GRIPLA

Gripla er alþjóðlegur vettvangur fyrir rannsóknir á sviði íslenskra og norrænna fræða. Birtar eru útgáfur á stuttum textum, greinar og ritgerðir og stuttar fræðilegar athugasemdir. Greinar skulu að jafnaði skrifaðar á íslensku en einnig eru birtar greinar á öðrum norrænum málum, ensku, þýsku og frönsku. Leiðbeiningar um frágang handrita er að finna á heimasíðu Árnastofnunar: arnastofnun.is/page/arnastofnun_timarit_gripla_leidb. Allt efni sem birtast á er lesið yfir af sérfræðingum. Greinum og útgáfum (öðrum en stuttum athugasemdum o. þ. h.) skal fylgja útdráttur á ensku. Greinum á öðrum málum en íslensku skal einnig fylgja útdráttur á íslensku. Hverju bindi Griplu fylgir handritaskrá.

GRIPLA

RITSTJÓRAR

GÍSLI SIGURÐSSON OG VIÐAR PÁLSSON

XXIII

REYKJAVÍK
STOFNUN ÁRNA MAGNÚSSONAR Í ÍSLENSKUM FRÆÐUM
2012

STOFNUN ÁRNA MAGNÚSSONAR Í ÍSLENSKUM FRÆÐUM
RIT 85

Prófarkalestur
HÖFUNDAR, RITSTJÓRAR OG SVANHILDUR GUNNARSDÓTTIR

Aðstoð við prófarkalestur
EMILY LETHBRIDGE, REYNIR ÞÓR EGGERTSSON
KATELIN PARSONS OG ANDREW WAWN

Umbrot
SVERRIR SVEINSSON

Prentun og bókband
LITLAPRENT EHF.

Prentþjónusta og dreifing
HÁSKÓLAÚTGÁFAN

Handritaskrá:
Sameiginleg handritaskrá Griplu 23 (2012) og 24 (2013) fylgir hinu síðara hefti

Meginmál þessarar bókar er sett með 10,5 punkta Andron Mega Corpus letri
á 13,4 punkta fæti og bókin er prentuð á 115 gr. Munken Pure 13 pappír

PRINTED IN ICELAND

ISSN 1018-5011
ISBN 978 9979 654 24 7

EIRÍKUR RÖGNVALDSSON, ANTON KARL INGASON,
EINAR FREYR SIGURÐSSON, JOEL C. WALLENBERG

SÖGULEGI ÍSLENSKI TRJÁBANKINN

1. Inngangur

HAUSTIÐ 2011 LAUK gerð ***Sögulegs íslensks trjábanka*** (*Icelandic Parsed Historical Corpus, IcePaHC*),[1] safns þáttaðra (setningafræðilega greindra) texta frá 12. til 21. aldar – alls ein milljón lesmálsorða[2] úr rúmum 60 textum eða textabútum (Wallenberg o.fl. 2011). Frumgreining textanna var vélræn en meginvinnan við bankann fólst í handvirkri þáttun textanna sem var mikið verk. Eins og venja er í trjábönkum er bæði greind formgerð setninga og setningafræðileg vensl. Framsetningin er oft í formi setningafræðilegra hríslna eða trjáa, og af því er dregið heitið **trjábanki** (e. *treebank*).

1 Smíði *Sögulega íslenska trjábankans* var kostuð af styrk Rannsóknasjóðs til verkefnisins „Hagnýt máltækni utan ensku“ (nr. 090662011); U.S. National Science Foundation (NSF) International Research Fellowship Program (IRFP), grant #OISE-0853114, „Evolution of Language Systems: a comparative study of grammatical change in Icelandic and English“; styrk Rannsóknasjóðs Háskóla Íslands til verkefnisins „Sögulegur íslenskur trjábanki“; og styrk frá EU ICT Policy Support Programme sem hluta af „Competitiveness and Innovation Framework Programme“, styrknúmer 270899 (META-NORD). Við stöndum í þakkarskuld við ýmsa fræðimenn og rithöfunda sem létu okkur í té texta sem þeir eru að gefa út eða hafa skrifað. Við þökkum Hrafni Loftssyni dósent sem er aðalhöfundur *IceNLP* hugbúnaðarpakkans, Brynhildi Stefánsdóttur og Huldu Óladóttur sem unnu við þáttun textans, og stúdentum sem slógu inn allmarga texta. – Trjábankinn hefur verið kynntur á ýmsum vettvangi, s.s. á RILiVS-vinnustofu í Osló í september 2009 (Eiríkur Rögnvaldsson, Anton Karl Ingason og Einar Freyr Sigurðsson 2011), í fyrirlestrum við University of Pennsylvania, University of Massachusetts og New York University í maí 2010, á Hugvísindaþingi í Reykjavík í mars 2011 og 2012, á MENOTA-fundi í Reykjavík í ágúst 2011, á ACRH-vinnustofunni í Heidelberg í janúar 2012 (Eiríkur Rögnvaldsson o.fl. 2011), o.v. Við þökkum áheyrendum á þessum stöðum fyrir gagnlegar umræður og athugasemdir. Síðast en ekki síst þökkum við samstarfsfólki okkar við Pennsylvaníuháskóla, einkum Tony Kroch og Beatrice Santorini, fyrir ómetanlegt framlag til verksins. Að auki fá tveir nafnlausir ritrýnar þakkir fyrir ýmsar gagnlegar ábendingar.

2 Hér er orðið ***lesmálsorð*** notað yfir það sem nefnist „running word“ eða „token“ á ensku, eins og gert er í *Íslenskri orðtíðnibók* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991). Fjöldi lesmálsorða er þannig mælikvarði á lengd texta.

Gripla XXIII (2012): 331–352.

Sögulegi íslenski trjábankinn er gerður í tvennum tilgangi. Annars vegar til nota í máltækni, en nákvæmar upplýsingar um setningagerð eru mikilvæg forsenda fyrir gerð ýmiss konar máltæknibúnaðar, svo sem leiðréttingarforrita, vélrænna þýðinga, tölfræðilegra þáttara o.fl. Hins vegar er bankinn ætlaður til málrannsókna, einkum á setningagerð og setningafræðilegum breytingum, og hefur þegar sannað gildi sitt í ýmsum rannsóknum af því tagi.

Á undanförnum árum hefur verið unnið að smíði viðamikilla trjábanka fyrir ýmis tungumál en *Sögulegi íslenski trjábankinn* er einstakur að ýmsu leyti, að því er við teljum:

- Í fyrsta lagi er hann frá upphafi ætlaður til nota bæði í máltækni og málfræðilegum rannsóknum. Flestir trjábankar eru annaðhvort gerðir til nota innan máltækni (s.s. *Penn Treebank*, sjá 2. kafla) eða til setningafræðilegra rannsókna (s.s. sögulegu ensku trjábankarnir, sjá 2. kafla), en ekki hvors tveggja.
- Í öðru lagi spannar trjábankinn á tíundu öld – elstu textarnir eru frá lokum 12. aldar en þeir yngstu frá fyrsta áratug 21. aldar. Flest tungumál hafa breyst svo mikið á undanförnum þúsund árum að það væri hvorki gagnlegt né raunhæft að hafa texta frá svo löngum tíma í einum og sama trjábankanum.
- Í þriðja lagi hefur trjábankinn að geyma eina milljón lesmálsorða og er því eitt stærsta safn þáttaðra texta sem til er fyrir nokkurt tungumál. Til eru mun stærri trjábankar sem hafa verið þáttaðir á vélrænan hátt, en stærri handleiðréttir trjábankar munu aðeins vera til fyrir tvö tungumál – ensku (*Penn Treebank*, sjá 2. kafla) og tékknesku (*Prague Dependency Treebank*, sjá Hajič 2005 og http://ufal.mff.cuni.cz/pdt2.0/).
- Í fjórða lagi er trjábankinn algerlega opinn og aðgengilegur öllum, án nokkurra leyfa eða skráningar, og sama máli gegnir um allan hugbúnað sem notaður var til að smíða hann, svo og þann hugbúnað sem varð til innan verkefnisins. Bæði hugbúnaðinum og trjábankanum sjálfum er dreift með stöðluðu leyfi (LGPL; sjá http://www.gnu.org/licenses/lgpl.html).

Í þessari grein er gerð grein fyrir forsendum trjábankans og vinnunni við gerð hans, og tekið dæmi um hugsanlega nýtingu. Í 2. kafla er sagt frá

aðdraganda að smíði bankans. Í 3. kafla er gerð grein fyrir efnivið bankans og ýmsum vafamálum í meðferð og túlkun hans. Í 4. kafla er vinnulagi við gerð bankans lýst, svo og greiningaraðferðum. Í 5. kafla er tekið dæmi um hvernig hægt er að nota bankann við rannsóknir á setningafræðilegum breytingum – í þessu tilviki notkun á fornafninu/töluorðinu *einn*. Í 6. kafla er gerð grein fyrir dreifingu bankans. 7. kafli er svo lokaorð.

2. Aðdragandi

Saga trjábanka hefst með *Penn Treebank* (http://www.cis.upenn.edu/~treebank/) sem gerður var um og upp úr 1990 við Pennsylvaníuháskóla í Bandaríkjunum (Marcus, Santorini og Marcinkiewicz 1993). Þar starfar rannsóknarhópur undir stjórn Anthony Krochs prófessors sem hefur rutt brautina í því að koma upp málfræðilega og setningafræðilega greindum forntextum til að nota við rannsóknir á eldri málstigum og málbreytingum. Þekktasta dæmið eru sögulegu ensku trjábankarnir, *Penn Parsed Corpora of Historical English*, *PPCHE* (http://www.ling.upenn.edu/hist-corpora/). Þar er um að ræða tvö söfn; *Penn-Helsinki Corpus of Middle English* (1,3 milljónir lesmálsorða; Kroch og Taylor 2000) og *Penn-Helsinki Parsed Corpus of Early Modern English* (1,8 milljónir lesmálsorða; Kroch, Santorini og Delfs 2004), en fleiri eru í vinnslu.

Á undanförnum áratugum hafa komið upp fjölmargar mismunandi stefnur og greiningaraðferðir í setningafræði og þessi fjölbreytni endurspeglast í þeim trjábönkum sem eru til eða í smíðum. Sumir þeirra eru gerðir út frá ákveðnum fræðikenningum en í öðrum er leitast við að hafa greininguna hlutlausari og óháða tilteknum kenningakerfum. Trjábankar af síðarnefndu gerðinni skiptast í tvo meginflokka; þá sem byggjast á venslamálfræði (e. *dependency grammar*), t.d. áðurnefndur Prague *Dependency Treebank* (sjá Inngang) og *PROIEL*-bankinn sem hefur að geyma gríska texta Nýja testamentisins og þýðingar hans á ýmis indóevrópsk mál (http://www.hf.uio.no/ifikk/english/research/projects/proiel/publications/; Haug og Jøhndal 2008), og þá sem byggjast á liðgerðarmálfræði (e. *phrase structure grammar*), t.d. Penn Treebank og sögulegu ensku trjábankarnir áðurnefndu. Í fyrrnefndu tegundinni er meginákhersla lögð á að sýna innbyrðis vensl orða en í síðarnefndu tegundinni er áhersla lögð á að sýna stigveldisuppbyggingu setningarliða og línulega röð.

Tildrögin að smíði trjábankans voru þau að Eiríkur Rögnvaldsson hafði lengi haft í huga að gera trjábanka fyrir íslensku til nota í ýmsum máltækniverkefnum, og m.a. fengið undirbúningsstyrk til þess verks úr Rannsóknasjóði Háskólans 2003. Skriður komst þó fyrst á málið þegar verkefnið „Hagkvæm máltækni utan ensku – íslenska tilraunin“ fékk öndvegisstyrk Rannsóknasjóðs í ársbyrjun 2009. Innan þess verkefnis voru þrír meginþættir og sá stærsti þeirra gerð íslensks trjábanka. Sá trjábanki átti einkum að innihalda texta úr nútímamáli, enda ætlaður til nota innan máltækni, en þó var gert ráð fyrir að hafa fáeina eldri texta í honum einnig. Sumarið 2009 hófu meistaranemarnir Anton Karl Ingason og Einar Freyr Sigurðsson undirbúning að gerð þessa trjábanka.

Í ársbyrjun 2010 kom Joel Wallenberg til Íslands sem nýdoktor við Málvísindastofnun með styrk frá National Science Foundation (NSF) í Bandaríkjunum til að bera saman málbreytingar í ensku og íslensku. Um svipað leyti fékk Eiríkur Rögnvaldsson styrk úr Rannsóknasjóði Háskólans til að vinna að sögulegum trjábanka. Við þetta gerbreyttust aðstæður. Joel var að koma frá Pennsylvaníuháskóla þar sem hann hafði nýlokið doktorsritgerð undir leiðsögn Anthony Kroch prófessors, sem er brautryðjandi í gerð sögulegra trjábanka eins og áður segir.

Eins og nefnt er í Inngangi eru flestir trjábankar hannaðir til notkunar annaðhvort í setningafræðirannsóknum eða máltækni en yfirleitt ekki til hvors tveggja. Setningafræðileg nýting krefst yfirleitt dýpri og nákvæmari greiningar, auk ítarlegra upplýsinga um textana sem greindir eru. Fyrir hagnýtingu í máltækni skiptir textamagnið oft meira máli en dýpt greiningarinnar. Trjábankar eru þar ekki síst nýttir til að þjálfa greiningarforrit, þáttara, sem síðan er hægt að nota til að þátta aðra texta, og til að slík þjálfun verði árangursrík þarf mikið magn þáttaðra texta.

Þetta þýðir þó ekki endilega að grundvallarmunur sé eða þurfi að vera á trjábönkum eftir því hvernig ætlunin er að nota þá. Að athuguðu máli varð niðurstaðan sú að sameina framangreind verkefni, þannig að Joel gekk til liðs við Eirík, Anton og Einar. Við sáum að með því móti væri hægt að smíða trjábanka sem yrði mun stærri en ella hefði verið hægt, og gæti gagnast bæði innan máltækni og til setningafræðilegra rannsókna. Uppbygging gagnasafna á borð við trjábanka er mjög dýr og tímafrek og því mikilvægt að nýta vinnuna sem best og koma í veg fyrir tvíverknað. Við teljum að það hafi tekist einkar vel í þessu tilviki.

3. Efniviður

3.1 Textaflokkar

Upphafleg hugmynd okkar var að hafa fimm textategundir í trjábankanum – frásagnartexta, trúarlega texta, ævi- og ferðasögur, laga- og dómatexta, og fræðitexta. Við stefndum að því að vera með u.þ.b. 20 þúsund lesmálsorð af hverri textategund frá hverri öld – alls 100 þúsund lesmálsorð frá hverri öld, um milljón lesmálsorð í heildina. Við gerðum okkur þó grein fyrir því að það yrði erfitt eða útilokað að afla texta af öllum þessum tegundum frá öllum öldum. Að hluta til er það vegna þess að textarnir eru alls ekki til. En önnur ástæða er sú að við vorum bundnir við að nota texta sem voru til rafrænir, eða a.m.k. prentaðir. Við létum reyndar slá inn allnokkra texta en eingöngu eftir prentuðum bókum. Það hefði verið of kostnaðarsamt að láta slá inn texta eftir handritum.

Við endurskoðun upphaflegrar áætlunar var ákveðið að leggja megináherslu á frásagnartexta. Á fyrri öldum eru það Íslendingasögur, samtíðarsögur, riddarasögur o.fl., en á 19.-21. öld einkum skáldsögur. Niðurstaðan varð sú að frásagnartextar eru u.þ.b. 2/3 af heildinni – 675 þúsund lesmálsorð. Einnig stefndum við að því að vera með trúarlega texta frá öllum öldum. Það tókst að mestu leyti þótt eina öld vanti inn í, og trúarlegir textar eru tæpur fjórðungur af heildinni. Ævisögur og ferðabækur eru um 75 þúsund lesmálsorð, en orðafjöldinn úr þeim tveimur textaflokkum sem eftir standa er óverulegur. Dreifing texta eftir öldum og textaflokkum er sýnd í Töflu 1.

Öld	Frásagnir	Trúarrit	Ævisögur	Fræði	Lög	Samtals
12.	0	40.871	0	4.439	0	45.310
13.	93.463	21.196	0	0	6.183	120.842
14.	77.370	21.315	0	0	0	98.685
15.	111.560	0	0	0	0	111.560
16.	35.733	60.464	0	0	0	96.197
17.	46.281	28.134	52.997	0	0	127.412
18.	63.322	22.963	22.099	0	0	108.384
19.	100.921	20.370	0	3.258	0	124.000
20.	103.921	21.234	0	0	0	125.155
21.	43.102	0	0	0	0	43.102
Samtals	675.114	236.547	75.096	7.707	6.183	1.000.647

Tafla 1: Fjöldi lesmálsorða úr hverjum textaflokki frá hverri öld.

Í trjábankanum eru rúmlega 60 textar eða bútar úr textum sem fengnir eru úr ýmsum áttum. U.þ.b. 20 textar fengust úr ýmsum textasöfnum á netinu, einkum frá *Netútgáfunni* (*http://snerpa.is/net*) en einnig frá *Gutenberg-verkefninu* (*http://www.gutenberg.org*), *Internet Archive* (*http://www.archive.org/*) og *Medieval Nordic Text Archive* (*Menota*, *http://www.menota.org/*). Um 10 textar komu úr textasafni Stofnunar Árna Magnússonar (*http://www.lexis.hi.is/corpus/*). Við fengum 10 texta beint frá fræðimönnum sem vinna að útgáfu þeirra eða frá forlögum sem hafa gefið þá út. Afgangurinn, um 20 textar, var svo sleginn inn fyrir okkur af stúdentum sem unnu að verkefninu. Fjórir textar frá 20. og 21. öld eru enn í höfundarrétti en við fengum leyfi höfundanna til að nota þá og dreifa þeim.

Gæði textanna voru misjöfn. Í flestum tilvikum höfðum við þó aðgang að traustum útgáfum, ýmist stafréttum eða með samræmdri stafsetningu. Þannig fengum við t.d. nýja útgáfu Ármanns Jakobssonar og Þórðar Inga Guðjónssonar á *Morkinskinnu*, útgáfu Þórunnar Sigurðardóttur á *Fimmtíu heilögum hugvekjum*, nýja útgáfu Jóhannesar Bjarna Sigtryggssonar á *Ævisögu Jóns Steingrímssonar*, útgáfu Matthews J. Driscoll á *Fimmbræðra sögu*, og óprentaða útgáfu Svanhildar Óskarsdóttur á *Júditarbók*, svo að dæmi séu nefnd. Útgáfur Agnete Loth á *Late Medieval Icelandic Romances* og *Reykjahólabók* komu okkur einnig að góðu gagni. Síðastnefndu textarnir voru slegnir inn fyrir verkefnið, en hina fengum við alla rafræna hjá útgefendum.

Stafsetning textanna var með ýmsu móti. Sumir voru með nútímastafsetningu – að sjálfsögðu 20. og 21. aldar textarnir, en einnig ýmsir þeirra eldri sem hafa verið gefnir þannig út. Aðrir voru með samræmdri stafsetningu fornri, t.d. *Morkinskinna*. Enn aðrir voru stafréttir eftir handritum, bæði ýmsir fornir textar (t.d. úr *Late Medieval Icelandic Romances*) og yngri (t.d. *Fimmbræðra saga*). Við ákváðum að færa alla texta til nútímastafsetningar. Það auðveldar mjög leit að einstökum orðum – nægilegt er að slá inn nútímamálsmyndina til að finna öll dæmi um tiltekið orð, þótt það sé skrifað á ýmsan hátt í textunum.

Meginástæðan var þó sú að við vildum geta notað greiningarforrit sem hafa verið skrifuð fyrir íslensku – markarann *IceTagger* (Hrafn Loftsson 2008), þáttarann *IceParser* (Hrafn Loftsson og Eiríkur Rögnvaldsson 2007) og lemmunarforritið *Lemmald* (Anton Karl Ingason o.fl. 2008), sem saman mynda hugbúnaðarpakkann *IceNLP* (hægt er að sækja hann á

http://icenlp.sourceforge.net/). Þessi forrit flýttu mjög fyrir vinnunni og því töldum við mjög mikilvægt að geta notað þau. Þau miðast hins vegar öll við nútímastafsetningu og það hefði verið tímafrekt og erfitt að breyta þeim þannig að þau réðu sæmilega við öll þau stafsetningartilbrigði sem fram koma í textunum.

Það er tiltölulega einfalt að breyta samræmdri stafsetningu fornri í nútímastafsetningu – vörpunin í þá átt er einkvæm í flestum tilvikum og því hægt að nota einfaldar skriftur eða leit og skipti. Stafréttu útgáfurnar voru hins vegar mun erfiðari viðfangs og tímafrekari. Vissulega er hægt að nota leit og skipti að vissu marki, en þó var nauðsynlegt að fara yfir textana orð fyrir orð og breyta þeim. Sú vinna var tafsöm og hætt við villum. Sumir textarnir höfðu líka verið skannaðir og þá bættust skönnunarvillur við. Það er nokkuð víst að eitthvað af villum og ósamræmi er enn í stafsetningu textanna. Það ætti þó yfirleitt ekki að hafa áhrif á notagildi þeirra til setningafræðirannsókna.

3.2 Aldur og tímasetning

Einn megintilgangurinn með gerð trjábankans er að koma upp tæki til að rannsaka og tímasetja setningafræðilegar breytingar. Þess vegna er vitanlega mikilvægt að hægt sé að ákvarða aldur textanna með sæmilegri nákvæmni. Á því eru hins vegar mikil vandkvæði eins og alkunna er og óþarft að rekja hér. Forntextar, aðrir en bréf, eru ekki varðveittir í frumriti og erfitt að vita hvort og hvaða breytingar skrifarar hafa gert í uppskriftum. Haraldur Bernharðsson hefur vissulega skoðað málblöndun í handritum en sú rannsókn takmarkast við breytingar á stafsetningu og beygingum (Haraldur Bernharðsson 1999). Setningafræðilegar breytingar skrifara hafa hins vegar lítt verið kannaðar skipulega (sjá þó Kjartan Ottósson 2001).

Það er því oft erfitt að vita hvaða málstigi tiltekinn texti tilheyrir. Hvað á t.d. að gera við texta sem talinn er saminn á miðri 14. öld en varðveittur í handriti frá miðri 15. öld? Sýnir hann setningagerð 14. aldar, 15. aldar, eða einhverja blöndu af þessu tvennu? Um það er útilokað að fullyrða nokkuð. Segjum nú samt að við gætum slegið því föstu að setningagerð frumritsins skilaði sér óbreytt í varðveittri gerð. Eftir sem áður vitum við ekki hversu dæmigerð setningagerð upphaflegs skrifara var – hann gæti t.d. hafa fyrnt mál sitt, orðið fyrir áhrifum frá þýðingum úr erlendum málum, o.s.frv.

Í tímasetningu okkar á textunum höfum við í flestum tilvikum

miðað við það sem talið er líklegur ritunartími frumtextans. Þar er þó vissulega margt á huldu og líklega er ekki fullt samræmi í þessu hjá okkur. Meginatriðið er þó að á heimasíðu trjábankans er að finna ítarlegar upplýsingar um hvaðan hver texti er fenginn. Þegar um er að ræða prentaðar útgáfur, eins og er í langflestum tilvikum, getur notandinn þá farið í útgáfuna og kynnt sér það sem þar er sagt um uppruna textans. Viðfangsefni okkar var að búa til rannsóknartæki og leiðbeina mönnum um notkun þess, en við leggjum áherslu á að notendur bankans beri ábyrgð á túlkun niðurstaðna sinna.

Ef trjábankinn er notaður til að rekja sögulega þróun, t.d. sýna hvernig tiltekin málbreyting breiðist út, skiptir tímasetning textanna vissulega máli. Röng tímasetning getur auðvitað skekkt niðurstöðurnar verulega. Á það er hins vegar að líta að í bankanum eru a.m.k. fimm textar sem taldir eru vera frá hverri öld. Yfirleitt er ekki heppilegt að byggja mikið á einstökum textum, heldur eðlilegt að taka nokkra texta saman, t.d. alla texta frá hverri öld, og skoða niðurstöður úr þeim sem heild. Jafnvel þótt allir textar væru örugglega rétt tímasettir eru margir aðrir þættir sem valda því að oft er óskynsamlegt að miða mikið við einstaka texta. Það gæti t.d. verið hálfrar aldar aldursmunur á höfundum tveggja texta sem skrifaðir eru sama árið. Þegar nokkrir textar eru teknir saman ættu sérkenni einstakra texta að jafnast út og heildin að gefa sæmilega rétta mynd.

En það er líka hugsanlegt að styðjast við trjábankann til að tímasetja texta. Ef tiltekinn texti sker sig mjög úr, fellur ekki inn í setningafræðilega þróun sem aðrir textar sýna, gæti það verið vísbending um að hann sé eldri eða yngri en hann hefur verið talinn. Það getur þá verið forvitnileg tilraun að færa textann til í tíma þangað til hann fellur nokkurn veginn að þróuninni, og skoða hvort hugsanlegt sé að sú tímasetning geti staðist. Vitanlega er yfirleitt ekki hægt að nota þetta eitt sér til að tímasetja texta, en það ætti að geta komið að gagni með öðru.

4. Vinnulag

Sem fyrr segir hófst smíði trjábankans fyrir alvöru í ársbyrjun 2010. Þá lá fyrir að velja greiningaraðferð, ákveða hvað og hvernig skyldi greint, og móta verklag við smíði bankans. Eins og áður segir er greining í trjábönkum með ýmsu móti, en við ákváðum að byggja íslenska trjábank-

ann á liðgerðargreiningu (e. *phrase structure annotation*). Meginástæða þess var sú að við vorum í samstarfi við rannsóknarhóp Anthony Kroch, en sá hópur hefur staðið að gerð sögulegu ensku trjábankanna sem eru af þessari gerð. Þar sem mikil líkindi eru með íslensku og fornensku gátum við haft mjög mikið gagn af greiningunni sem þar hafði verið unnin. Ítarleg handbók er til fyrir þessa greiningu þar sem lýst er hvernig farið er með margvíslegar setningagerðir (Santorini 2010), og sú lýsing gagnaðist okkur vel.

Annar ávinningur við að nýta þessa greiningu er sá að til er góður leitarhugbúnaður sem miðast við hana, og nýtist því beint í íslenska trjábankanum. Þar sem greining íslensku og eldri málstiga ensku er hliðstæð opnast möguleiki til margvíslegs samanburðar á setningagerð málanna og þróun hennar. Nú eru líka til eða í smíðum trjábankar þar sem sama greining er notuð fyrir ýmis önnur mál, s.s. frönsku (Martineu o.fl. 2010), portúgölsku (Galves og Faria 2010), snemmháþýsku (Light 2010), forngrísku (Beck 2011), færeysku (Anton Karl Ingason o.fl. 2012; Eiríkur Rögnvaldsson o.fl. 2012) og fleiri, og samanburður íslensku við þau mál verður þá einnig mögulegur. Einnig má nefna að þetta greiningarskema felur í sér meiri upplýsingar en mörg önnur, t.d. flest þeirra sem byggjast á venslagreiningu (e. *dependency*).[3]

Eftir að textunum hafði verið breytt í nútímastafsetningu tóku aðstoðarmenn úr hópi stúdenta við þeim og settu inn málsgreina- og setningaskil. Það er ekki alltaf auðvelt að gera slíkt vélrænt, en hins vegar eykur það mjög nákvæmni vélrænnar greiningar ef hún getur byggt á slíkum skilum. Að þessu loknu voru textarnir keyrðir í gegnum forritin í *IceNLP*-hugbúnaðarpakkanum – *IceTagger*, *IceParser* og *Lemmald*. Þessi forrit skiluðu grófri málfræðilegri og setningafræðilegri greiningu textans. Síðan voru keyrð ýmis heimasmíðuð forrit sem færðu úttakið úr *IceNLP* í það snið sem notað er í sögulegu ensku trjábönkunum. Í því fólst m.a. að afmarka setningarliði og breyta markamenginu (e. *tagset*), þ.e. þeim skammstöfunum sem notaðar eru til að tákna einstaka orðflokka og setn-

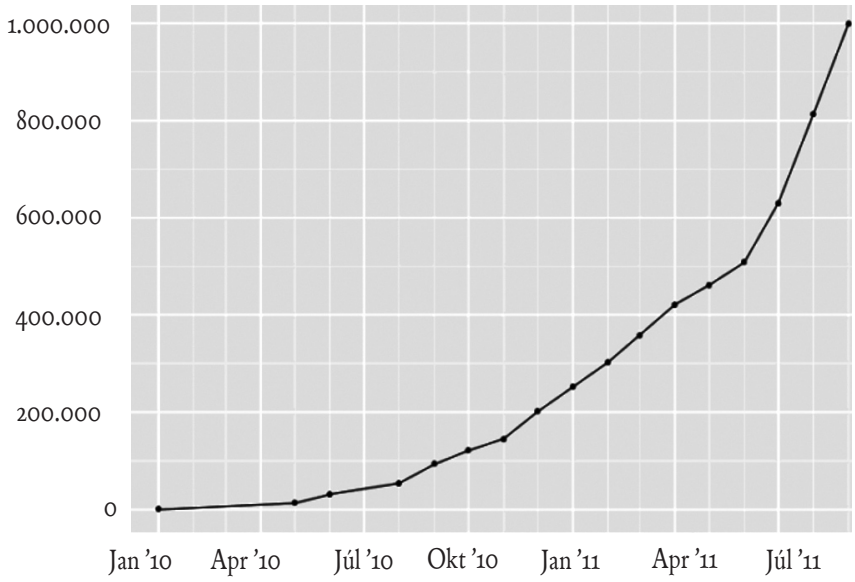
3 Þar er einkum um að ræða upplýsingar um orðaröð – venslagreiningin leggur megináherslu á vensl orða og liða, eins og áður er nefnt, en skeytir minna um línulega röð þeirra. Þetta er þó ekki algilt því að til eru ýmis afbrigði af venslagreiningu, rétt eins og liðgerðargreiningu, og sum þeirra hafa að geyma svipaðar upplýsingar og liðgerðargreining. Dag Haug, aðalhöfundur *PROIEL*-bankans (sjá 2. kafla), hefur t.d. skrifað forrit sem breytir venslagreiningu *PROIEL* í liðgerðargreiningu sögulegu Penn-trjábankanna (persónulegar upplýsingar).

ingarliði. Síðan tóku við aðstoðarmenn úr hópi stúdenta og settu inn setningaskil.

Notaðar eru enskar skammstafanir í greiningunni. Fyrir því eru nokkrar ástæður. Ein er sú að þannig er hægt að nota hugbúnað sem gerður hefur verið fyrir ensku trjábankana óbreyttan. Önnur ástæða er að það auðveldar samanburð við trjábanka fyrir önnur mál sem nýta sama greiningarskema. En meginástæðan er sú að trjábankinn á að geta nýst erlendum fræðimönnum þótt þeir kunni lítið í íslensku og þekki ekki íslensk heiti orðflokka og setningarliða. Þótt byggt sé á enska greiningarskemanu er vikið frá því í nokkrum atriðum þar sem það skiptir máli til að auðvelda skilvirka leit í trjábankanum. Þannig eru nefnimyndir (uppflettimyndir) orða ævinlega látnar fylgja í íslenska bankanum en það er ekki gert í þeim ensku. Enn fremur eru föll sýnd.

Síðan tók aðalvinnan við – að leiðrétta og endurskoða vélrænu greininguna. Hún var mjög seinleg og krafðist vitanlega margvíslegrar þekkingar – á íslensku máli og íslenskri setningagerð á öllum tímum, á greiningarskemanu, og á setningafræðilegri greiningu almennt. Til að flýta fyrir vinnunni var skrifaður sérstakur hugbúnaður, *Annotald* (Beck, Ecay og Anton Karl Ingason 2011), sem miðast við það að ná hámarkshraða við greiningu og leiðréttingu. Í forritinu er mikill fjöldi flýtiskipana sem eru hannaðar þannig að notandinn getur alltaf haft vinstri höndina á lyklaborðinu en þá hægri á músinni. Í fljótu bragði virðist þetta e.t.v. ekki skipta miklu máli, en þegar greina þarf eina milljón orða munar um hverja handahreyfingu. Það kom líka í ljós að hraðinn jókst talsvert þegar þetta forrit var tekið í notkun.

Þrír greinendur, Anton, Einar og Joel, unnu við þáttunina. Í upphafi fóru þeir hver yfir greiningu annars og tóku góðan tíma í að kynna sér vandlega greiningarleiðbeiningar með ensku sögulegu trjábönkunum (Santorini 2010) ásamt því að laga þær að íslensku eftir þörfum (http://www.linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)#Annotation_guidelines). Verkinu miðaði hægt í fyrstu á meðan verið var að móta greininguna en þegar greinendur voru orðnir þjálfaðir og höfðu tileinkað sér greiningarskemað og komið sér upp greiningarreglum fyrir flestar setningagerðir þar sem íslenska víkur frá eldri ensku hættu þeir að fara yfir greininguna hver hjá öðrum og lögðu áherslu á að hraða þáttun eftir mætti.



Mynd 1: Framvinda þáttunar.

Á Mynd 1 sést hvernig þáttuninni vatt fram þá 20 mánuði sem hún stóð yfir (frá janúar 2010 til ágústloka 2011). Við gerum okkur fulla grein fyrir því að hraði þáttunarinnar, ásamt því að talsverður hluti af henni er óyfirfarinn, leiðir til þess að í henni er án efa fjöldi af villum og margs kyns ósamræmi. Við teljum þó að villufjöldinn sé ekki ýkja mikill miðað við stærð bankans.Við höfum leitað kerfisbundið að villum og keyrt bankann í gegnum margvísleg próf. Við slíka keyrslu kom t.d. í ljós að 51 setning í bankanum hefur að geyma tvo nafnliði sem báðir eru merktir sem beint andlag (NP-OB1). Þetta er u.þ.b. 1% af þeim 4.727 setningum sem innihalda tvö andlög og í flestum tilvikum ætti annar liðurinn að vera merktur sem óbeint andlag (NP-OB^2) í staðinn.

Þótt vitanlega sé stefnt að því að hafa greininguna villulausa teljum við að meiri ávinningur sé að því að hafa bankann jafnstóran og raun ber vitni en hafa algerlega réttan banka sem þá væri ekki nema brot af stærð þessa. Við vonumst til að geta leiðrétt villur í bankanum smátt og smátt og gefið út nýjar og réttari gerðir af honum. Við viljum þó leggja áherslu á að bankinn er fyrst og fremst ætlaður til nota í megindlegum rannsóknum, ekki eigindlegum. Þegar verið er að skoða stöðu tiltekinnar setningagerðar í

málinu á tilteknum tíma, eða þróun hennar yfir ákveðið tímabil, má ætla að einstakar villur jafnist út og hafi ekki veruleg áhrif á niðurstöður.

Það er hins vegar ekki hægt að nota t.d. eitt dæmi í bankanum umhugsunarlaust til að sýna fram á að tiltekin setningagerð hafi verið til í málinu á tilteknum tíma. Þar gæti verið um að ræða villu í greiningu, setningin gæti verið tekin rangt upp úr útgáfunni eða útgáfan verið ónákvæm, tímasetning textans gæti verið vafasöm, o.s.frv. Á hinn bóginn er alltaf ljóst hvaðan hver setning er tekin, eins og áður segir. Notandinn þarf því ekki – og má ekki – treysta á greininguna í bankanum, heldur getur hann alltaf farið í frumheimildir til að sjá hvort setningin er rétt tekin upp, og beitt eigin setningafræðilegri kunnáttu til að endurskoða og leiðrétta greininguna.

5. Nýting

Við leit í bankanum er notað forritið *CorpusSearch* (Randall 2005; http://corpussearch.sourceforge.net/) sem er upphaflega gert til að nota með sögulegu ensku trjábönkunum. Þetta forrit fylgir með trjábankanum þegar hann er sóttur (sjá 6. kafla). Það notar sérstakt skipanamál sem gefur kost á mjög fjölbreyttri og nákvæmri leit að tilteknum setningagerðum. Fjölhæfni skipanamálsins hefur vissulega þann ókost að það tekur tíma að læra að nýta sér alla möguleika þess, en ítarlegar leiðbeiningar eru á netinu þannig að allir ættu að geta komist upp á lag með að nota forritið.

Sem dæmi um hvernig hægt er að nota trjábankann til að skoða sögulega þróun setningafræðilegra fyrirbæra skal hér litið á breytingar á tíðni fornafnsins/töluorðsins *einn*. Það er alkunna að á tímabili í málsögunni er orðið töluvert notað eins og óákveðinn greinir (sjá Eirík Rögnvaldsson 2005, 608–609):

(1) Það var **einn mann** í Englandi sem fleiri aðrir þó frá þessum verði nú sagt heldur en öðrum ...

(2) ... og sem veislan er sett og allir eru komnir í sitt sæti kemur inn **einn maður** ókenndur og settist niður og er heldur fölur ...

Þessi dæmi eru úr *Miðaldaævintýrum þýddum úr ensku* sem talin eru þýdd á seinni hluta 15. aldar og Einar G. Pétursson gaf út 1976. Slíkar setningar eru mjög áberandi í þýðingum, t.d. riddarasagna og trúarrita, og er varla óvarlegt að álykta að um erlend áhrif sé að ræða. Eftir því sem við best

vitum hefur ekki verið gerð nákvæm athugun á tíðni og þróun þessarar notkunar. Með hjálp trjábankans er hins vegar auðvelt að fá yfirlit um þetta. Þar má nota eftirfarandi leitarskipun í *CorpusSearch*:

(3) NP* **idoms** ONE-*
AND NP* **idoms** N-*
AND ONE-* **iprecedes** N-*

* (stjarna) er algildisstafur sem stendur fyrir hvaða staf eða stafastreng sem er –NP* táknar því nafnlið af hvaða gerð sem er (frumlag, andlag, o.s.frv.). Orðið *einn* fær sérmerkingu í bankanum, ONE, enda oft erfitt að greina það í orðflokk. Stjarnan er sett aftast í ONE-* til að fá öll dæmi um orðið, óháð falli. Sömuleiðis er stjarna sett á N-* til að fá öll nafnorð í eintölu, óháð falli. *idoms* táknar „immediately dominates", þ.e. „beint yfirskipað", og *iprecedes* táknar „immediately precedes", þ.e. „fer næst á undan". Hér er því leitað að dæmum þar sem *einn* og nafnorð í eintölu eru beinir stofnhlutar í nafnlið, og einn fer næst á undan nafnorðinu.

Þessi leit skilar á skjáinn öllum setningum sem falla að leitarskilyrðunum – alls 1518. Ein þeirra er „Í annan stað leit hann á einn kastala" úr *Ectors sögu* frá 15. öld, sem lítur svona út í trjábankanum:[4]

(4) ((IP-MAT (PP (P Í-í)
(NP (OTHER-A annan-annar) (N-A stað-staður)))
(VBDI leit-líta)
(NP-SBJ (PRO-N hann-hann))
(PP (P á-á)
(NP (ONE-A einn-einn) (N-A kastala-kastali)))
(. .-.))
(ID 1450.ECTORSSAGA.NAR-SAG,.1016))

Aftan við setningarnar kemur svo yfirlitstafla þar sem sést hversu mörg dæmi fundust í hverjum hinna 60 texta í bankanum. Þessa töflu er auðvelt að líma inn í töflureikni eins og Excel og vinna þar með tölurnar á ýmsan hátt – reikna út tíðni og hlutfall dæma eftir öldum, textategundum o.fl.

4 Athugið að mörkin (skammstafanirnar) tákna ýmist setningarliði eða orðflokka (tegundir orða). Í mörkum sumra setningarliða kemur fram bæði eðli liðarins (NP = „Noun Phrase", nafnliður) og hlutverk (SBJ = „Subject", frumlag). IP-MAT táknar aðalsetningu („Inflection Phrase", „matrix"), VBDI sögn (VB) í þátíð (D) og framsöguhætti (I). Aðrar skammstafanir eru væntanlega auðskildar; -A aftan við táknun orðflokks merkir þolfall.

Öld	Fjöldi lesmálsorða	Fjöldi setninga	Á 1000 lesmálsorð	Á 100 setningar
12. öld	45.310	2.291	0,77	1,53
13. öld	120.842	9.283	0,84	1,10
14. öld	104.742	9.000	1,27	1,48
15. öld	105.502	9.663	2,66	2,91
16. öld	89.197	5.944	1,61	2,42
17. öld	127.412	8.342	2,24	3,42
18. öld	108.384	6.212	1,98	3,46
19. öld	124.068	9.439	1,06	1,39
20. öld	125.155	9.785	1,15	1,47
21. öld	43.102	3.005	1,11	1,57

Tafla 2: Hlutfall dæma um einn af fjölda lesmálsorða og setninga eftir öldum

Í Töflu 2 kemur fram samanlagður fjöldi lesmálsorða og setninga í textum frá hverri öld. Síðan er reiknaður út fjöldi dæma um *einn* á hver 1.000 lesmálsorð, og á hverjar 100 setningar. Eins og sjá má hækka tölurnar mjög mikið á 15. öld og haldast háar næstu þrjár aldir. Á 19. öld verður talsverð lækkun, og enn meiri á þeirri 20. Reyndar hækka tölurnar aftur á 21. öldinni en þar er aðeins um tvo texta að ræða og óvarlegt að draga miklar ályktanir af þeim.

Það er líka hægt að skipta textunum niður á annan hátt. Í Töflu 3 eru nokkrar aldir teknar saman og skipt í þrjá hópa – 12.–14. öld, 15.–18. öld, og 19.–21. öld.

Öld	Fjöldi lesmálsorða	Fjöldi setninga	Á 1000 lesmálsorð	Á 100 setningar
12.–14.	295.833	22.898	1,08	1,39
15.–18.	405.556	27.837	2,16	3,15
19.–21.	292.355	22.279	1,10	1,45

Tafla 3: Hlutfall dæma um einn af fjölda lesmálsorða og setninga eftir tímabilum.

Hér sést munurinn enn betur. Textarnir frá 15.–18. öld eru að verulegu leyti þýðingar (riddarasögur, trúarrit) eða undir sterkum erlendum áhrifum (t.d. *Reisubók Jóns Indíafara*). Hlutfallsleg tíðni orðsins *einn* á þessum öldum er meira en tvöföld á við tíðnina á öldunum á undan og eftir. Það er athyglisvert að hlutfallið á 12.–14. öld og 19.–21. öld er nokkurn veginn hið sama. Það gæti bent til þess að í raun og veru hafi engin breyting orðið á notkun *einn* í „ómengaðri" íslensku – sú hækkun sem sést í rituðum textum frá 15.–18. öld stafi af eðli textanna en beri ekki vott um neina raunverulega málbreytingu.[5]

Um þetta er þó ekkert hægt að fullyrða án nákvæmari skoðunar sem ekki er á dagskrá hér. Tilgangur þessarar athugunar er ekki sá að fara í saumana á notkun *einn* sem e.k. óákveðins greinis, heldur að benda á hvernig hægt er að nota trjábankann til setningafræðilegra athugana og kalla á svipstundu fram niðurstöður sem áður hefði tekið óratíma að komast að.

6. Opin og frjáls dreifing

Við teljum að opinn aðgangur að rannsóknargögnum og niðurstöðum sé mjög mikilvægur fyrir eflingu vísinda og fræða og lögðum því frá upphafi áherslu á að bankinn yrði sem flestum að sem mestu gagni. Við settum fram 10 viðmið þar að lútandi:

1. Hægt er að sækja frumgögn í heild af netinu (gögnin eru ekki falin bak við leitarviðmót).
2. Opinn vefaðgangur er að allri skjölun.
3. Notendur þurfa ekki að skrá sig eða auðkenna eða skrifa undir samning til að fá aðgang.
4. Smíðaferlið notar eingöngu frjáls og opin hugbúnaðartól svo að hægt sé að endurnýta ferlið á gagnsæjan hátt.

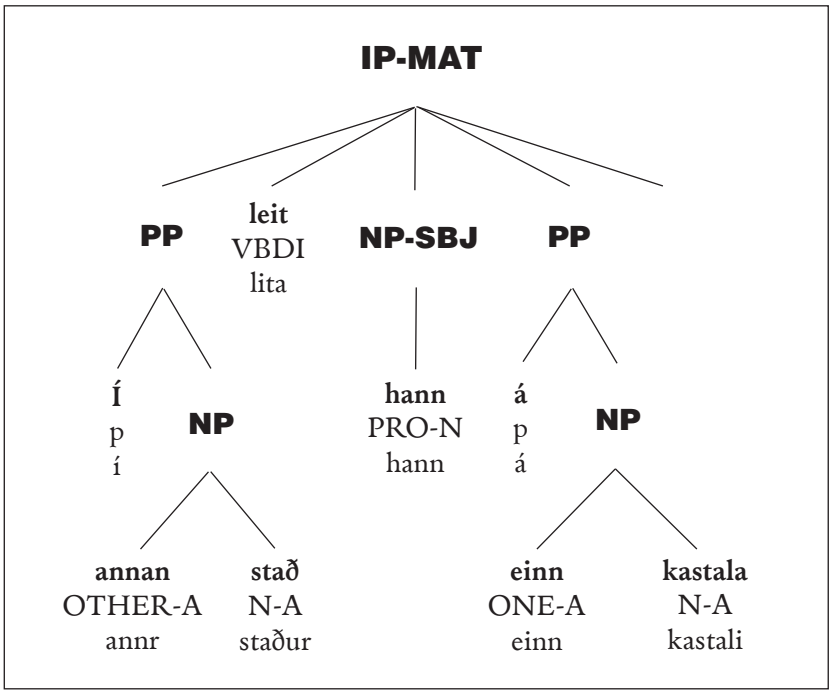
5 Ritrýnir nefnir að forvitnilegt væri að skoða hvort notkun *einn* sé mismunandi eftir textaflokkum. Vegna þess að fjöldi texta í flokkunum er mjög mismunandi eftir öldum – frá 15. öld t.d. einvörðungu frásagnartextar – er ekki hægt að sýna marktækar tölur um þetta. Tafla 2 sýnir þó að hlutfall *einn* er nokkuð svipað frá 15. til 18. aldar, þrátt fyrir ólík hlutföll textategunda á þessum öldum, og bendir það til þess að textategundin skipti ekki meginmáli.

5. Smíðin fer fram í opnu útgáfustýringarkerfi.
6. Tölusettar útgáfur eru settar reglulega á netið (á þriggja mánaða fresti).
7. Notendur geta breytt trjábankanum og gefið út breytingarnar án sérstaks leyfis.
8. Bankinn er ókeypis fyrir fræðimenn.
9. Bankinn er ókeypis fyrir fyrirtæki.
10. Bankinn notar staðlað frjálst notandaleyfi ((L)GPL – http:// www.gnu.org/ licenses/), CC – http://creativecommons.org/, o.s.frv.).

Bankinn er því algerlega opinn og ókeypis, og hver sem er getur gert hvað sem er með hann. Þetta er í góðu samræmi við íslenska málstefnu sem samþykkt var á Alþingi fyrir þremur árum (*Íslenska til alls* 2009), þar sem segir að stefnt skuli að því „[a]ð málleg gagnasöfn og hugbúnaður til að vinna með íslenskt mál verði gerð opin og frjáls eftir því sem kostur er (sbr. stefnu stjórnvalda um frjálsan og opinn hugbúnað)". Í þeirri stefnu segir m.a.: „Stefnt skal að því að hugbúnaður sem smíðaður er og fjármagnaður af opinberum aðilum, m.a. í rannsóknar- og þróunarverkefnum, verði endurnýtanlegur. Liður í því er að hugbúnaðurinn sé frjáls" (Forsætisráðuneytið 2007).

Hægt er að sækja bankann í heild, ásamt leitarhugbúnaði fyrir Windows og Linux, á slóðinni http://www.linguist.is/icelandic_treebank/Download. Notandinn setur bankann upp á eigin tölvu og er því ekki háður netsambandi við notkun hans. Þótt við teljum að gerð trjábankans sé lokið er núverandi gerð hans tölusett 0.9 en ekki 1.0. Ástæðan er sú að við vinnum enn að ýmsum leiðréttingum og samræmingu sem við viljum ljúka áður en við setjum endapunktinn aftan við verkið.

Bankinn hefur einnig verið settur upp í INESS-trjábankamiðstöðinni við Háskólann í Bergen (http://iness.uib.no). Þar er hægt að leita í honum og skoða hríslumyndir af setningum, sem getur verið hentugt því að úttakið úr *CorpusSearch* er ekki sérlega þægilegt aflestrar. Setningin sem sýnd var hér að framan birtist í INESS eins og sýnt er á Mynd 2. Hins vegar sýnir INESS ekki heildarfjölda dæma um það sem leitað er að og nýtist því ekki við tölfræðilega úrvinnslu.



Mynd 2: Setning úr Sögulega íslenska trjábankanum í INESS-miðstöðinni.

Meðan á vinnslu trjábankans stóð gengum við frá nýrri útgáfu á netinu á þriggja mánaða fresti – fyrst í byrjun júlí 2010. Tilgangurinn var sá að kynna verkið frá byrjun, koma því í gagnið og fá viðbrögð notenda. Þetta tókst vel – þannig var útgáfu 0.4 sem var gefin út í apríl 2011 og hafði að geyma 440 þúsund lesmálsorð hlaðið niður meira en 450 sinnum. Útgáfu 0.9 hefur verið hlaðið niður um 660 sinnum á tæpu ári.

Þetta skilaði sér í því að trjábankinn hafði þegar verið nýttur í ýmsum rannsóknum áður en núverandi gerð hans kom í ágústlok 2011 (sjá t.d. Sapp 2011; Anton Karl Ingason, Einar Freyr Sigurðsson og Wallenberg 2011; Light og Wallenberg 2011; Maling, Kroch og Sigríður Sigurjónsdóttir 2011). Fleiri greinar um trjábankann og notkun hans hafa birst síðan eða eru væntanlegar (Eiríkur Rögnvaldsson o.fl. 2011, 2012; Anton Karl Ingason, Einar Freyr Sigurðsson og Wallenberg 2012; Jóhanna Barðdal og Þórhallur Eyþórsson 2012; o.fl.).

7. Lokaorð

Í þessari grein hefur verið sagt frá aðdragandanum að gerð *Sögulega íslenska trjábankans*, efniviði bankans og greiningaraðferð. Enn fremur var vinnunni við smíði bankans lýst og sýnt dæmi um hvernig hann má nota til athugana á setningafræðilegum breytingum. Eins og fram kom í inngangi teljum við að bankinn sé óvenjulegt verk fyrir ýmissa hluta sakir. Mikilvægast er að þar komu saman fræðimenn úr mismunandi greinum og með ólík áhugasvið, sem áttuðu sig á gildi þess að sameina krafta sína við gerð mikilvægs gagnasafns sem þjónar tvennum tilgangi.

Eins og fram kom í upphafi á bankinn annars vegar að vera tæki til rannsókna á íslenskri setningafræði, bæði samtímalegri og sögulegri, og hins vegar tól til þróunar ýmiss konar máltæknibúnaðar, s.s. villuleitarforrita, þýðingarforrita, leitarforrita o.fl. Hann hefur þegar sannað gildi sitt til setningafræðirannsókna eins og fram hefur komið, en enn sem komið er hefur hann ekki verið nýttur í síðarnefnda tilgangnum. Við erum þó ekki í vafa um að þar getur hann komið að góðu gagni. Í honum eru hátt í 300 þúsund lesmálsorð sem kalla má nútímamál – textar frá 19., 20. og 21. öld – og það er meira en nóg til að þjálfa tölfræðilegan þáttara til vélrænnar setningagreiningar.

Að lokum leggjum við áherslu á mikilvægi þess að trjábankinn sé algerlega opinn og ókeypis og dreift með opnum leyfum. Þetta er sérstaklega mikilvægt í tungumálum eins og íslensku vegna þess að uppbygging málfanga er dýr en mannafli og fé af skornum skammti, og því ber að forðast tvíverknað. Við vonum að aðrir fræðimenn fylgi þessu fordæmi og opni aðgang að gögnum sínum og niðurstöðum eftir því sem kostur er. Það er allra hagur.

HEIMILDIR

Anton Karl Ingason, Einar Freyr Sigurðsson og Joel Wallenberg. 2011. „Distinguishing Change and Stability: a Quantitative Study of Icelandic Oblique Subjects.“ Erindi flutt á DiGS 13, University of Pennsylvania, Philadelphia, 3. júní.

Anton Karl Ingason, Einar Freyr Sigurðsson og Joel Wallenberg. 2012. „Antisocial Syntax. Disentangling the Icelandic VO/OV Parameter and its Lexical Remains.“ Erindi flutt á DiGS 14, Lissabon, 6. júlí.

Anton Karl Ingason, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson og Joel C. Wallenberg. 2012. *Faroese Parsed Historical Corpus (FarPaHC).* Version 0.1. <http://linguist.is/farpahc>

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

Beck, Jana E. 2011. *Penn Parsed Corpora of Historical Greek (PPCHiG). <http://www.ling.upenn.edu/~janabeck/greek-corpora.html>*

Beck, Jana E., Aaron Ecay og Anton Karl Ingason. 2011. *Annotald.* Version 11.11. *<http://github.com/janabeck/Annotald>*

Eiríkur Rögnvaldsson. 2005. „Setningafræðilegar breytingar í íslensku." *Setningar. Handbók um setningafræði*. Ritstj. Höskuldur Þráinsson. Íslensk tunga 3. Reykjavík: Almenna bókafélagið, 602–635.

Eiríkur Rögnvaldsson, Anton Karl Ingason og Einar Freyr Sigurðsson. 2011. „Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC)." *Language Variation Infrastructure. Papers on selected projects*. Ritstj. Janne Bondi Johannessen. Oslo Studies in Language 3.2. Osló: University of Oslo, 97–111.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson og Joel Wallenberg. 2011. „Creating a Dual-Purpose Treebank." Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics* 26:2: 141–152.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson og Joel Wallenberg. 2012. „The Icelandic Parsed Historical Corpus (IcePaHC)." *Proceedings of LREC 2012*. Istanbúl: ELRA, 1977–1984.

Forsætisráðuneytið. 2007. *Frjáls og opinn hugbúnaður. Stefna stjórnvalda.* Desember 2007. <http://www.ut.is/media/verkefnisstjorn-radstefna-rafraen-framtid/Frjals _og_opinn_hugbunadur_-_Stefna_stjornvalda. pdf>

Galves, Charlotte, og Pablo Faria. 2010. *Tycho Brahe Parsed Corpus of Historical Portuguese.*<http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>

Hajič, Jan. 2005. „Complex Corpus Annotation: The Prague Dependency Treebank." *Insight into Slovak and Czech Linguistics*. Ritstj. Mária Šimková. Bratislava: Veda, 54–73.

Haraldur Bernharðsson. 1999. *Málblöndun í sautjándu aldar uppskriftum íslenskra miðaldahandrita.* Reykjavík: Málvísindastofnun Háskólans.

Haug, Dag Trygve Truslew og Marius Jøhndal. 2008. „Creating a Parallel Treebank of the Old Indo-European Bible Translations." *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Marrakech, 27–34.

Hrafn Loftsson. 2008. „Tagging Icelandic text: A linguistic rule-based approach." *Nordic Journal of Linguistics* 31:1: 47–72.

Hrafn Loftsson og Eiríkur Rögnvaldsson. 2007. „IceParser: An Incremental Finite-State Parser for Icelandic." *NODALIDA 2007 Conference Proceedings*. Ritstj. Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek og Mare Koit. Tartu: University of Tartu, 128–135.

Íslenska til alls. 2009. Tillögur íslenskrar málnefndar að íslenskri málstefnu samþykktar á Alþingi 12. mars 2009. Reykjavík: Mennta- og menningarmálaráðuneytið.

Jóhanna Barðdal og Þórhallur Eyþórsson. 2012. „Hungering and Lusting for Women and Fleshly Delicacies: Reconstructing Grammatical Relations for Proto-Germanic." Væntanlegt í *Variation and Change in Argument Realization*, *Transactions of the Philological Society* 110:3.

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

Kjartan Ottosson. 2001. „Kven sitt språk ser vi i avskrifter? – eller Éloge de la variante grammaticale." Erindi flutt á ráðstefnunni Nordiske middelaldertekster: Utgivere og brukere. Senter for høyere studier & Det Norske Videnskaps-Akademi, Osló, 28. apríl.

Kroch, Anthony, Beatrice Santorini og Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME).* Philadelphia: Department of Linguistics, University of Pennsylvania. CD-ROM. <*http://www.ling.upenn. edu/hist-corpora/*>

Kroch, Anthony, og Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2).* Philadelphia: Department of Linguistics, University of Pennsylvania. CD-ROM. 2. útg. <*http://www.ling.upenn.edu/hist-corpora/*>

Light, Caitlin. 2010. *Parsed Corpus of Early New High German. <http://enhgcorpus. wikispaces.com/home>*

Light, Caitlin og Joel Wallenberg. 2011. „On the Use of Passives across Germanic." Erindi flutt á DiGS 13, University of Pennsylvania, Philadelphia, 4. júní.

Maling, Joan, Anthony Kroch og Sigríður Sigurjónsdóttir. 2011. „The Icelandic Challenge: a System-Internal Syntactic Change." Erindi flutt á Comparative Germanic Syntax and the Challenge from Icelandic. Workshop at 33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Georg-August-Universität Göttingen, 24. febrúar.

Marcus, Mitch P., Beatrice Santorini og Mary Ann Marcinkiewicz. 1993. „Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics* 19: 313–330.

Martineu, France, Paul Hirschbühler, Anthony Kroch og Yves Charles Morin. 2010. *Corpus MCVF (parsed corpus), Modéliser le changement: les voies du français*. University of Ottawa. CD-ROM. <*http://www.arts.uottawa.ca/voies/voies _fr.html*>

Randall, Beth. 2005. *CorpusSearch 2 Users Guide.* Philadelphia: University of Pennsylvania.<*http://corpussearch.sourceforge.net/CS-manual/Contents.html*>

Santorini, Beatrice. 2010. *Annotation manual for the Penn historical corpora and the PCEEC.* Philadelphia: University of Pennsylvania. <*http://www.ling.upenn.edu/hist-corpora/annotation/index.html*>

Sapp, Christopher D. 2011. „A Relative Pronoun in Old Norse?" Erindi flutt á DiGS 13, University of Pennsylvania, Philadelphia, 5. júní.

Wallenberg, Joel, Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. <*http://www.linguist.is/icelandic_treebank*>

SUMMARY

The Icelandic Parsed Historical Corpus

Keywords: Treebank, parsing, historical corpus, diachronic syntax, language technology.

The article describes the background for and construction of *Icelandic Parsed Historical Corpus, IcePaHC*, a million word parsed historical corpus of Icelandic that has just been completed (Wallenberg et al. 2011). This corpus contains fragments of 60 texts ranging from the late twelfth century to the present day and serves the dual purpose of being both a cornerstone of Icelandic language technology and also an invaluable tool in Icelandic diachronic syntax research.

The corpus is unusual in many ways. First, it was designed to serve as a tool for both language technology and syntactic research, and was developed by scholars with research experience in both diachronic syntax and computational linguistics. Secondly, the corpus spans almost ten centuries – the oldest texts were written in the final decades of the twelfth century and the youngest are from the first decade of the present century. Thirdly, the corpus contains over one million words and is thus among the largest of the parsed corpora that have been published for any language. Fourthly, access to the corpus is completely open and free and thus requires no registration or paperwork, and the same is true for all the software used in its construction and also for other software developed within the project.

In the present paper, we follow the Introduction by describing the background to the treebank, whose origins lie in three different projects. Several aspects of the material in the treebank are then discussed – the selection of texts, their quality, and their conversion to modern Icelandic spelling. We then explain our decision to build a Penn style treebank and we offer an overview of the annotation process. Following a case study which shows how the treebank can be used to investigate

historical differences in use of the numeral/indefinite pronoun *einn* "one, a", we present our open source policy and set out "10 basic types of user freedom" for language resources.

The corpus has been made available via free download (*http://linguist.is/icelandic_treebank/Download*). Both the software and the corpus itself are distributed under the LGPL license (http://www.gnu.org/licenses/lgpl.html).

Eiríkur Rögnvaldsson
Háskóla Íslands
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is

Anton Karl Ingason
University of Pennsylvania
619 Williams Hall
Philadelphia, PA 19104-6305
ingason@ling.upenn.edu

Einar Freyr Sigurðsson
Háskóla Íslands
Árnagarði við Suðurgötu
IS-101 Reykjavík
einarfs@gmail.com

Joel C. Wallenberg
Newcastle University
Percy Building
Newcastle Upon Tyne, NE1 7RU
joel.wallenberg@gmail.com