitale tekster. Hver enkelt tekst er forsynet med oplysninger om teksten som kaldes *metadata*, for eksempel titel, udgivelsesår, genre og forfatterens navn (eventuelt køn og fødselsår) for udgivne tekster. Hvert ord er opmærket med oplysninger om ordklasse, bøjning og lemma, og korpusset er gemt i standardiseret format, sædvanligvis xml-format. For at få et balanceret korpus er det en fordel, at teksterne kommer fra forskellige kilder, som skal give indtryk af, hvordan et sprog bliver brugt i en bestemt periode.

Fra korpusset kan man finde information om frekvens af ordklasser, ord og bøjningsformer, om ordforbindelser, syntaks og semantik osv. Korpusser er også nyttige, når man laver for eksempel ordbøger, programmer til stave- og grammatikkontol, maskinoversættelse, talegenkendelse og talesyntese, til støtte for handicappede (blinde, døve og hørehæmmede, bevægelseshæmmede og ordblinde) og i sprogundervisning.

MÍM korpusprojektet

Formålet med korpusprojektet (Sigrún Helgadóttir m.fl. 2012) var at samle tekster, skrevet af mennesker som har islandsk som modersmål, fra en række forskellige kilder med i alt 25 millioner ord fra perioden 2000–2010. Kun tekster som var elektronisk tilgængelige, blev samlet, og man skulle sikre licens til brug af teksterne i korpusset fra indehavere af ophavsret. Det var tænkt som yderst vigtigt at sikre licens for ophavsretsbeskyttet materiale. Alle tekster skulle opmærkes automatisk med oplysninger om ordklasse og bøjning og med lemma.

I tabel 1 vises hvordan teksterne i korpusset fordeles mellem de forskellige kilder af tekster. Størstedelen af teksterne kommer fra trykte bøger og fra aviser, både trykte og webaviser. Offentlige tekster udgør en stor del af korpusset og dernæst kommer tekster fra tidsskrifter, trykte og elektroniske. Man fandt det også vigtigt at få tekster fra blogs. Vi fik tekster fra almene bloggere, teologer og politikere. Bloggene er anonyme. I gruppen "Usorteret" findes der nogle tekster fra e-post lister, som også ser ud til at være uformel tekst ligesom bloggen. Det var meget vigtigt at få hjælp fra redaktøren til universitetets videnskabsweb[1]. Han sikrede licens fra alle som skriver tekster for webben og afleverede teksten til kor-

1 http://visindavefur.is/

pusset. Videnskabswebben indeholder tekster om alt muligt fra socialvidenskab til astronomi.

Genre i MÍM	Antal tekster (filer)	Antal ord	%
Trykte bøger	168	5.972.893	23,89
Aviser (trykte og elektroniske)	12.725	5.779.509	23,12
Offentlige tekster (rapporter, domme, forslag, love, drøftelser fra Alþingi)	1.246	3.513.990	14,06
Tidsskrifter (trykte og elektroniske)	311	2.501.222	10,00
Blogs	8.998	1.976.706	7,91
Artikler fra Universitetets videnskabsweb (Vísindavefurinn)	4.949	1.838.909	7,36
Tekster fra websites for virksomheder, organisationer og institutter	106	1.337.764	5,35
Tekster til oplæsning (blandt andet fra radio og tv)	1.196	694.506	2,78
Stile og skriftlige opgaver fra gymnasieelever og studenter	51	666.042	2,66
Talesprog	4	504.318	2,02
Usorteret	46	214.663	0,86
Totalt	29.800	25.000.522	100,00

Tabel 1. Tekster i MÍM efter genre.

I alt har man 58 % af tekstmaterialet i korpusset fra trykte kilder, 40 % fra webben og talesprog udgør 2 %. Ophavsretsbeskyttet tekst udgør omtrent 85,9 % af korpusteksterne.

Projektet havde ikke resurser til at samle talesprog. Talesprog blev skaffet fra fire forskellige talesprogsprojekter (Höskuldur Þráinsson m.fl. 2007), i alt 54 timer af transskriberet tale. Disse projekter omfatter monologer (taler fra Alþingi, delprojekt af ScanDiaSyn), interviews (MIN - Moderne importord i sprogene i Norden) og spontane samtaler mellem voksne mennesker, mænd og kvinder (ÍSTAL og Hvordan taler unge islændinge i begyndelsen af det 21. århundrede?). Alle transskriberede tekster er blevet opmærkede og inkluderet i korpusset, men er anonyme. Man planlægger at gøre talesproget tilgængeligt til søgning med lydfiler snart, men kun talerne fra Alþingi bliver tilgængelige for almenheden.

Ophavsret

I korpusprojektet arbejdede man med to juridiske dokumenter. Alle indehavere af ophavsretsbeskyttet udgiven tekst underskrev en deklaration. Indehaverne fik information om projektet og en kopi af brugerlicensen. De fleste indehavere har givet licens til brug af alle deres tekster, men enkelte kun de tekster som blev inkluderet i korpusset. Forfattere af tekst som ikke var udgivet, fik en meget enklere deklaration til underskrift. Deklarationen foreskriver blandt andet at alle tekster skulle være tilgængelige uden betaling, kun 80 % af udgivne tekster inkluderes i korpusset, og brugere må acceptere brugerlicensen.

Brugerlicensen foreskriver at brugeren kan bruge sine resultater (dvs. det som han lærer fra korpusset) frit. Man kan for eksempel forestille sig at man laver n-grams fra korpusset og bruger disse til at lave skrivehjælp til ordblinde. Men teksterne må ikke kopieres eller videregives til andre, undtagen det, som er omfattet af citatretten.

I begyndelsen af projektet sikrede man samarbejde fra Den islandske forfatterforening[2] (Rithöfundasamband Íslands), Islandsk forfatterforening for faglitteratur og undervisning[3] (Hagþenkir) og Den islandske forlæggerforening[4] (Félag íslenskra bókaútgefenda). Alle tre foreninger anbefalede deres medlemmer at de skulle samarbejde med Árni Magnússon-instuttet om korpusprojektet. Det var yderst vigtigt at sikre at udgiverne var positive overfor projektet siden de bevarer de elektroniske kopier af udgivne bøger. Projektet fik et meget fint samarbejde med nogle af de største udgivere i Island.

Forberedelse af tekster for korpusset

Tekst blev skaffet i forskellige formater, som pdf, xml eller Word filer, som tekst fra databaser, webtekst og så videre. Teksten måtte uddrages af formatet. Man fjernede også fremmedsprogede og oldnordiske citater, fodnoter, indholdsfortegnelser, indekser, digte, tabeller, billeder og så videre samt bindestreg. Alle tekster er tilgængelige i UTF-8 tegnkodningstabel.

Den sædvanlige arbejdsgang ved korpustekst, når teksten er blevet renset, er at segmentere i sætninger og tokenisere, dvs. fordele teksten i tokens eller ord. Det blev gjort med *IceNLP* softwaresystemet (Hrafn

2 http://rsi.is/
3 http://hagthenkir.is/
4 http://www.bokatidindi.is/index.php/forsiea

Loftsson og Eiríkur Rögnvaldsson 2007). Derefter blev teksten ordklasseopmærket og lemmatiseret.

Opmærkningen giver hvert ord en bogstavsstreng, et tag, som viser ordklasse og bøjning og lemma. Det første bogstav i taggen angiver ordklassen, og de andre (op til 5) angiver for eksempel køn, tal og kasus for substantiver og person, tal og tempus for verber. Lemma er ordets grundform eller opslagsform og er for eksempel nominativ, ental for substantiver og infinitiv for verber. Som eksempel kan man analysere sætningen *ég sagði* (jeg sagde). Lemma til ordet *ég* er *ég* og taggen bliver **fp1en**, hvor **f** står for pronomen, **p** står for personligt pronomen, **1** står for første person, **e** står for ental og **n** står for nominativ. Lemma til ordet *sagði* er *segja* og taggen bliver **sfg1eþ** hvor **s** står for verb, **f** står for indikativ, **g** står for aktiv, **1** står for første person, **e** står for ental og **þ** står for datid.

Analysen bygger på det tidligere arbejde med *Íslensk orðtíðnibók* [Den islandske frekvensordbog] (Jörgen Pind m.fl. 1991) og tagsættet indeholder omtrent 700 tags.

Ordklasseopmærkning sker med fire ordklassetaggere og et system til at vælge imellem de fire tags. Til sidst bruges en lemmatiserer for at finde lemmaet for et ord. Hele processen er pakket ind i et system som vi kalder *CorpusTagger* (Hrafn Loftsson m.fl. 2010).

Alle tekster i korpusset er forsynet med metadata. For udgivne tekster består metadata af bibliografiske oplysninger som titel, udgivelsesår, genre og forfatterens navn (eventuelt køn og fødselsår). For andre tekster registreres metadata, som identificerer teksten. For talesprog er oplysninger om sessionen og talerne registreret. De fleste metadata er blevet registreret manuelt, for enkelte filer fra avisen *Morgunblaðið,* for eksempel, og taler fra Alþingi er metadata oprettet automatiskt. Når man søger i teksterne via søgegrænsefladen, bliver metadata vist og alle xml-filer, som kan downloades, indeholder metadata.

Tilgængelighed og brug

Korpusset er tilgængeligt på to forskellige måder. Man kan søge i teksterne og bruge taggene for at lave en mere nøjagtig søgning. Man kan finde eksempler på hvordan et ord eller en vending bliver brugt i islandsk. Resultatet vises i en konkordans, som er en række linjer, hvor søgeudtrykket optræder i den umiddelbare kontekst. For at få et overblik over konkordansen kan man evt. sortere den. Kilden, hvor teksten i hver enkelt linje i

konkordansen stammer fra, bliver også vist. Søgesiden for *MÍM*-korpusset er http://mim.arnastofnun.is/.

Søgegrænsefladen bygger på Glossa[5] fra Universitetet i Oslo og Glossa bruger "corpus search engine", IMS Corpus Workbench (CWB)[6] fra Universitetet i Stuttgart.

Brugere som vil undersøge og se eksempler på sproglige fænomener, sådan som de optræder i naturligt forekommende islandske tekster, bruger søgesiden. Det kan være nyttigt både for dem, der beskæftiger sig professionelt med sprog (fx journalister, lærere og sprogforskere), og for dem, der bare synes sprog er interessant og sjovt. *MÍM* er allerede blevet brugt i undervisningen i Háskóli Íslands. Vi har planer om at præsentere, hvordan korpusset kan bruges i undervisning. Der er skrevet en artikel om *MÍM* og dets brug i undervisning i et tidsskrift udgivet af sammenslutningen af islandske modersmålslærere (Sigrún Helgadóttir 2013).

I forbindelse med META-NORD projektet har den islandske META-NORD gruppe etableret webpladsen www.málföng.is, hvor forskellige sprogresurser er tilgængelige. *MÍM*-korpusset er tilgængeligt for download fra denne webplads i en pakke som består af 29.800 filer i TEI-konform[7] xml-format. Brugere, som vil hente korpusset, må acceptere en særlig brugerlicens. Denne pakke er hovedsagelig tænkt for dem, som laver sprogteknologiske værktøjer. Korpusset er allerede blevet brugt til at komplettere BÍN (database for moderne islandske fleksioner) (Bjarnadóttir 2012), for at lave en database med semantiske relationer for at udarbejde et program til stavekontol og for at udforme "word prediction" til ordblinde.

Andre korpusser

På *MÍM* søgesiden, http://mim.arnastofnun.is/, findes der to andre korpusser. Det første er Korpus for den islandske frekvensordbog (IFD Korpusset). Frekvensordbogen blev udgivet 1991 (Jörgen Pind m.fl. 1991) og er lavet af en samling af tekster udgivet 1980–1989, hovedsagelig litterære tekster. Teksterne giver et billede af sproget, som det blev skrevet i udgivne tekster i perioden 1980–1989. Tekstsamlingen indeholder omtrent 500.000 ord fra 100 tekster, opmærkede med oplysninger om ordklasse,

5 http://www.hf.uio.no/tekstlab/glossa.html
6 http://cwb.sourceforge.net/
7 http://www.tei-c.org/index.xml

bøjning og lemmaer plus metadata om hver tekst. Korpusset er velegnet til undervisning i morfologi, siden taggene blev korrigeret manuelt. Korpusset er tilgængeligt på webpladsen http://www.málföng.is/ til søgning og download med særlig brugerlicens (omtrent identisk brugerlicensen for *MÍM*). IFD korpusset er også blevet brugt til træning af "data-driven" taggere og til udvikling af regelbaserede taggere. Korpusset er tilgængeligt til dette formål på webpladsen http://málföng.is/.

Det andet korpus indeholder tekster fra de islandske sagaer og kaldes *Sagakorpusset* (Eiríkur Rögnvaldsson og Sigrún Helgadóttir 2011). Korpusset indeholder 44 digitale tekster fra sagaer, dvs. fra 41 islandske sagaer samt værkene Sturlunga, Heimskringla og Landnámabók. Korpusset indeholder 1.659.385 ord, opmærkede med oplysninger om ordklasse, bøjning og lemma plus metadata om hver tekst. Teksterne er blevet normaliseret til moderne retskrivning og nogle bøjningsendelser blev ændret. Korpusset er tilgængeligt på webpladsen http://málföng.is/ til søgning og download med brugerlicens CC BY 3.0[8] (Creative Commons Atribution 3.0 Unported). Sagakorpusset kan bruges blandt andet til at undersøge brug af ord og konstruktioner i sagaerne.

Eksempel på søgning

Man kan søge efter enkelte ord eller ordforbindelser í *MÍM* og finde eksempler på sprogbrug i formelle eller uformelle tekster, som blev skrevet i perioden 2000–2010. Man kan bruge taggene for en mere præcis søgning. Det kan for eksempel være nyttigt at finde, hvilket adjektiv bruges med forskellige substantiver.

Vi kan for eksempel søge efter hvilke adjektiver bruges med ordet "kjóll" (kjole). I så fald kan man definere søgning efter alle bøjningsformer af ordet "kjóll" med et adjektiv, som står foran substantivet.

Figur 1 viser hvordan søgningen er defineret. Det første søgeord bør være et adjektiv (lýsingarorð) og det andet søgeord en bøjningsform af ordet "kjóll". Det realiseres ved at specificere "nefnimynd" (lemma)[9].

Søgningen giver 198 konkordanslinjer, som alle stammer fra udgivne bøger. Figurer 2 og 3 viser nogle konkordanslinjer, som er resultat af søgningen specificeret i Figur 1. Med en mouse-overfunktion kan man se or-

8 http://creativecommons.org/licenses/by/3.0/

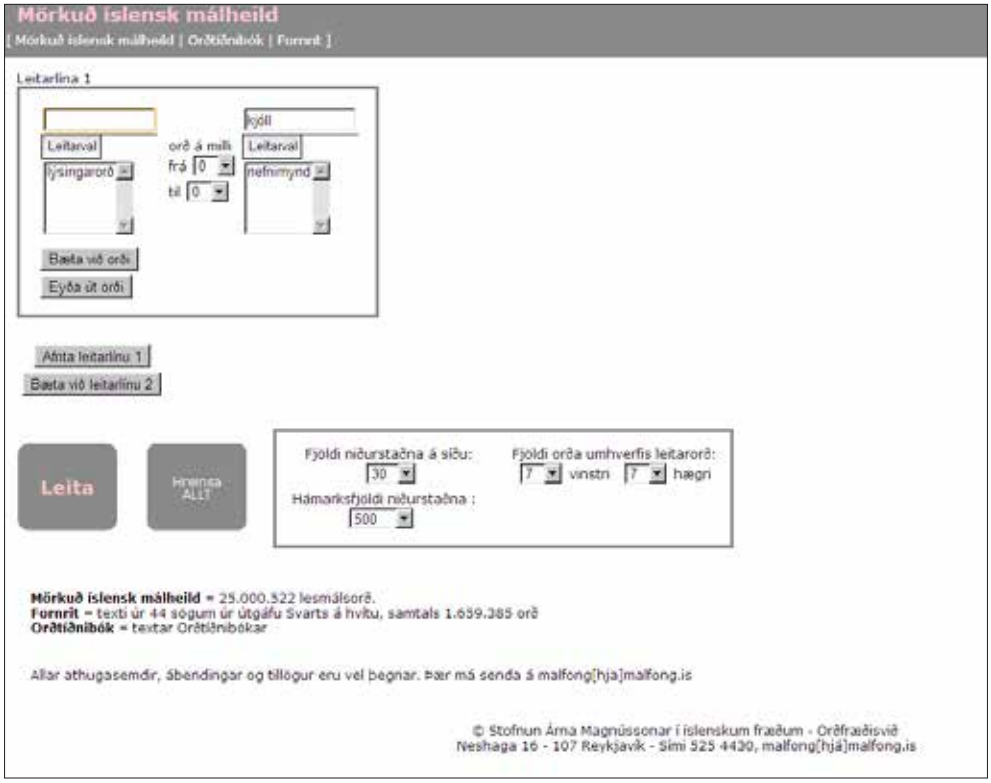
9 Søgegrænsefladen er på islandsk men man har planer om en engelsk grænseflade i den nærmeste fremtid.

denes analyse. På Figur 2 kan man se en analyse for ordet "hvítum" (maskulin, dativ af ordet "hvítur", dansk hvid).

På venstre side af hver konkordanslinje er en tegnstreng for kilden af tekstudsnittet. Med en mouse-overfunktion på tegnstrengen kan man se oplysninger om kilden. På Figur 3 kan man se at teksten „...klæða sig . Fyrst fóru þær í **ljósbláan kjól** eða kyrtil sem náði þeim niður á..." stammer fra bogen *Valkyrjan,* som er skrevet af forfatteren Elías Snæland Jónsson, udgivet af Vaka-Helgafell i 2003.

Konklusion

Det er allerede klart at *MÍM*-korpusset er meget vigtigt for dem, som beskæftiger sig med lingvistiske undersøgelser. Det er også vigtigt for dem, som laver sprogteknologiske værktøjer, da det er det eneste store islandske ordklasseopmærkede korpus. Det var uhyre vigtigt at sikre licens til brug af ophavsretsbeskyttet tekst således at forskere, eventuelt med nogle restriktioner, kan bruge korpusset til forskellige formål.



Figur 1. Søgning for adjectiver med ordet kjóll.

Figur 2. Nogle linjer fra resultat af søgningen specificeret i Figur 1.

Figur 3. Nogle linjer fra resultat af søgningen specificeret i Figur 1

Finansiering og tak

Korpusprojektet var hovedsagelig finansieret af undervisningsministeriets sprogteknologiske projekt, som lanceredes i 2000. Men det er også blevet finansieret fra andre kilder. Man kan nævne forskningsprojektet "Tilbrigði í setningagerð"[10], som har leveret talesprogsmateriale til *MÍM*, Nordisk Ministerråd (Nordisk Netordbog), Rannís[11] – Det islanske forskningsfond (Viable Language Technology beyond English – Icelandic as a test case[12]), Islandske studenters innovationsfond[13] (nogle stipendier), Universitetets forskningsfond (nogle stipendier) og META-NORD[14]. Forfatteren vil også gerne takke Eiríkur Rögnvaldsson og Halldóra Jónsdóttir for at læse manuskriptet med henblik på indhold og sprog.

Emneord: korpus, opmærkning, taggers, sprogteknologi, islandsk

Sigrún Helgadóttir er statistiker og har arbejdet med sprogteknologi på Árni Magnússon-instituttet for islandske studier i mange år.

Summary

The article describes the development of a morphosyntactically tagged corpus of Icelandic, the MÍM corpus. The corpus consists of about 25 million tokens of contemporary Icelandic texts from the years 2000–2010, collected from varied sources during the years 2006–2010. The corpus is intended for use in Language Technology projects and for linguistic re-

10 http://malvis.hi.is/tilbrigdi_i_setningagerd
11 http://rannis.is/
12 http://iceblark.wordpress.com/
13 http://rannis.is/funding/icelandic-student-innovation-fund/
14 http://www.meta-nord.eu/

search. Text collection and permission clearance is described. Text cleaning and annotation phases are also described. The corpus has been available since spring 2013 for search through a web interface and for download in TEI-conformant XML format. Two other corpora available from the same web pages are also described.

Litteratur

Eiríkur Rögnvaldsson og Sigrún Helgadóttir, 2011: Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change. I: Caroline Sporleder, Antal P.J. van den Bosch og Kalliopi A. Zervanou (red.): *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. s. 63–76. Berlín: Springer.

Hrafn Loftsson og Eiríkur Rögnvaldsson, 2007: IceNLP: A Natural Language Processing Toolkit for Icelandic. I: *Proceedings of Interspeech 2007, Special Session: "Speech and language technology for less-resourced languages"*. Antwerpen.

Hrafn Loftsson, J.H. Yngvason, S. Helgadóttir, E. Rögnvaldsson, 2010: Developing a PoS-tagged corpus using existing tools. I: *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation (LREC 2010).* Valetta, Malta.

Hrafn Loftsson, S. Helgadóttir, E. Rögnvaldsson, 2011: Using a morphological database to increase the accuracy in PoS tagging. I: *Proceedings of Recent Advances in Natural Language Processing, RANLP*. Hissar, Bulgaria.

Höskuldur Thráinsson, Á. Angantýsson, Á. Svavarsdóttir, T. Eythórsson, J. G. Jónsson, 2007. The Icelandic (Pilot) Project in ScanDiaSyn. I: *Nordlyd*, 34(1), s. 87–124.

Jörgen Pind, F.Magnússon, S. Briem, 1991: *Íslensk orðtíðnibók* [Den islandske frekvensordbog]. Reykjavik: Leksikografisk institut ved Islands Universitet.

Kristín Bjarnadóttir, 2012. The Database of Modern Icelandic Inflection. I: *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT2012*. Istanbul, Tyrkiet.

Rögnvaldur Ólafsson, E. Rögnvaldsson, Þ. Sigurðsson, 1999: *Tungutækni* [Sprogteknologi]. Skýrsla starfshóps [Komité rapport], Reykjavik:

Menntamálaráðuneytið [Undervisnings-, forsknings- og kulturministeriet].
Sigrún Helgadóttir, Á. Svavarsdóttir, E. Rögnvaldsson, K. Bjarnadóttir, H. Loftsson, 2012: The Tagged Icelandic Corpus (MÍM). I: *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages -SaLTMiL 8 – AfLaT2012*. Istanbul, Tyrkiet.
Sigrún Helgadóttir, 2013: Mörkuð íslensk málheild. *Skíma,* 36(1), s. 26–27. Reykjavík.