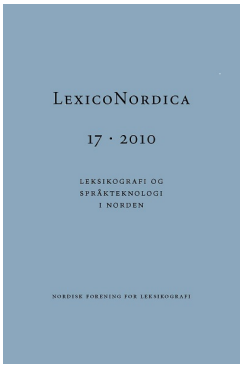
LexicoNordica

Titel: Sprogteknologiske ressourcer for islandsk leksikografi

Forfatter: Eiríkúr Rögnvaldsson

Kilde: LexicoNordica 17, 2010, s.181-195

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



Søgbarhed

Artiklerne i de ældre LexicoNordica (1-16) 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.

Sprogteknologiske ressourcer for islandsk leksikografi

Eiríkur Rögnvaldsson

Ten years ago, the Icelandic government launched a special Language Technology Program with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in the creation and development of several important resources and tools that have had profound influence on Icelandic language technology, and are also valuable for Icelandic lexicography and linguistic research in general. The present paper describes briefly some of the most important of these products, such as a morphological database (260,000 lemmas), a 25 million word balanced PoS tagged corpus, a lemmatiser, a rule-based tagger, and a shallow parser. Finally, it is pointed out that all the tools that the Icelandic Language Technology Community has developed in the past few years have been made Open Source, and the importance of adopting Open Source Policy for small language communities is emphasized.

1. Indledning

Denne artikel handler om de vigtigste sprogteknologiske ressourcer som eksisterer for islandsk.[1] Selv om disse ressourcer oprindelig blev lavet for at benyttes i sprogteknologiske værktøjer, så kan de fleste af dem også være nyttige for leksikografien. Der begyndes med at opridse baggrunden for at de fleste af disse ressourcer blev lavet – det islandske sprogteknologiprojekt som begyndte i 2001. Derefter beskrives de vigtigste af disse ressourcer kort – en morfo-

1 Mange tak til redaktørerne for *LexicoNordica* for kommentarer og rettelser.

logisk database, et balanceret korpus og sprogteknologiske værktøjer som en lemmatiser, en morfologisk tagger og en syntaktisk parser. Til slut berøres spørgsmålet om open source policy som forfatteren anser for at være meget vigtig for et lille sprogsamfund som det islandske.

2. Islandsk sprogteknologi omkring århundredskiftet

Man kan godt sige at islandsk sprogteknologi simpelthen ikke eksisterede for ti år siden. Der fandtes ganske vist et godt stavekontrolsystem og en talesyntese, omend den var relativt primitiv. Men det var alt. Der eksisterede ingen programmer eller enkelte kurser i sprogteknologi eller datalingvistik ved islandske universiteter eller højskoler, og der fandtes ikke nogen som bedrev akademisk forskning inden for dette felt. Der var heller ikke nogen private firmaer som arbejdede med sprogteknologi.

Dette har nu ændret sig. I efteråret 1998 oprettede ministeren for undervisning og forskning et udvalg som skulle undersøge situationen for islandsk sprogteknologi og komme med forslag til hvordan man kunne styrke islandsk sprogteknologi. Der var tre medlemmer i udvalget: Rögnvaldur Ólafsson, docent i fysik, Eiríkur Rögnvaldsson, professor i islandsk sprog, og Þorgeir Sigurðsson, ingeniør og lingvist.

Udvalget leverede sin rapport til ministeren i april 1999 (Ólafsson m.fl. 1999), og to år senere, i år 2001, oprettede regeringen et specielt sprogteknologiprojekt (Arnalds 2004, Ólafsson 2004). Projektets hovedmål var at give institutioner og firmaer støtte til at opbygge grundressourcer for islandsk sprogteknologi. Dette initiativ har ført til oprettelsen af forskellige projekter som har haft stor indflydelse på dette område.

Sprogteknologiudvalget foreslog fire typer af aktiviteter til at styrke islandsk sprogteknologi (Ólafsson m.fl. 1999):

- Sproglige ressourcer skulle udvikles og opbygges til anvendelse for firmaer som ville udvikle sprogteknologiske værktøjer og andre produkter.
- Praktisk forskning inden for sprogteknologi skulle støttes.
- Firmaer skulle støttes til at udvikle sprogteknologiske produkter.
- Universitetsprogrammer og kurser i sprogteknologi skulle oprettes.

Alle disse aktiviteter er blevet gennemført, i det mindste til en vis grad (Arnalds 2004, Ólafsson 2004, Rögnvaldsson 2005). Vigtige ressourcer er blevet opbygget, sprogteknologisk forskning er blevet igangsat, støtte til at udvikle sprogteknologiske værktøjer er blevet givet til firmaer, og en tværfaglig mastergrad i sprogteknologi er blevet oprettet i samarbejde mellem Islands Universitet og Reykjavik University. De vigtigste ressourcer som blev udviklet inden for sprogteknologiprojektet, er følgende:

- En morfologisk database med omkring 260.000 ord.
- Et balanceret og tagget korpus med 25 millioner ord.
- En statistisk PoS-tagger.
- Islandsk talesyntese.
- Islandsk talegenkender.
- Et forbedret stavekontrolsystem.

I det følgende gives der en kort beskrivelse af de af disse ressourcer som kan være til nytte for islandsk leksikografi (se også Rögnvaldsson 2008, Rögnvaldsson m.fl. 2009).

3. Sprogteknologiske ressourcer

3.1. Morfologisk database

En af de vigtigste ressourcer som blev sat i gang under regeringens sprogteknologiprojekt, er en morfologisk database for islandsk sprog, Beygingarlýsing íslensks nútímamáls (bøjningsbeskrivelse af islandsk nutidssprog, BÍN; Bjarnadóttir 2005) som der arbejdes på ved Leksikografisk institut ved Islands Universitet, der nu er blevet en del af Árni Magnússon-instituttet for islandske studier. Projektet blev påbegyndt i 2002 og var i begyndelsen finansieret af den islandske sprogteknologifond. Den første fase af projektet blev afsluttet i 2004, men arbejdet fortsættes – nu finansieret af instituttet. Kristín Bjarnadóttir har været projektleder fra begyndelsen. Databasen indeholder nu paradigmer for næsten 260.000 ord, mens der er mere end 5,6 millioner ordformer.

Databasen blev oprindelig oprettet med to formål. Først og fremmest skulle den kunne bruges i forskellige sprogteknologiske produkter og værktøjer. Det var faktisk hovedgrunden til at den blev lavet, for uden finansiering fra sprogteknologifonden havde det været umuligt at sætte projektet i gang. Men databasen har også været til stor nytte for almene brugere efter at den er kommet på internettet og alle kan slå bøjningen af islandske ord op i den. Denne mulighed er faktisk blevet meget populær, og både skoleelever og andre bruger databasen meget.

Databasen er allerede blevet benyttet i nogle praktiske produkter, særlig søgeprogrammer. Når man bruger disse programmer til søgning efter islandske ord, behøver man ikke indtaste alle mulige bøjningsformer fordi programmerne har adgang til databasen og derfor automatisk kan søge efter alle former – også de uregelmæssige. Databasen bruges fx i telefonbogen på internettet (http://ja.is) hvilket betyder at man kan taste hvilken som helst bøjnings-

form af et givent navn ind – telefonbogen vil vise navnene i den almindelige opslagsform (dvs. nominativ). Databasen bruges også i læremateriale*t Icelandic Online* på internettet (http://icelandic.hi.is). Desuden bruges den i forskellige projekter der vedrører tagging og lemmatisering af islandske tekster.

Databasen var oprindelig tekstbaseret og lagret som XML-filer, men nu er den lige blevet omstruktureret og filerne lagt ind i en MySQL-database. Kristín Bjarnadóttir har været ansvarlig for den lingvistiske del af omstruktureringen, mens Hjálmar Gíslason har været ansvarlig for al programmering. Formålet med omstruktureringen er på den ene side at gøre det nemmere at føje nye ord til databasen og rette fejl i den og på den anden side at gøre søgning i databasen hurtigere og mere effektiv.

3.2. Balanceret tagget korpus

En anden ressource som er blevet udviklet med støtte fra sprogteknologiprojektet, er et balanceret PoS-tagget korpus på omkring 25 millioner ord (Helgadóttir 2005, 2009; Loftsson m.fl. 2010). Dette korpus er også blevet udviklet på Leksikografisk institut, og projektlederen er statistikeren Sigrún Helgadóttir. Projektet blev påbegyndt i 2004 og skal være færdigt i 2011. Indsamlingsfasen var meget sværere og langsommere end man havde forventet. Grunden hertil var hovedsagelig problemer angående ophavsret. Det tog vældig lang tid at få lov hos alle ophavsretsindehavere til at bruge deres tekster. Det viste sig også at bogforlagene var vældig skeptiske over for projektet og ikke særlig villige til at udlevere de tekster som de var i besiddelse af. Men til sidst lykkedes det at forhandle en aftale med forlagene på plads, og indsamlingen af tekster er nu færdig.

Målet var at få fat i så mange teksttyper som muligt, og der findes i alt omkring 20 forskellige teksttyper i korpusset. Størstedelen af materialet kommer fra de følgende teksttyper: Avistekster

(28%), trykte bøger (romaner osv.) (22%), blog (7,5%), forskellige tidsskrifter (8,5%), tekst fra den islandske "Videnskabsweb" (http://visindavefur.hi.is) (7,5%), webtekster fra institutter, firmaer, foreninger etc. (6%), love og andre tekster fra Altinget (3%) samt talesprog (2%) (Helgadóttir 2009; se også Loftsson m.fl. 2010). Størstedelen af talesprogsmaterialet er samtaler fra projektet Ístal – islandsk talesprogsbank – og blev indsamlet i år 2000. Alle teksterne er fra det 21. århundrede, dvs. årene 2000–2009.

I løbet af 2010 bliver teksterne PoS-tagget og lagret i TEI-kompatibelt XML-format (http://www.tei-c.org/release/doc/tei-p4-doc/html/). Nu arbejdes der på en brugergrænseflade for korpusset. Det skal være søgbart gennem internettet, og man skal kunne søge efter enkelte ord og ordforbindelser, og også efter tags – fx alle substantiver i femininum, singularis, dativ. Projektet har fået adgang til søgesystemet Glossa hos Tekstlaboratoriet i Oslo og er i gang med at tilpasse det til korpusset. Dette arbejde vil forhåbentlig blive færdigt i begyndelsen af 2011, men en præliminær version findes allerede på webben (http://mim.hi.is). Så vil alle kunne søge i korpusset, men af ophavsretlige grunde vil man kun se et brudstykke af hver tekst ad gangen. De som har brug for hele korpusset, enten til lingvistisk forskning eller for at udvikle sprogteknologiske værktøjer, vil kunne få hele korpusset udleveret, men må i givet fald underskrive en kontrakt. Dette er en model der kendes andre steder fra, og ligner fx den som bruges til den danske orddatabase STO.

Dette korpus vil forhåbentlig blive af stor nytte for alle som arbejder med islandsk sprog og islandsk leksikografi. Det vil blive muligt at få meget bedre information end man har haft hidtil om ordfrekvens, forskellen mellem teksttyper, nye ord, ændringer i bøjningen osv. Korpusset er halvtreds gange så stort som det største taggede korpus man har haft hidtil, nemlig det som *Íslensk orðtíðnibók* (islandsk frekvensordbog, Pind m.fl. 1991) er baseret på, og det indeholder mange teksttyper som ikke fandtes i basen til frekvensordbogen. Særlig værdifuldt er det at have tekster af

uformelt register, såsom blog og talesprog. I disse tekster kan man finde mange sprogændringer som ikke forekommer i de mere formelle tekster og ikke hidtil har været registreret i ordbøger. Dette handler naturligvis om nye ord, både låneord og nydannelser, men også om ændringer i ordenes betydning og brug, ændringer i bøjningen osv.

Til tagningen bliver der brugt et detaljeret tagsæt som giver brugerne mulighed for at bruge teksten til syntaktisk forskning som også kan benyttes ved leksikografisk beskrivelse. Tagsættet er ganske vist morfologisk baseret og ikke syntaktisk, men da der er en nær forbindelse mellem morfologien og syntaksen, kan man også hente en del syntaktiske oplysninger i de morfologiske tags (se fx Rögnvaldsson og Helgadóttir 2008).

3.3. Andre sproglige ressourcer

To andre ressourcer som er blevet lavet inden for andre projekter som støttedes af regeringens sprogteknologiprogram, bør også nævnes. Den ene er en ordliste med lydskrift. I 2003 gik Islands Universitet og nogle private firmaer sammen om at lave en islandsk talegenkendelse. En vigtig del af projektet var at lave en fonetisk ordliste for islandsk. Der blev lavet en ordliste som indeholder 56.000 af de hyppigste ordformer i islandsk. Listen byggede på frekvenslister som blev sammenstillet ud fra forskellige tekster som blev stillet til rådighed for projektet – avistekster fra Islands største avis, *Morgunblaðið*, omkring 100 bøger fra Islands største forlag og flere mindre tekstsamlinger. Fire studerende inden for lingvistik og sprogteknologi fik til opgave at transskribere listen med lydsskriftsystemet SAMPA. Senere er transskriptionen blevet konverteret til IPA, så nu findes der en liste med to typer af fonetisk transskription. Denne liste er allerede blevet benyttet både til islandsk talegenkendelse og islandsk talesyntese. Den kan naturligvis også bruges til traditionelle ordbøger.

Den anden ressource er resultatet af et syntaktisk parsing-projekt som desværre ikke blev afsluttet. Det er en liste på 28.000 linjer med verber og deres argumentstruktur. Listen viser hvilke argumenter hvert verbum kan have, i hvilke kasus argumenterne står, hvilke præpositioner verberne tager osv. Hvert verbum får ofte mange linjer da mange verber kan have flere forskellige argumentstrukturer. Listen er lavet ud fra eksisterende ordbøger, og den kan naturligvis være af stor nytte i leksikografisk arbejde.

4. Islandsk sprogteknologi i dag

4.1. Sprogteknologiske værktøjer

Foruden de sproglige ressourcer som der er blevet gjort rede for ovenfor, er der blevet udviklet nogle sprogteknologiske værktøjer som naturligvis også kan være af stor nytte for leksikografien. Heldigvis var originalfilerne fra Islandsk frekvensordbog blevet opbevaret på Leksikografisk institut og kunne bruges til at træne statistiske taggere. Den første tagger som blev trænet ved brug af listerne, var μ-tbl (Lager 1999), men den tagger som gav det bedste resultat, var TnT (Brants 2000). Den bedste score man har fået, er 92% korrekt (Helgadóttir 2007). Det er vistnok ikke særlig godt hvis man sammenligner det med resultater for mange andre sprog, som engelsk eller svensk, men tagsættet er også usædvanlig stort – omkring 700 forskellige tags.

Hrafn Loftsson, lektor ved Reykjavik Universitet, har udviklet en regelbaseret tagger, IceTagger, som giver lidt bedre resultat end TnT (Loftsson 2008). Hrafn og flere har også eksperimenteret med forskellige kombinationer af taggere og forenkling af tagsættet (Henrich m.fl. 2009). Dette giver op til 93,70% korrekt tagging.

Hrafn har også skrevet en syntaktisk parser, IceParser (Loftsson og Rögnvaldsson 2007). Dette er en såkaldt "shallow parser"

som ikke udfører en fuld syntaktisk analyse, men genkender de vigtigste syntaktiske fraser, såsom nominalfraser, præpositionsfraser osv. Parseren markerer også de vigtigste syntaktiske funktioner, såsom subjekt, objekt osv. IceParser udfører en analyse som minder om Constraint Grammar (CG, se Karlsson 1990), men den er dog ikke CG-baseret. Hrafn har imidlertid planer om at omskrive sin parser så at den bliver ren CG-parser.

Indtil for nylig har man ikke haft nogen lemmatiser for islandsk. Der er blevet eksperimenteret med at træne CST's lemmatiser (http://cst.dk/online/lemmatiser/) på islandsk tekst, og det gav et godt resultat. Anton Karl Ingason, som er studerende inden for sprogteknologi og lingvistik, har nu skrevet en lemmatiser for islandsk, Lemmald, som giver lidt bedre resultater (Ingason m.fl. 2008).

Alle disse værktøjer er nu online på http://nlp.cs.ru.is. De kan bruges én ad gangen eller alle samtidig (dvs. man kan få teksten PoS-tagget, parset og lemmatiseret på én gang).

4.2. Nuværende projekter

Efter at regeringens sprogteknologiprojekt blev afsluttet i slutningen af 2004, så det ud til at udviklingen af sprogteknologiske værktøjer for islandsk ikke ville kunne fortsættes. Men heldigvis havde man på dette tidspunkt fået samlet en solid gruppe forskere og studerende fra tre institutioner – Islands Universitet, Reykjavik University og Árni Magnússon-instituttet for islandske studier. Denne gruppe har stået bag næsten alle sprogteknologiske projekter på Island i løbet af de sidste ti år.

I begyndelsen af 2009 fik sprogteknologigruppen tildelt et stort treårs stipendium fra den islandske forskningsfond til projektet *Viable Language Technology beyond English – Icelandic as a Test Case* (http://iceblark.wordpress.com). Formålet med dette projekt er at fortsætte med at opbygge ressourcer for islandsk sprogtekno-

logi på den billigste og mest effektive måde og derved bidrage til en islandsk såkaldt BLARK (Basic Language Resource Kit, se Krauwer 2003). Projektet har tre delprojekter. Ét er det semantiske projekt som Anna Björk Nikulásdóttir og Matthew Whelpton (2010) beskriver i deres artikel i dette nummer af *LexicoNordica*. Et andet er maskinoversættelse fra islandsk til engelsk ved hjælp af Apertium-programmet der er udviklet ved universitetet i Alicante (http://www.apertium.org/). En prøveudgave af oversættelsessystemet er nu tilgængelig på http://nlp.cs.ru.is/ApertiumISENWeb/.

Det tredje og største delprojekt er en træbank, dvs. et syntaktisk analyseret korpus. Ved opbygningen af korpusset samarbejder projektet med en gruppe fra University of Pennsylvania som har stået bag nogle af de største og bedst kendte projekter af denne slags – Penn Treebank (Marcus m.fl. 1993) og Penn Parsed Corpora of Historical English (PPCME2, PPCEME). Det er kostbart og tidkrævende at opbygge en træbank, og derfor er det meget vigtigt at automatisere processen så vidt som muligt. Der eksperimenteres p.t. med at bruge IceTagger og IceParser til præliminær analyse og koble dem til programmer fra sprogteknologigruppens medarbejdere i University of Pennsylvania. En præliminær version af træbanken (IcePaHC) er allerede blevet lagt ud på internettet og kan downloades fra http://linguist.is/icelandic_treebank.

5. Konklusion

I denne artikel er der blevet gjort rede for de vigtigste sprogteknologiske ressourcer og værktøjer som findes for islandsk. Men det er ikke nok at disse ting eksisterer; et stort og velkendt problem ved mange sproglige ressourcer at de er kommercielle og må købes – ofte til en høj pris. Da den islandske regerings sprogteknologiprojekt begyndte i 2001, blev det samtidig besluttet at staten skulle finansiere opbygningen af forskellige ressourcer for islandsk

sprogteknologi da det islandske marked er alt for lille til at kommercielle firmaer har råd til at bygge sådanne ressourcer op.

Det blev desværre ikke besluttet fra begyndelsen at disse ressourcer skulle være gratis, kun at alle som ville bruge dem til sprogteknologiske projekter skulle kunne få dem til rimelig pris. Man håbede at de penge som firmaer ville betale for at få lov til at benytte dem, ville være nok til at vedligeholde dem. Det har dog vist sig at det er vældig svært at finde ud hvad der er en rimelig pris, og at selv en lav pris er en stor tærskel for benyttelse af ressourcerne. Der er også mange som gerne ville eksperimentere med ressourcerne, men ikke kan betale for dem og ikke bryder sig om eller finder det værd at læse og underskrive alle mulige kontrakter for at kunne bruge dem til videnskabelige formål. Dette har ført til at ressourcerne ikke er brugt så meget som man havde håbet.

Derfor har den islandske sprogteknologigruppe indset at det er nødvendigt at alle sproglige og sprogteknologiske ressourcer for islandsk bliver open source i videst muligt omfang. Dette er allerede blevet gjort ved de fleste af de sprogteknologiske værktøjer som medlemmer af gruppen har lavet. Taggeren IceTagger, parseren IceParser og lemmatiseringsprogrammet Lemmald indgår alle i programpakken IceNLP som er licenseret under GNU LGPL (Lesser General Public License, se http://www.gnu.org/licenses/lgpl.html) og findes på http://sourceforge.net/projects/icenlp.

Det er klart at disse ressourcer og værktøjer kan blive af stor nytte for leksikografien. Men der mangler endnu forskellige grundlæggende ressourcer, ikke mindst en god sprogteknologisk orddatabase som fx den danske STO, med morfologiske, syntaktiske og semantiske oplysninger. Der eksisterer faktisk meget materiale til en orddatabase af denne slags, og man kan vel sige at en af de vigtigste opgaver for islandsk sprogteknologi i de næste år, er at finde en måde at koble alle disse ressourcer sammen. Det ville åbne nye og fascinerende muligheder for islandsk leksikografi.

Litteratur

Arnalds, Ari 2004: Language Technology in Iceland. I: Henrik Holmboe (red.): *Nordisk Sprogteknologi. Årbog 2003*. København: Museum Tusculanums Forlag, 41–43.

Bjarnadóttir, Kristín 2005: Modern Icelandic Inflections. I: Henrik Holmboe (red.): *Nordisk Sprogteknologi. Årbog 2005*. København: Museum Tusculanums Forlag, 49–50.

Brants, Thorsten 2000: TnT – A Statistical Part-of-Speech Tagger. I: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, 224–231.

Henrich, Verena, Timo Reuter og Hrafn Loftsson 2009: CombiTagger: A System for Developing Combined Taggers. I: *Proceedings of the 22nd International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*. Sanibel Island, Florida. http://aaai.org/ocs/index.php/FLAIRS/2009/paper/view/67/296

Helgadóttir, Sigrún 2004: Mörkuð íslensk málheild. I: *Samspil tungu og tækni*. Reykjavík: Ministeriet for undervisning, forskning og kultur, 67–71.

Helgadóttir, Sigrún 2007: Mörkun íslensks texta. I: *Orð og tunga 9*, 75–107.

Helgadóttir, Sigrún 2009: Mörkun texta og markaðar málheildir. Foredrag på Hugvísindaþing, Reykjavík, 14. marts.

Ingason, Anton Karl, Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson 2008: A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). I: Bengt Nordström og Aarne Ranta (red.): *Advances in Natural Language Processing*. Lecture Notes in Computer Science, Vol. 5221. Berlin: Springer, 205–216.

Karlsson, Fred 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. I: Hans Karlgren (red.): *Proceedings*

of the 13th International Conference of Computational Linguistics, Vol. 3. Helsingfors, 168–173.

Krauwer, Steven 2003: The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. I: *Proceedings of SPECOM 2003.* Moskva, 8–15.

Lager, Torbjörn 1999: The μ-TBL System: Logic Programming Tools for Transformation-Based Learning. I: *Proceedings of the Third International Workshop on Computational Natural Language Learning* (CoNLL'99), Bergen. http://www.cnts.ua.ac.be/conll99/programme.html

Loftsson, Hrafn 2008: Tagging Icelandic text: A linguistic rule-based approach. I: *Nordic Journal of Linguistics 31*, 47–72.

Loftsson, Hrafn, og Eiríkur Rögnvaldsson 2007: IceParser: An Incremental Finite-State Parser for Icelandic. I: Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek og Mare Koit (red.): *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, 128–135.

Loftsson, Hrafn, Jökull H. Yngvason, Sigrún Helgadóttir og Eiríkur 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, 53–60.

Marcus, Mitch, Beatrice Santorini og Mary Ann Marcinkiewicz 1993: Building a Large Annotated Corpus of English: The Penn Treebank. I: *Computational Linguistics 19(2)*, 313–330.

Nikulásdóttir, Anna Björk og Matthew Whelpton 2010: Lexical Acquisition through Noun Clustering. I: *LexicoNordica 17* (i dette bind).

Ólafsson, Rögnvaldur 2004: Tungutækniverkefni menntamálaráðuneytisins. I: *Samspil tungu og tækni*. Reykjavík: Ministeriet for undervisning, forskning og kultur, 7–13.

Ólafsson, Rögnvaldur, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson 1999: *Tungutækni. Skýrsla starfshóps.* Reykjavík: Ministeriet for undervisning, forskning og kultur.

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

Rögnvaldsson, Eiríkur 2005: Staða íslenskrar tungutækni við lok tungutækniátaks *Tölvumál*, 24. februar.

Rögnvaldsson, Eiríkur 2008: Icelandic Language Technology Ten Years Later. I: *Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages.* SALTMIL workshop, LREC 2008. Marrakech, 1–5.

Rögnvaldsson, Eiríkur og Sigrún Helgadóttir 2008: Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change. I: *2nd Workshop on Language Technology for Cultural Heritage Data.* LREC 2008 workshop. Marrakech, 40–46.

Rögnvaldsson, Eiríkur, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna Björk Nikulásdóttir, Matthew Whelpton og Anton Karl Ingason 2009: Icelandic Language Resources and Technology: Status and Prospects. I: Rickard Domeij, Kimmo Koskenniemi, Steven Krauwer, Bente Maegaard, Eiríkur Rögnvaldsson og Koenraad de Smedt (red.): *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Tartu: Northern European Association for Language Technology (NEALT), Tartu University Library, 27–32.

Internethenvisninger

BÍN = Bjarnadóttir, Kristín (red.): *Beygingarlýsing íslensks nútímamáls.* http://bin.arnastofnun.is

Icelandic Online. Undervisningsmateriale i islandsk for begyndere. http://icelandic.hi.is/

IcePaHC = Wallenberg, Joel, Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson 2010. *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.1. http://linguist.is/icelandic_treebank

PPCEME = Kroch, Anthony, Beatrice Santorini og Lauren Delfs 2004: *Penn-Helsinki Parsed Corpus of Early Modern English.* http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/

PPCME2 = Kroch, Anthony og Ann Taylor 2000: *Penn-Helsinki Parsed Corpus of Middle English, second edition.* http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/

STO = Braasch, Anna m.fl. (red.): *Sprogteknologisk Ordbase.* København: Center for Sprogteknologi 2001–2004. http://cst.dk/sto

Eiríkur Rögnvaldsson
professor
Islands Universitet
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is