EIRÍKUR RÖGNVALDSSON

# Tungumál, tölvur og tungutækni

## 0. Inngangur

Tengsl tölva og tungumáls má rekja alveg aftur til fyrstu ára tölvunnar um miðja 20. öld. Menn áttuðu sig snemma á því að það væri hægt að hafa gagn af tölvum við ýmiss konar málfræðilegar rannsóknir. Það á ekki síst við um ýmiss konar talningar, sem tölvur hafa alltaf verið góðar í – jafnvel þær fyrstu og frumstæðustu (Gazdar & Mellish 1989:3). Fljótlega var farið að nota tölvur til að gera ýmiss konar orðaskrár, skoða tíðni orða í mismunandi textum o.s.frv. Þannig var t.d. talsvert gert að því að leita höfunda texta eða skoða áhrif eins höfundar á annan með því að bera saman orðaforða þeirra og orðtíðni (Butler 1985:17-24).

Skemmtilegt dæmi um þetta er að finna í skáldsögunni *Lítill heimur* (*Small World*) eftir David Lodge (1996:58), í íslenskri þýðingu Sverris Hólmarssonar. Þar ræða þeir saman Persse McGarrigle lektor frá Limerick og Robin Dempsey prófessor frá Darlington, en samtalið á að fara fram vorið 1979:

> „Ja, rannsóknarverkefnið mitt var Shakespeare og T.S. Eliot," sagði Persse.
>
> „Þar hefði ég getað orðið þér að liði," greip Dempsey frammí. Hann var nýkominn á barinn ásamt Angelicu, sem var undrafögur í skósíðum serk úr vínrauðri bómull ofinni daufu mynstri af öðrum litum. „Þar hefði tölvuvinnsla einmitt verið vel við hæfi," hélt Dempsey áfram. „Þú hefðir ekki þurft annað en koma textanum á tölvutækt form og þá hefðirðu getað fengið tölvuna til að gera skrá yfir hvert einasta orð, orðasamband og setningarbyggingu sem er að finna hjá báðum þessum höfundum. Þú hefðir getað reiknað nákvæmlega út áhrif Shakespeares á T.S. Eliot."

Prófessor Dempsey er reyndar ekki sérlega hátt skrifaður fræðimaður í bókinni (eyðir t.d. löngum stundum í samræður við forritið ELIZU, sjá t.d. Gazdar & Mellish 1989:64-66), en þessi ræða hans er þó líklega dæmigerð fyrir þá trú sem margir höfðu fyrir nokkrum áratugum á möguleikum tölvutækninnar í bókmenntarannsóknum. Slíkar rannsóknir eru vissulega enn stundaðar, en menn hafa þó fyrir löngu áttað sig á því að það er síður en svo einfalt að túlka niðurstöður úr talningum af þessu tagi.

Annað svið tungumálsins þar sem snemma var reynt að nýta tölvur voru þýðingar. Á 6. áratug 20. aldar og fram á þann 7. var miklu fé varið í tilraunir með tölvuþýðingar, einkum í Bandaríkjunum. Mikið af þeim tilraunum var gert í hernaðarlegum tilgangi og kostað af bandaríska hernum. Fyrstu forritin þýddu texta í meginatriðum orð fyrir orð, og lítið var um að stuðst væri við málfræðilegar kenningar eða líkön. En þótt reynt hefði verið að nota slíkar kenningar hefði það komið að litlu gagni vegna þess að tölvur þeirra tíma hefðu tæpast getað hagnýtt þær (Gazdar & Mellish 1989:3).

Árið 1966 birti bandaríska vísindaakademían (National Academy of Sciences) „svarta skýrslu" um tölvuþýðingar í Bandaríkjunum, þar sem fram kom að þrátt fyrir gífurlegan kostnað hefði árangurinn verið ákaflega lítill. Eftir þetta var sáralítið fé veitt til tölvuþýðinga, í Bandaríkjunum a.m.k., og það var ekki fyrr en eftir 1980 sem aftur fór að færast líf í þær samfara framförum í tölvutækni.

Á undanförnum 10-20 árum hefur orðið gífurleg útþensla í hvers kyns tölvunotkun í tengslum við tungumál, og **tungutækni** (e. language technology) orðið að mikilvægum iðnaði víðast á Vesturlöndum. Þessi þróun hefur þó ekki náð til Íslands, eins og m.a. er rakið í nýlegri skýrslu starfshóps sem gerði úttekt á íslenskri tungutækni (sjá *Tungutækni* 1999). Í framhaldi af þeirri skýrslu ákvað ríkisstjórnin haustið 2000 að verja verulegum fjármunum til að efla starfsemi á þessu sviði.

Markmið þessarar greinar er að lýsa því hvað tungutækni er, hver staða hennar er á Íslandi nú, og hverjar eru framtíðarhorfur í íslenskri tungutækni. Í fyrsta kafla eru helstu hugtök á þessu sviði skilgreind, og gerð stutt grein fyrir því hvernig tölvur hafa verið notaðar í íslenskum málrannsóknum. Í öðrum kafla er lýst helstu forsendum fyrir því að öflug íslensk tungutækni rísi upp á næstu árum. Þriðji kafli er svo lokaorð.

## 1. Samspil tölva og tungumáls

Áður en lengra er haldið er rétt að huga að merkingu ýmissa orða og hugtaka sem tengjast tölvunotkun í málrannsóknum (sjá líka Eirík Rögnvaldsson 2001). Ekki er mikil hefð fyrir íslensku orðafari á þessu sviði, og orðanotkun hefur verið nokkuð á reiki. Hér verða tekin fyrir hugtökin **máltölvun**, **tölvufræðileg málvísindi**, **gagnamálfræði** og **tungutækni**. Það fyrsttalda er nokkuð þekkt en hefur verið notað í víðri og óljósri merkingu. Síðastnefnda orðið er nýlegt, en hefur hlotið mikla útbreiðslu á skömmum tíma. Hin tvö eru svo nýyrði sem ekki hafa fengið fastan sess enn, og ekki er ljóst hvernig reiðir af.

### *1.1 Máltölvun, tölvufræðileg málvísindi, gagnamálfræði*

Elsta orðið um tölvunotkun við málrannsóknir er líklega **máltölvun**, sem Baldur Jónsson (1975) bjó til fyrir 30 árum. Þetta orð samsvarar einna helst því sem nefnist **linguistic computing** á ensku, eða **linguistic and literary computing**. Það á við hvers kyns notkun tölva við lausn mállegra verkefna. Þar getur verið um að ræða talningar orða og bókstafa, gerð tíðniskráa, orðstöðulykla, orðabókagerð o.s.frv. Þetta eru verkefni sem í sjálfu sér væri hugsanlegt að vinna án aðstoðar tölvunnar, en yrðu mörg hver aldrei unnin í höndunum vegna þess hversu umfangsmikil þau eru. Oft þarf litla tölvuþekkingu til að leysa þessi verk af hendi, heldur eru þau unnin með hjálp tilbúinna forrita eða forritapakka s.s. *WordSmith*, *WordCruncher* o.s.frv.

Með **tölvufræðilegum málvísindum**, eða **computational linguistics**, er aftur á móti frekar átt við það að matreiða tungumálið þannig handa tölvum að þær geti framkvæmt málfræðilega greiningu. Slík greining er aftur undirstaða þess að hægt sé að nota tölvur við vélrænar þýðingar, lemmun, talgreiningu o.fl. Þar er sem sé verið að semja nýja tegund mállýsingar, þar sem taka þarf tillit til ýmissa atriða sem ekki skipta máli þegar málfræðingar sjá um greininguna. Stundum kemur gerð hugbúnaðar inn í þetta, en venjulega eru þó sérstakir forritarar sem sjá um þá hlið mála. Það breytir ekki því að þeir sem semja mállýsinguna þurfa að hafa góða hugmynd um það hvernig tölvur vinna; hvað þær geta gert og hvað ekki.

Því sem á ensku heitir **corpus linguistics** og reynt hefur verið að þýða sem **gagnamálfræði** er oft stillt upp sem andstæðu fræðilegra málvísinda (e. theoretical linguistics), þar sem megináhersla er lögð á að setja fram kenningar og prófa þær síðan á tungumálinu sjálfu. Gagnamálfræðin byrjar hins vegar á því að skoða textana vandlega og setja fram lýsingu að þeirri skoðun lokinni. Viðfangsefni og markmið gagnamálfræði

og fræðilegra málvísinda eru að sumu leyti ólík, en að öðru leyti geta þessar greinar bætt hvor aðra upp (sjá Jurafsky & Martin 2000:10-14).

Í sjálfu sér er hægt að stunda gagnamálfræði án hjálpar tölvutækninnar, og ýmsar rannsóknir byggðar á stórum textasöfnum voru gerðar fyrir daga tölvunnar (sjá t.d. McEnery & Wilson 1996). En hæfileikar tölvunnar til að leita að gögnum, sækja þau, raða þeim og reikna út niðurstöður gera hana að gífurlega mikilvægu hjálpartæki við málrannsóknir, og því hvarflar vart að nokkrum lengur að stunda rannsóknir af þessu tagi án þess að nota tölvu. Á síðustu áratugum hafa verið að koma upp risastór tölvutæk textasöfn samfara gífurlegum framförum í tölvutækni, og þetta hefur hvorttveggja stuðlað að mikilli uppsveiflu í gagnamálfræði. Þetta tengist svo uppgangi tungutækninnar, sem fjallað er um í næsta kafla.

### *2.2 Tungutækni*

Það orð sem nú heyrist mest þegar rætt er um tölvur og tungumál er **tungutækni**, sem vikið var að í inngangi og er þýðing á **language technology** (eða **language engineering**, sem einnig hafði verið þýtt sem **tungumálaverkfræði**). Þetta er ekki gamalt orð í íslensku; mun fyrst hafa sést á prenti fyrir 4-5 árum. Það vísar til hvers kyns samvinnu tungumáls og tölvutækni, að einu skilyrði uppfylltu: Samvinnan verður að hafa einhvern hagnýtan tilgang, beinast að því að hanna eða útbúa einhvern hugbúnað eða tæki sem nýtast mönnum í starfi eða leik.

Þessi samvinna er tvenns konar, eins og lýst er nánar hér á eftir, og felst annars vegar í notkun tölvutækninnar í þágu tungumálsins; hins vegar í notkun tungumálsins innan tölvutækninnar. Undir merkjum tungutækni eru stundaðar margvíslegar rannsóknir í gagnamálfræði, tölvufræðilegum málvísindum og tölvunarfræði, með það að markmiði að gagnast í ákveðnum iðnaði. En tungutækni er líka oft hrein iðnaðarstarfsemi, sem nýtir sér fyrirliggjandi gagnasöfn og rannsóknaniðurstöður tölvufræðilegra málvísinda og gagnamálfræði við smíði hvers kyns forrita og tóla.

Það er hægt að nýta tölvu- og upplýsingatækni á ýmsan hátt til þess að auðvelda mönnum að nota tungumálið. Þar má nefna ýmiss konar hugbúnað til að leiðrétta og leiðbeina um stafsetningu og málfar. Slíkur búnaður fylgir t.d. algengum forritapökkum eins og *Microsoft Office* á ýmsum tungumálum. Íslensk stafsetningarleiðréttingarforrit eru til, einkum *Púki* Friðriks Skúlasonar, en ekkert málfræðileiðréttingarforrit er til fyrir íslensku. Vélrænar þýðingar af einu máli á annað falla undir þetta. Einhverjar tilraunir hafa verið gerðar til að láta tölvur þýða texta milli íslensku og annarra mála (sjá Stefán Briem 1998, 1991), en ekkert nothæft forrit af því tagi er til, a.m.k. ekki á markaði.

Þá má nefna ýmiss konar tölvuorðabækur. Nokkrar slíkar hafa verið til milli íslensku og erlendra mála, en engin sérlega fullkomin. Nú er *Íslensk orðabók* (2000) þó komin á rafrænu formi. Hér má einnig telja ýmiss konar hjálpartæki handa þeim sem eiga erfitt með mál eða lestur sökum einhvers konar fötlunar. Til er íslenskur talgervill, þ.e. hugbúnaður sem gerir tölvu kleift að lesa upp ritaðan (tölvutækan) texta (sjá Pétur Helgason 1990; Carlson o.fl. 1990). Hann var upphaflega gerður um 1990, en hefur nýlega verið endurbættur út frá annarri tækni en upphaflega var notuð. Að endingu má nefna hér ýmiss konar kennsluforrit til að þjálfa notkun tungumálsins. Þótt vísir að sumum þeim tegundum hugbúnaðar sem hér hafa verið nefndar sé til hér á landi er það ekki nóg; þróunin á þessu sviði er mjög ör, og við drögumst sífellt lengra aftur úr.

En tungumálið er ekki bara þiggjandi í þessari samvinnu við tölvutæknina. Það er líka notað á margvíslegan hátt til að gera tæknina aðgengilegri og auðvelda mönnum að nýta sér hana. Þar má í fyrsta lagi nefna notkun tungumálsins við leit á netinu og í ýmiss konar gagnabönkum. Í stað þess að bera spurningar fram á staðlaðan hátt, og nota takmarkaðan orðaforða, er nú víða hægt að spyrja á venjulegu máli, rétt eins og maður talar við mann. Þetta er að verða algengt víða erlendis þar sem svara þarf mörgum svipuðum fyrirspurnum á ákveðnu sviði, t.d. í sambandi við flug- og lestarsamgöngur. Þar eru nýtt svonefnd **samræðukerfi** (e. dialogue systems) þar sem maður og tölva ræða saman á mannamáli, ýmist rituðu eða töluðu. Mikið er nú lagt í rannsóknir og þróun á þessu sviði (sjá t.d. Jurafsky & Martin 2000:719-721; Stork 1996).

Í öðru lagi má nefna notkun málsins við stjórn tölva og ýmiss konar tölvustýrðra tækja. Það fer mjög í vöxt að slíkum tækjum sé stjórnað með venjulegu máli, annaðhvort rituðu eða töluðu. Skipanir eru þá ýmist slegnar inn á lyklaborð eða talaðar í hljóðnema, í stað þess að ýtt sé á þartilgerða takka. Þetta mun á næstunni taka til sífellt fjölbreyttari tækja, s.s. ýmiss konar framleiðslutækja, heimilistækja og bíla. En engin slík tæki skilja íslensku – enn sem komið er a.m.k. Til að svo megi verða þarf m.a. að leggja mikla vinnu og fé í að gera nákvæma íslenska hljóðlýsingu og hljóðgreiningu. Fyrirtækið Voice Era vinnur nú að slíkri greiningu, og verður fróðlegt að sjá hvort og hvernig sú vinna skilar sér í nothæfum samræðukerfum og öðrum tungutæknitólum.

### *1.3 Tölvunotkun í íslenskum málrannsóknum*

Á Íslandi hafa verið gerðar nokkrar rannsóknir sem telja má til máltölvunar. Fyrsta rannsókn á íslensku máli sem unnin var í tölvu var könnun Baldurs Jónssonar og samstarfsmanna hans á tíðni orða í *Hreiðrinu* eftir Ólaf Jóhann Sigurðsson, og í framhaldi af því ólemmaður orðstöðulykill[1] yfir sama texta (sjá Baldur Jónsson 1975, 1978; Baldur Jónsson, Björn Þ. Ellertsson og Sven Þ. Sigurðsson 1980). Langsamlega viðamesta og vandaðasta rannsóknin á þessu sviði er sú tíðnikönnun á íslenskum textum sem unnið var að hjá Orðabók Háskólans á árunum 1986-1991. Niðurstöður hennar birtust í *Íslenskri orðtíðnibók* (1991) eftir Jörgen Pind (ritstjóra), Friðrik Magnússon og Stefán Briem; sjá einnig Friðrik Magnússon (1988).

Um miðjan síðasta áratug komu svo út tveir viðamiklir orðstöðulyklar sem báðir voru að nokkru leyti tölvuunnir og höfðu verið í vinnslu um nokkurra ára skeið. Annar var *Biblíulykill*, sem gengur að útgáfunni frá 1981. Þetta var samstarfsverkefni fimm aðila; Íslenskrar málstöðvar, Orðabókar Háskólans, Málvísindastofnunar Háskólans, Guðfræðistofnunar Háskólans og Baldurs Pálssonar forritara. Biblíulykillinn kom út árið 1994, og var fyrsti lemmaði og tölvuunni orðstöðulykillinn að texta á íslensku sem prentaður hefur verið.[2]

---

[1] Orðstöðulykill tiltekins texta er skrá þar sem öll dæmi um sérhverja orðmynd textans eru sýnd í samhengi, þ.e. með nokkrum næstu orðum á undan og eftir, þannig að hvert dæmi fær sérstaka línu. Lemmun texta felst í því að flokka saman þær myndir sem tilheyra sama flettiorði, og greina sundur samhljóma orð sem tilheyra mismunandi flettiorðum. Þannig eru færðar saman myndir eins og *á*, *eiga*, *ætti* o.fl. af so. *eiga*, en dæmi um orðmyndina *á* greind sundur eftir því hvaða flettiorði þau tilheyra (fs. *á*, so. *eiga*, no. *á*, no. *ær*). Í ólemmuðum orðstöðulykli hefur slík greining ekki farið fram.

[2] Áður höfðu verið gefnir út lemmaðir orðstöðulyklar sem voru algerlega handunnir; *Orðalykill að Nýja testamentinu* (Björn Magnússon 1951) og *A Concordance to Eddic Poetry* (Kellogg 1988).

En líklega jafnframt sá síðasti. Það eru sennilega úrelt vinnubrögð að gefa skrár af þessu tagi út á prenti, því að rafrænar útgáfur eru mun ódýrari og sveigjanlegri í alla staði. Fyrsti lykillinn sem birtist á því formi var *Orðstöðulykill Íslendinga sagna,* sem gefinn var út á geisladiski 1996 og er viðamesti orðstöðulykill sem hér hefur verið gerður. Ritstjórar hans eru Bergljót S. Kristjánsdóttir, Eiríkur Rögnvaldsson, Guðrún Ingólfsdóttir og Örnólfur Thorsson. Nú er svo unnið að gerð orðstöðulykils yfir skáldsögur Halldórs Laxness, í samvinnu Orðabókar Háskólans og Vöku-Helgafells.

Aðferðum máltölvunar hefur lítið verið beitt við rannsóknir á íslenskum bókmenntum. Þó má nefna athuganir Örnólfs Thorssonar (1993, 1994) á orðaforða *Grettis sögu* og annarra íslenskra fornsagna.

## 2. Íslensk tungutækni

Eins og vikið var að í inngangi er íslensk tungutækni skammt á veg komin, og sárafá tæki og tól eru til þar sem íslenskt mál og tölvutækni vinna saman. Þróunarstarf á þessu sviði er dýrt, og smæð hins íslenska markaðar gerir það að verkum að ekki hefur verið fýsilegt fyrir hugbúnaðarfyrirtæki að leggja í slíkan kostnað (sjá *Tungutækni* 1999:25). Sú staða kann þó að vera að breytast, eftir að ákveðið var að veita ríkisstyrki til uppbyggingar á þessu sviði.

En þetta fé, þótt nauðsynlegt sé, leysir ekki sjálfkrafa allan vanda. Frumforsendur fyrir því að til verði íslensk tungutækni eru þrjár; menntað fólk, málsöfn og málgreiningarforrit (sjá einnig *Tungutækni* 1999). Hér á eftir verður fjallað um þessar þrjár forsendur og rökstutt hvers vegna framtíð íslenskrar tungutækni hvílir á þeim.

### *2.1 Tungutækni í háskólamenntun*

„Eitt stærsta vandamál sem Íslendingar standa frammi fyrir ef þeir ætla að hefja öflugt og markvisst starf á sviði tungutækni er skortur á fólki með menntun, reynslu og þekkingu á þessu sviði“ segir í skýrslu starfshóps um tungutækni (*Tungutækni* 1999:29). Hér er mikilvægt að hafa í huga að enda þótt skammt sé síðan tungutækni varð að iðnaði í grannlöndum okkar styðst það starf við margra ára rannsóknir og kennslu á háskólastigi. Í enskumælandi löndum er **Computational Linguistics** víða sérstök grein í háskólum, en einnig sums staðar innan málvísindadeilda eða tölvunarfræðideilda. Svipuðu máli gegnir um **Komputerlinguistik** í þýskumælandi löndum, **datalingvistik** á Norðurlöndum o.s.frv.

Mikill vöxtur hefur verið í þessum greinum á undanförnum árum. Vitaskuld er beint samband milli þenslunnar á þessi svið í háskólum og hinnar öru þróunar sem hefur verið í tungutækni sem iðngrein sem veltir háum fjárhæðum. Þörf atvinnulífsins fyrir fólk með menntun á þessu sviði hefur stóraukist, og þar með vilji yfirvalda til að efla slíka kennslu á háskólastigi. En um leið hafa áherslur í kennslunni breyst. Í stað þess að þessi fræði séu fyrst og fremst stunduð sem hefðbundin akademísk grein hefur áherslan færst yfir á hagnýtingu þeirra í ýmiss konar tækjum og tólum; leiðréttingarforritum, vélrænum þýðingum, hjálpartækjum fyrir fatlaða o.s.frv.

Þessar breyttu áherslur endurspeglast í þeim heitum sem nú eru höfð yfir kennslugreinar á þessu sviði í mörgum háskólum grannlandanna. Við hliðina á Computational Linguistics, datalingvistik o.s.frv. er nú komið **Language Technology**, **sprogteknologi** o.þ.h. Þetta kom t.d. vel í ljós í fyrra þegar sænska ríkisstjórnin ákvað að koma á fót 16 nýjum brautum í rannsóknanámi, forskarskolar, sem eru samstarfsverkefni

sænskra háskóla. Ein þessara brauta er **Sveriges nationella forskarskola i språkteknologi –** ekki „forskarskola i datalingvistik“ (sjá http://www.gslt.hum.gu.se).

Hér á landi hefur lítið sem ekkert verið fengist við kennslu og rannsóknir á þessu sviði. Innan íslensku og almennra málvísinda við Háskóla Íslands hafa einstöku sinnum verið haldin námskeið um tölvur og tungumál, en því fer fjarri að þar hafi verið um skipulagðar námsleiðir að ræða. Enginn íslenskur málfræðingur fæst nú svo að heitið geti við rannsóknir á sviði tungutækni. Þess eru fáein dæmi að íslenskir stúdentar í tölvunarfræði og rafmagnsverkfræði hafi unnið lokaverkefni sem falla undir tungutækni (sjá t.d. Þorgeir Sigurðsson 1981; Kjartan R. Guðmundsson 1986; Trausti Þór Kristjánsson 1995), en ekki er vitað til að neinn þeirra hafi haldið áfram námi eða rannsóknum á því sviði. Í skýrslu starfshóps um tungutækni sagði um þetta (*Tungutækni* 1999:31):

> Óráðlegt er að ætla að Íslendingar geti byggt upp öflugt starf á sviði tungutækni án þess að hyggja að fræðilegum undirstöðum slíks starfs. Nauðsynlegt er að fá sem fyrst til starfa vel menntað fólk á sviði íslensks máls og tölvunarfræði sem gerir sér grein fyrir sérkennum íslenskrar málfræði og íslensks málsamfélags. Ef ekki verður byggð upp innlend þekking á þessu sviði innan menntastofnana verðum við um ófyrirsjáanlega framtíð þiggjendur á þessu sviði og höfum miklu minni möguleika á að bregðast við breyttum aðstæðum og nýjungum, og þróa þau tól og tæki sem henta best íslenskum aðstæðum.

Starfshópurinn lagði því til í skýrslu sinni að komið yrði upp þverfaglegu meistaranámi á þessu sviði við Háskóla Íslands. Sú tillaga kemst væntanlega til framkvæmda haustið 2002, í framhaldi af samningi um þjónustu á sviði tungutækni sem menntamálaráðuneytið gerði við Háskóla Íslands sumarið 2001. Þar er gert ráð fyrir að taka inn nemendur með bæði málfræðilegan og tölvufræðilegan bakgrunn. Þess verður að vænta að þannig skapist frjó samvinna þessara tveggja sviða og til verði öflug sveit fólks sem komi til starfa innan íslenskrar tungutækni og byggi hana upp.

### *2.2 Málsöfn*

Önnur forsenda fyrir þróun íslenskrar tungutækni er gerð ákveðinna málgrunna sem nefndir hafa verið **tungutæknieiningar**. Þetta eru gagnasöfn og greiningartæki sem nýtt eru sem hráefni í tungutæknitól. Langflest verkefni innan tungutækni byggjast á einhvers konar míllegum gagnasöfnum. Hráefnið í þessum söfnum getur verið af ýmsu tagi og það getur verið flokkað og greint á ýmsa vegu. Þegar um er að ræða tól til afmarkaðra nota getur einfalt gagnasafn dugað. Í undirstöðu að rafrænni rímorðabók er t.d. meginatriði að hafa upplýsingar um rímatkvæði og atkvæðafjölda; orðflokkur og merking skiptir minna máli. Æskilegast er þó að til verði vönduð grundvallarsöfn sem séu óháð ákveðinni nýtingu, heldur geti nýst í hvers kyns tungutæknitólum. Þrenns konar söfn skipta mestu máli; **orðasöfn**, **textasöfn** og **hljóðsöfn**.

#### 2.2.1 Orðasöfn

Þau orðasöfn sem hér um ræðir eru í grundvallaratriðum svipuð hefðbundnum orðabókum á rafrænu formi, en munurinn er einkum tvenns konar. Í orðasöfnum til nota í tungutækni þurfa upplýsingarnar að vera miklu ítarlegri og fjölbreyttari en í venjulegum orðabókum. Þar þurfa að vera upplýsingar um stafsetningu, orðflokk, beygingu og merkingu orðanna, en einnig um setningafræðilega eiginleika þeirra og tengsl við önnur

orð í setningu (t.d. fallstjórn sagna). Einnig þurfa að vera þarna upplýsingar um orðastæður (collocations),[3] orðtök og málshætti; stílgildi orða; o.s.frv. Í orðasöfnum af þessu tagi er líka nauðsynlegt að setja allar upplýsingar fram á staðlaðan og samræmdan hátt til að tölvur geti unnið með þær. Því getur þurft að koma upp nákvæmu og flóknu flokkunarkerfi fyrir hinar ýmsu tegundir upplýsinga sem þarna verða að vera.[4]

Orðasöfn af þessu tagi hafa víða verið byggð upp eða eru í smíðum. Evrópusambandið hefur t.d. fjármagnað stór verkefni á þessu sviði, s.s. PAROLE og LE-PAROLE (sjá http://www.ub.es/gilcub/SIMPLE/simple. html). Í Danmörku er nú unnið að stóru orðasafni til nota í tungutækni, STO, eða SprogTeknologisk Ordbase (sjá Braasch o.fl. 1998). Þetta er samvinnuverkefni ýmissa stofnana, stjórnað af Center for sprogteknologi í Kaupmannahöfn. Í þessu safni eiga að vera u.þ.b. 50 þúsund orð, þar af um 35 þúsund úr almennu máli og 15 þúsund íðorð af sex mismunandi sviðum. Merkingareiningar (semantiske enheder) verða aftur á móti helmingi fleiri, eða um 100 þúsund.[5] Lögð er áhersla á hinn setningafræðilega þátt safnsins, þ.e. upplýsingar um innbyrðis tengsl orðanna. Sá þáttur skiptir mjög miklu máli fyrir hagnýtingu á ýmsum sviðum tungutækni, s.s. í forritum til málfræðileiðréttingar og þýðingarforritum.

Það er fljótsagt að ekkert íslenskt orðasafn af þessu tagi er til. Vissulega má segja að bæði *Íslensk orðabók* (2000) og íslenskur orðabókargrunnur Orðabókar Háskólans („norræni stofninn“ svokallaði) gætu orðið upphaf að slíku safni. Bæði söfnin skortir þó þá tvo meginþætti sem áður voru nefndir og greina tungutæknisöfn frá venjulegum orðabókum; upplýsingarnar í þeim eru alltof takmarkaðar, og framsetning þeirra er hvorki nógu formleg né stöðluð. Þar að auki er orðaforði beggja safnanna of ósamstæður og götóttur.

### 2.2.2 Málheildir

Í grannlöndum okkar eru víðast hvar til eða í uppbyggingu stórar **málheildir** (e. corpora, et. corpus). Með málheild er átt við textasafn sem er sett saman eftir ákveðnum reglum um t.d. efnisflokka, kyn og aldur höfunda o.s.frv., þannig að málheildin gefi sem besta mynd af því sem verið er að rannsaka. Málheild þarf því að greina frá **textasafni** (e.

---

[3] Með orðastæðum er átt við tvö eða fleiri orð sem standa stundum eða oftast saman – oftar en svo að hægt sé að líta á það sem tilviljun. „Orðastæðu má lauslega skilgreina sem samband orða sem mynda merkingarlega heild og koma iðulega fyrir sem samstæða innan setningar. Oftast nær skiptist orðastæða í tvo aðalliði þar sem annar er **kjarni** en hinn **ákvæði**“ (Jón Hilmar Jónsson 1994:xvi).

[4] Í lýsingu á STO-safninu danska sem sagt er frá hér á eftir segir: „En morfologisk enhed giver som minimum oplysning om opslagsordets stavning, bøjning, ordklasse og køn (disse suppleres løbende med orddannelsesoplysninger); en syntaktisk enhed indeholder oplysninger om opslagsordets konstruktions-potentiale (funktionel og kategoriel valens, mm.), syntaktiske funktion i konstruktioner, brug af hjælpe-verbum osv. Endelig indeholder en semantisk enhed som minimum oplysning om domæne. […] Oplysningerne udtrykkes i attribut/værdi-par der er formaliserede i koder; hver unik kombination af et sæt sammenhørende koder udgør et mønster“ (http://cst.dk/sto/beskrivelsesmodel/index.html).

[5] Hér má sjá dæmi úr setningafræðilegri lýsingu sagnarinnar *vente* í STO, þar sem upphafsstrengirnir eru „Mønstre der angiver de forskellige konstruktionstyper“ (http://cst.dk/sto/leksikalskindgang/index.html):

| | |
|---|---|
| Dv2Pntis-paa | (Han venter på svar; Han venter på at hun kommer) |
| Dv2Pnis-med | (Han venter med at gøre arbejdet) |
| Dv1f | (Hun venter sig) |
| Dv2N0 | (Forældrene ventede det værste) |
| Dv2t | (Jeg venter, at det kommer til at fungere) |
| Dv3fNP-af | (Han venter sig meget af den nye medarbejder) |

collection of texts), sem er tilviljanakennt samsafn texta, án þess að hugað hafi verið að neins konar hlutföllum. Meðal þekktra málheilda má nefna British National Corpus (BNC, sjá http://info.ox.ac.uk/bnc/) og hinn norska Nasjonalt korpus for språkteknologi. Það eru hvorttveggja 100 milljón orða söfn, sem eru samsett úr sem fjölbreyttustum textum í ákveðnum hlutföllum.[6]

Eina íslenska málheildin sem sett hefur verið saman í rannsóknaskyni er sú sem liggur til grundvallar *Íslenskri orðtíðnibók* (Jörgen Pind, Friðrik Magnússon & Stefán Briem 1991). Hún er ekki stór, aðeins 500 þúsund orð; sett saman úr 100 textabútum sem hver var u.þ.b. 5000 lesmálsorð. Þeir voru úr fimm textaflokkum; 1) íslensk skáldverk; 2) þýdd skáldverk; 3) ævisögur og endurminningar; 4) fræðslutextar; 5) barna- og unglingabækur. Í 4. flokki var helmingurinn af sviði hugvísinda og hinn helmingurinn úr raunvísindum og tækni; og í 5. flokki var helmingur textanna frumsaminn á íslensku, en hinn helmingurinn þýddur.

Hér eru vissulega margs konar textar, en þó er hæpið að segja að þessi málheild sé dæmigerð fyrir íslenskt mál í lok 20. aldar. Það vekur athygli að meirihluti textanna (60 af 100) eru skáldverk; en hins vegar vantar algerlega texta úr dagblöðum og tímaritum almenns efnis. Þá má nefna að aðeins rúmur fjórðungur textanna er frumsaminn eða þýddur af konum. Ekki er auðvelt að segja til um uppruna eða aldur höfunda, en þeir þættir geta líka skipt máli. [7] Það er sem sé ljóst að þarna eru ýmsar breytur sem hægt væri að hugsa sér að taka tillit til, og án efa yrðu niðurstöður um orðaforða og orðtíðni aðrar ef aðrir textar lægju að baki.

Síðast en ekki síst vantar talmál algerlega í þessa málheild. Sáralítið er í raun vitað um íslenskt talmál og að hvaða leyti það er frábrugðið ritmáli, t.d. í beygingum, setningagerð, orðavali o.s.frv. Þó má fullyrða, bæði út frá óformlegum athugunum og erlendum rannsóknum, að þarna sé talsverður munur á. Undanfarin tvö ár hefur verið unnið að söfnun til íslenskrar talmálsheildar sem nefnist *Íslenskur talmálsbanki*, eða *ÍS-TAL* (sjá http://www.hi.is/~eirikur/istal). Frumkvæði að þessu verki átti Þórunn Blöndal, lektor við Kennaraháskóla Íslands, en þátttakendur í því eru alls sjö frá Kennaraháskólanum, Háskóla Íslands og Orðabók Háskólans.

Markmið verkefnisins er að koma á fót gagnabanka með íslensku talmáli sem getur orðið grundvöllur rannsókna á einkennum talaðs máls og leitt í ljós muninn á talmáli og ritmáli. Með tilkomu þessa banka skapast í fyrsta sinn tækifæri til að rannsaka ýmsa þætti íslensks talmáls og bera það saman við ritmálið (sjá Ástu Svavarsdóttur 2001; Þórunni Blöndal 2001).

---

[6] Í tillögum undirbúningsnefndar norsku málheildarinnar var t.d. gert ráð fyrir eftirfarandi skiptingu (sjá http://www.tele.ntnu.no/users/svendsen/korpus/kortrapport.pdf):

1. Fjölmiðlaefni (20%); dagblöð, héraðsblöð, textavarp, texti við fréttir.
2. Fagurbókmenntir (25%); skáldsögur, smásögur, leikrit, texti við sjónvarpsþætti og kvikmyndir.
3. Nytjatextar (50%); fræðibækur af ýmsu tagi, tímarit og vikublöð, kennslubækur, uppflettibækur, stjórnvaldstextar.
4. Óprentað efni og smáprent (5%); auglýsingar, notendaleiðbeiningar, sölu- og kynningarefni; viðskiptatextar (minnisblöð, fundargerðir), bréfaskipti (bréf og tölvupóstur).

[7] Til samanburðar má nefna samsetningu textasafnsins sem liggur til grundvallar COBUILD orðabókinni ensku, sem sagt er frá í bókinni *Looking Up* (sjá Renouf 1987). Þar var m.a. gætt að hlutföllum í aldri, kyni og uppruna höfunda. Ákveðið var að hafa 75% textans eftir karla en 25% eftir konur; 70% breska ensku, 20% ameríska ensku og 5% önnur afbrigði málsins; og 75% ritmál, 25% talmál.

Stórar málheildir eru grundvallarforsenda fyrir þróun ýmissa tungutæknitóla. Undanfarinn áratug hafa aðferðir við þá þróun breyst verulega, samfara breytingum á tölvum og tölvutækni. Vinnslugeta og vinnsluhraði tölva hefur margfaldast og minni og diskpláss einnig. Jafnframt hefur tölvutækum textum af ýmsu tagi fjölgað gífurlega um leið og tölvutæknin tekur til sífellt fleiri þátta í daglegu lífi.

Þetta leiðir til þess að nú er raunhæft að vinna með geysistór textasöfn tiltölulega hrá. Í stað þess að vinna upp úr textasöfnunum söfn með grunnorðaforða, skrá upplýsingar um þann orðaforða og smíða síðan reglur út frá þeim er núna mögulegt að nota textasöfnin milliliðalaust og láta tölvuna vinna upplýsingar úr þeim jafnóðum. Um leið hafa tölfræðilíkön að nokkru leyti komið í stað málfræðireglna. Þessa sér m.a. stað í þýðingarforritum, en **hliðstæðar málheildir** (e. parallel corpora) eru mikið nýttar í vélrænum þýðingum. Þá er komið upp safni texta sem til eru á tveimur tungumálum, og þeir bornir saman til að reyna að finna mynstur í samsvöruninni sem síðan er hægt að nýta í þýðingum.

Málheildir eru einnig mjög þýðingarmiklar í orðabókarvinnslu, og nýjar orðabækur byggjast flestar á slíkum söfnum. Upplýsingar um notkun einstakra orða og stöðu þeirra í setningu skipta miklu máli fyrir orðabókarlýsinguna og auðvelda orðabókarmönnum að finna góð notkunardæmi. Það er líka mikilvægt að hafa upplýsingar um það t.d. að tiltekið nafnorð komi eingöngu fyrir í fleirtölu, tiltekin sögn sé aldrei notuð í nútíð, o.s.frv. Slíkar upplýsingar má auðveldlega fá í góðri málheild, en mjög erfitt er að afla þeirra á annan hátt. Fyrsta stóra orðabókin sem byggð var á málheild var enska COBUILD-orðabókin (1987) sem olli byltingu í orðabókagerð (sjá http://titania.cobuild.collins.co.uk/); hún þótti svo merkileg að gefin var út sérstök bók, *Looking Up*, um gerð hennar (Sinclair 1987).

### 2.2.3 Hljóðsöfn

Þriðja tegund almennra gagnasafna sem nauðsynleg er í tungutækni eru svo hljóðsöfn; upptökur af íslenskum orðum og málhljóðum ásamt margvíslegri greiningu. Slík söfn eru forsenda þess að hægt sé að þróa **talgreiningu** (e. speech analysis) og **raddþekkingu** (e. voice recognition), þ.e. hugbúnað sem greinir talað mál. Hugbúnað af því tagi þarf að þjálfa á stóru gagnasafni þar sem fyrir koma öll íslensk málhljóð í fjölbreyttu hljóðfræðilegu umhverfi. Í safninu þurfa einnig að koma fyrir sem flest orð, a.m.k. öll algengustu orð málsins. Þá þarf að gæta þess að í safninu séu dæmi um ólíkar raddir, bæði karl- og kvenraddir; dæmi um allar framburðarmállýskur; dæmi um mismunandi talhraða, mismunandi skýr framburður o.s.frv.

Grunnur að safni af þessu tagi er fyrir hendi í *Íslenskum talmálsbanka*, sem nefndur var hér að framan. Sá grunnur er hins vegar bæði lítill og fábreyttur og dugir sennilega skammt sem þjálfunarsafn í talgreiningu. Fyrirtækið Voice Era í Bolungarvík hefur nýlega komið upp stóru íslensku hljóðsafni, þar sem hundruð manna voru fengin til að lesa nokkra tugi setninga í síma. Þetta hefur verið notað við gerð íslensks raddgreinis, sem sagt er að þekki „öll íslensk orð – í u.þ.b. 80% tilvika“ (sjá http://www.eravoice.com/index.phtml?go=faq#3), en ekki er alveg ljóst hvernig á að túlka það.

## *2.3 Mörkun og markarar*

Sé stór málheild tiltæk má beita ýmsum tölfræðilegum aðferðum til að finna mynstur í málnotkun, og nota þau mynstur við gerð tungutæknitóla. Til að málheild komi að sem

bestum notum þarf hún þó helst að vera málfræðilega **mörkuð** (e. tagged), en með **mörkun** (e. tagging) er átt við það að merkja eindir í samfelldum texta á kerfisbundinn hátt. Eindirnar geta verið bókstafir, orð, setningarliðir, setningar o.fl. Merkingarnar geta líka verið af ýmsum toga. Þannig er t.d. hægt að hugsa sér að öll mannanöfn séu merkt á ákveðinn hátt, öll staðanöfn á annan hátt, öll erlend orð í textanum séu sérmerkt, o.s.frv. Til að marka texta þarf sérstakan hugbúnað, **markara** (e. tagger). Málfræðileg greining og mörkun er nauðsynleg í margvíslegum tungutæknitólum. Hér á eftir er gerð stutt grein fyrir því hvaða gildi slík greining hefur og hvernig hún fer fram.

### 2.3.1 Tilgangur málfræðilegrar mörkunar

Grundvallaratriðið í mörkun málfræðilegra upplýsinga er **orðflokksmörkun** (e. PoS tagging), þar sem orðflokksmerki er hengt á hvert orð, t.d. *Gamla*<lo> *konan*<no> *mætti*<so> *þessum*<fn> *tveim*<to> *drengjum*<no> *í*<fs> *morgun*<no>. Síðan er hægt að ganga lengra og bæta inn hvers kyns málfræðilegum upplýsingum, s.s. um kyn, tölu, fall, persónu, tíð, stig o.s.frv. Einnig má marka orð með upplýsingum um setningafræðileg hlutverk, s.s. frumlag, andlag, umsögn o.þ.h.

Meginhluti málfarsleiðréttinga er óhugsandi án málfræðilegrar greiningar. Aðeins lítill hluti málfarsvillna felst í því að notaðar séu orðmyndir sem ekki eiga að koma fyrir í málinu (t.d. *föður**s*** í stað *föður*, *k**ey**ptu* í stað *k**au**ptu*). Langflestar villur felast í því að nota leyfilegar orðmyndir á óleyfilegum stöðum í setningu. Villur eins og *Ég hitti syst**ir** þína* (í stað *syst**ur***), *vegna þeirrar tilhneiging**u*** (í stað *tilhneiging**ar***), *fjöldi manna kom**u*** (í stað *kom*), ***mér** langar* (í stað ***mig** langar*) o.s.frv. er ekki hægt að finna og leiðrétta á vélrænan hátt án málfræðilegrar greiningar, því að *systir*, *tilhneigingu*, *mér* og *komu* eru allt fullkomlega leyfilegar íslenskar orðmyndir – bara ekki á þessum stöðum í setningu.

Ýmsar algengar stafsetningarvillur eru líka þess eðlis að þær finnast ekki nema með málfræðilegri greiningu. Mörg orð í málinu eru t.d. ýmist skrifuð með einu eða tveimur *n*-um eftir setningafræðilegri stöðu; *morgunn*/ *morgun*, *Kristinn*/*Kristin*, *farinn*/*farin* o.s.frv. Hér eru báðar myndirnar leyfilegar, og villuleitarforrit sem eingöngu skoðar hverja orðmynd fyrir sig finnur því ekki villurnar í *það er kominn morgu**n**, ég hitti Kristi**nn**, hann er fari**n***.

Vélrænar þýðingar byggjast einnig á málfræðilegri greiningu. Án slíkrar greiningar getur vélræn þýðing ekki orðið annað en einföld uppfletting í orðasafni, þar sem orð úr einu máli er sett í stað orðs í öðru máli og ekkert hirt um reglur um orðaröð, beygingar og annað slíkt. Þá koma upp alþekkt dæmi eins og *hot spring river this book* (fyrir *hver á þessa bók*).

Málfræðileg mörkun fer venjulega fram í tveimur þrepum. Í fyrra þrepinu er orðum í textanum flett upp í orðasafni með beygingarlegum upplýsingum, og þær upplýsingar síðan færðar inn í textann. Þannig eiga t.d. að fást upplýsingar um að *í* sé forsetning, *hesturinn* sé nafnorð í karlkyni, eintölu, nefnifalli, með greini, og *fóruð* sé sögn í annarri persónu, fleirtölu, þátíð, framsöguhætti, germynd.

Slík uppfletting dugir hins vegar ekki til að greina öll orð í samfelldum texta á ótvíræðan hátt. Það kemur t.d. í ljós við uppflettinguna að þótt *í* sé einrætt orð er *á* ekki bara forsetning, heldur líka sögnin *eiga* í 1. og 3. persónu, eintölu, nútíð, framsöguhætti, germynd; kvenkynsnafnorðið *á* í eintölu, nefnifalli, þolfalli og þágufalli; kvenkynsnafnorðið *ær* í eintölu, þolfalli og þágufalli; og fleira mætti nefna. Þótt greiningin á *hestur* sé ótvíræð getur *hesta* verið bæði þolfall og eignarfall fleirtölu. Þótt *fóruð* sé

einrætt er *fórum* tvírætt; getur ekki einungis verið fyrsta persóna, fleirtala, þátíð, framsöguháttur, germynd af *fara*, heldur líka þágufall fleirtölu af *fórur* (sem reyndar kemur tæpast fyrir í eintölu).

### 2.3.2 Markarar

Til að greiða úr tví- og margræðni orðmynda þarf annað þrep í vinnslunni. Í því lagi er önnur eða ein greiningin valin, en hinni eða hinum hafnað. Forrit sem framkvæma slíkt val vinna á ýmsa vegu, en í grundvallaratriðum má segja að þau skiptist í tvo flokka; **tölfræðimarkara** (e. statistical/stochastic taggers) og **reglumarkara** (e. rule-based taggers) (sjá Jurafsky & Martin 2000:300-307).

Tölfræðimarkarar byggjast á upplýsingum um tíðni einstakra beygingarmynda til að velja líklegustu greininguna. Slíkur markari myndi greina *á* rétt í setningunni *Ég er á leiðinni*, vegna þess að *á* er mun oftar forsetning en nokkuð annað. Hins vegar yrði *á* ranglega greint í setningunni *Ég á þetta*; þar veldi tölfræðimarkarinn forsetningu eins og áður. Sömuleiðis yrði *fórum* trúlega greint ranglega sem sögn í sambandinu *í fórum mínum*, því að þessi orðmynd er mun algengari sem sagnmynd en sem nafnorðsmynd.

Reglumarkarar nota reglur um gerð setninga og setningarliða til að marka orðin. Þeir búa t.d. yfir upplýsingum um það að forsetning kemur sjaldan næst á undan sögn, og þess vegna er ólíklegt að orðið *fórum* sé sögn í sambandinu *í fórum mínum*, þótt svo gæti verið ef litið er á orðið eitt og sér. Reglumarkari ætti líka að búa yfir upplýsingum um það að þegar eignarfornafn stendur næst á eftir nafnorði sambeygjast orðin venjulega; þ.e., standa í sama kyni, tölu og falli. Í sambandinu *hesta þinna* er *þinna* ótvírætt eignarfall, og þær upplýsingar eiga að nægja til að úrskurða að *hesta* sé líka eignarfall, en ekki þolfall eins og það gæti einnig verið ef litið er á orðið eitt og sér.

Báðar þessar tegundir hafa kosti og galla. Tölfræðimarkarar hafa þann kost að það er tiltölulega fljótlegt að koma þeim upp.[8] Á hinn bóginn er hægt að ná betri niðurstöðum, þ.e. færri röngum greiningum, með reglumörkurum en tölfræðimörkurum; og reglumarkarar ráða betur við margbrotna greiningu (stóra **markaskrá** (e. tagset)) en tölfræðimarkarar. Einnig eru til markarar sem nýta sér bæði reglur og tölfræðilegar upplýsingar, og þetta tvennt getur spilað saman á margvíslegan hátt.

Ein þekktasta útfærslan á mörkurum er kennd við Eric Brill og yfirleitt nefnd **Brill's tagger**, **Brill type tagger** eða eitthvað í þá átt (sjá einkum Brill 1995; Jurafsky & Martin 2000:307-312). Slíkur markari byggist á aðferð sem nefnist á ensku **transformation based learning**. Með því er átt við það að markarinn getur komið sér upp reglum um greiningu orðmynda í tilteknu umhverfi, og er síðan fær um að endurskoða þær reglur eftir á að fengnum nýjum og betri upplýsingum. Slíka markara má því kalla **námfúsa**.

Það er sameiginlegt bæði reglumörkurum og námfúsum mörkurum að þeir þurfa á að halda sérstöku **þjálfunarsafni** (e. training corpus). Það er texti sem hefur verið greindur handvirkt eftir sama kerfi og vélræna greiningin á að nota. Þetta safn nýtist við leit að þeim mynstrum í textanum sem hægt er að setja fram í regluformi, og eftir því sem það er stærra má búast við betri niðurstöðum.

[8] Anders Nøklestad frá Tekstlaboratoriet í Osló nefndi það í fyrirlestri sínum á ráðstefnunni *Sambúð tungu og tækni* 13. nóvember sl. að hann hefði gert tölfræðimarkara fyrir norsku á þremur mánuðum (sjá einnig Nøklestad 1998). Það fóru hins vegar fjögur ársverk í norska reglumarkarann.

Við gerð reglumarkara eru það málfræðingar sem leita að þessum mynstrum, en námfúsir markarar sjá sjálfir um að finna mynstrin og byggja reglur á þeim í samræmi við sérstök **sniðmát** (e. templates), sem leiðbeina um það hvað í umhverfinu gæti skipt máli. Tölfræðimarkarar þurfa á hinn bóginn ekki á þjálfunarsafni að halda, þótt þeir geti vissulega nýtt slíkt safn til að afla upplýsinga um tíðni einstakra orðmynda, og tíðni mismunandi greininga á sömu orðmynd. Nauðsynlegt er að þjálfunarsafnið sé stórt og samsett úr fjölbreyttum textum til að markarinn sem byggist á því skili sem réttustum niðurstöðum.

Markarann má síðan keyra á sérstakt **prófunarsafn** (e. test corpus) sem er texti með réttri greiningu allra orða. Þá er hægt að meta hversu fullkominn hann er, út frá því hversu oft hann skilar sömu greiningu og orðin hafa í prófunarsafninu. Nauðsynlegt er að greina hvaða villur markarinn gerir og reyna síðan að endurbæta hann út frá þeirri greiningu.

Málfræðimarkari fyrir íslensku er ekki til – a.m.k. ekki á almennum markaði. Stefán Briem skrifaði þó á sínum tíma greiningarforrit sem notuð voru við vinnslu *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon & Stefán Briem 1991; sjá Stefán Briem 1990). Nýlega hafa svo verið gerðar tilraunir með að koma upp námfúsum málfræðimarkara fyrir íslensku (Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir 2002). Þær tilraunir lofa góðu, en eru á algeru byrjunarstigi.

## 3. Lokaorð

Í þessari grein var annars vegar reynt að skilgreina og skýra nokkur algeng hugtök sem tengjast tungutækni, og hins vegar leitast við að gera grein fyrir því hvað þarf til að íslensk tungutækni verði meira en nafnið tómt. Þar var nefnt menntað fólk, málsöfn og málgreiningarforrit. Allt þetta ætti að geta orðið til á næstu árum með þeim stuðningi frá almannafé sem þegar hefur verið ákveðinn. Getum við þá treyst því að íslensk tungutækni standi í blóma innan fárra ára?

Fyrir því er reyndar engin trygging. Íslensk tungutækni stendur nefnilega og fellur með því að markhópurinn, fólkið í landinu, vilji nýta sér hana. Það er alls ekkert sjálfgefið að svo verði, eins og vikið er að í skýrslu starfshóps um tungutækni (*Tungutækni* 1999:23). Þetta veltur ekki síst á því hversu vel gengur að koma tungutæknitólum á markað, og hversu hratt erlend tól af því tagi ryðja sér til rúms. Þegar menn hafa vanist því að nota erlent mál á einhverju tilteknu sviði er hægara sagt en gert að koma íslenskunni þar inn aftur, eins og vel sést á ritvinnslukerfum á íslenskum markaði undanfarna tvo áratugi (sjá einnig Eiríks Rögnvaldsson 1998). Ef hér yrði komið upp samræðukerfum sem töluðu ensku er hætt við að slík kerfi myndu fljótt festast í sessi.

Um þetta skal engu spáð. Vissulega stendur íslensk tunga að mörgu leyti mjög sterkt, og er notuð á öllum sviðum þjóðlífsins (sjá *Tungutækni* 1999:14-15). Á hinn bóginn eru Íslendingar þekktir að nýjungagirni á tæknisviðinu, og ólíklegir til að bíða lengi með að taka í notkun hvers kyns tæki og tól sem byggjast á tungutækni. Líklegt er að slík tæki og tól komi með sívaxandi þunga inn á markaðinn á næstu árum. Það verður forvitnilegt að fylgjast með því hvernig íslenskri tungutækni – og íslenskri tungu – reiðir af í því umróti.

HEIMILDIR


Ásta Svavarsdóttir. 2001. Orðaforði talmáls og ritmáls. Frumathugun á orðaforðanum í ÍS-TAL með samanburði við ritmálstexta. Erindi á 5. málþingi Rannsóknarstofnunar Kennaraháskóla Íslands 13. október.

Baldur Jónsson. 1975. *Tíðni orða í Hreiðrinu.* Tilraunaverkefni í máltölvun. Rannsóknastofnun í norrænum málvísindum, Háskóla Íslands, Reykjavík.

Baldur Jónsson. 1978. *Orðstöðulykill að Hreiðrinu.* Háskóli Íslands, Reykjavík.

Baldur Jónsson, Björn Ellertsson & Sven Þ. Sigurðsson. 1980. *Tölvukönnun á tíðni orða og stafa í íslenskum texta.* Raunvísindastofnun Háskóla Íslands, Reykjavík.

*Biblíulykill.* 1994. Orðalyklar að Biblíunni 1981. Biblíulykilsnefnd og Hið íslenska Biblíufélag, Reykjavík.

Björn Magnússon. 1951. *Orðalykill að Nýja testamentinu.* Ísafoldarprentsmiðja, Reykjavík.

Braasch, Anna, Anni Buhr Christensen, Sussi Olsen & Bolette S. Pedersen. 1998. A Large Scale Lexicon for Danish in the Information Society. A. Rubio, N. Gallardo & A. Tejada (ritstj.): *Proceedings from the First Conference on Language Resources and Evaluation*, Granada. s. 249-255; http://www.cst.dk/sto/granada/uk/index.html

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* 21:543-566.

Butler, Christopher. 1985. *Computers in Linguistics.* Blackwell, Oxford.

Carlson, Rolf, Pétur Helgason, Björn Granström, Höskuldur Þráinsson & Páll Jensson. 1990. An Icelandic Text-to-Speech System for the Disabled. *Proceedings of ECART (European Conference on the Advancement of Rehabilitation Technology), Maastricht, the Netherlands, 5-8 November 1990,* kafli 3.7.

*COBUILD* = Collins COBUILD English Dictionary. 1987. Ritstj. John Sinclair. Collins, Birmingham.

Eiríkur Rögnvaldsson. 1998. Informationsteknologien og små sprogsamfund. *Sprog i Norden*, s. 82-93.

Eiríkur Rögnvaldsson. 2001. Mál og tölvur. Þórunn Blöndal og Heimir Pálsson (ritstj.): *Alfræði íslenskrar tungu*. [Margmiðlunardiskur.] Lýðveldissjóður og Námsgagnastofnun, Reykjavík.

Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir og Sigrún Helgadóttir. 2002. Vélræn málfræðigreining með námfúsum markara. Erindi á 16. Rask-ráðstefnu Íslenska málfræðifélagsins 26. janúar.

Friðrik Magnússon. 1988. Hvað er títt? Tíðnikönnun Orðabókar Háskólans. *Orð og tunga* 1:1-49.

Gazdar, Gerald, & Chris Mellish. 1989. *Natural Language Processing in Prolog.* An Introduction to Computational Linguistics. Addison-Wesley, Wokingham.

*Íslendinga sögur. Orðstöðulykill og texti.* 1996. [Geisladiskur.] Ritstj. orðstöðulykils Bergljót S. Kristjánsdóttir, Eiríkur Rögnvaldsson, Guðrún Ingólfsdóttir og Örnólfur Thorsson. Mál og menning, Reykjavík.

*Íslensk orðabók.* 2000. Tölvuútgáfa. [Geisladiskur.] Ritstj. Mörður Árnason. Mál og menning, Reykjavík.

Jón Hilmar Jónsson. 1994. *Orðastaður.* Orðabók um íslenska málnotkun. Mál og menning, Reykjavík.
Jurafsky, Daniel, & James H. Martin. 2000. *Speech and Language Processing.* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey.
Jörgen Pind (ritstj.), Friðrik Magnússon & Stefán Briem. 1991. *Íslensk orðtíðnibók.* Orðabók Háskólans, Reykjavík.
Kellogg, Robert. 1988. *A Concordance to Eddic Poetry.* Colleagues Press, East Lansing, MI.
Kjartan R. Guðmundsson. 1986. Tölvutal. Óprentuð BS-ritgerð í tölvunarfræði við Háskóla Íslands, Reykjavík.
Lodge, David. 1996. *Lítill heimur.* Háskólarómansa. Sverrir Hólmarsson þýddi. Uglan, Reykjavík.
McEnery, Tony, & Andrew Wilson. 1996. *Corpus Linguistics.* Edinburgh University Press, Edinburgh.
Nøklestad, Anders. 1998: Statistisk disambiguerende tagging av norsk. Jan Terje Faarlund, Britt Mæhlum & Torbjørn Nordgård (ritstj.): *MONS 7. Utvalde artiklar frå det 7. Møtet Om Norsk Språk i Trondheim 1997.* Novus Forlag, Osló.
Pétur Helgason. 1990. *Lokaskýrsla verkefnis um tölvutal.* Málvísindastofnun Háskóla Íslands, Reykjavík.
Renouf, Antoinette. 1987. Corpus Development. Sinclair (ritstj.), s. 1-40.
Sinclair, John M. (ritstj.). 1987. *Looking Up.* An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. Collins, London.
Stefán Briem. 1988. *Vélrænar tungumálaþýðingar. Rannsóknarskýrsla.* Reykjavík.
Stefán Briem. 1990. Automatisk morfologisk analyse af islandsk tekst. Jörgen Pind & Eiríkur Rögnvaldsson (ritstj.): *Papers from the Seventh Scandinavian Conference of Computational Linguistics, Reykjavík 1989*, s. 3-13. Institute of Lexicography & Institute of Linguistics, Reykjavík.
Stefán Briem. 1991. *A Dependency Syntax of Icelandic According to Guidelines of DLT.* Reykjavík.
Stork, David G. (ritstj.). 1996. *Hal's Legacy: 2001's Computer as Dream and Reality.* MIT Press, Cambridge, Mass.; http://mitpress.mit.edu/e-books/Hal/
Trausti Þór Kristjánsson. 1995. Íslenskur talgervill. Óprentuð BS-ritgerð í rafmagnsverkfræði við Háskóla Íslands, Reykjavík.
*Tungutækni.* 1999. Skýrsla starfshóps. [Höf. Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson.] Menntamálaráðuneytið, Reykjavík.
Þorgeir Sigurðsson. 1981. Talvélar „Voice Synthesisers". Óprentuð BS-ritgerð í rafmagnsverkfræði við Háskóla Íslands, Reykjavík.
Þórunn Blöndal. 2001. Samtöl – öflugt tæki til náms. Erindi á 5. málþingi Rannsóknarstofnunar Kennaraháskóla Íslands 13. október.
Örnólfur Thorsson. 1993. *Orð af orði.* Hefð og nýmæli í Grettlu. MA-ritgerð í íslenskum bókmenntum við Háskóla Íslands, Reykjavík.
Örnólfur Thorsson. 1994. Grettir sterki og Sturla lögmaður. *Samtíðarsögur.* The Contemporary Sagas. Níunda alþjóðlega fornsagnaþingið. Forprent, bls. 907-933. Akureyri.