Eiríkur Rögnvaldsson

Textasöfn og málrannsóknir

Erindi flutt í Ríkisútvarpinu 11. janúar 2011

Á undanförnum 20-30 árum hafa forsendur til margvíslegra málrannsókna gerbreyst vegna þróunar í tölvu- og upplýsingatækni. Þar kemur ýmislegt til. Í fyrsta lagi er nú hægt að nálgast gífurlegt magn texta á tölvutæku formi – allt efni sem gefið er út, bækur, blöð, auglýsingabæklingar o.s.frv. er unnið í tölvum og þótt það efni sé vitanlega misaðgengilegt er yfirleitt lítill vandi fyrir málfræðinga að fá yfrið magn tölvutækra texta til að vinna með. Í öðru lagi er textamagnið á vefnum yfirgengilegt og mikið bætist við á hverjum degi. Þar er því hægt að vera með fingurinn á púlsinum – alltaf eru til splunkunýir textar til skoðunar.

Þetta breytir á ýmsan hátt aðstöðu málfræðinga til að rannsaka viðfangsefni sitt, tungumálið. Með leitarvélum eins og Google er hægt að leita á augabragði að dæmum í gífurlegu textamagni. Fyrir daga netsins og tölvutækra texta höfðu menn ekki önnur ráð til dæmaleitar en lesa textana frá orði til orðs. Slíkt er vitanlega mjög seinlegt og á þann hátt verður ekki komist yfir nema örlítið brot af því textamagni sem nú er hægt að skoða. Tilvist tölvutækra texta gerir mönnum einnig kleift að vinna með textana í margvíslegum tölfræði- og greiningarforritum og fá þannig upplýsingar um tíðni orða, meðallengd setninga o.s.frv. Þetta eru upplýsingar sem í sjálfu sér er hægt að afla án þess að nota tölvur en vegna þess hversu tímafrekt slíkt væri er það tæpast raunhæft.

En það er ekki bara textamagnið sem er margfalt meira en áður var kostur á að rannsaka. Eðli textanna hefur líka breyst. Til skamms tíma a.m.k. byggðust flestar málrannsóknir á rituðu máli, aðallega á prentuðum bókum og blöðum. En prentað mál er í eðli sínu íhaldssamt. Útgefið efni hefur oftast verið vandlega yfirlesið og sýnir því venjulega eingöngu viðurkennt mál sem samræmist tiltölulega formlegu málsniði – ýmiss konar frávikum frá því hefur verið útrýmt í yfirlestrinum. Þar eru því ekki miklar líkur á að finna t.d. málbreytingar sem eru að koma upp og hafa ekki öðlast viðurkenningu. Slíkt er miklu frekar að finna í talmáli, en dæmasöfnun úr því er margfalt torveldari.

Á netinu getur hins vegar hver sem er skrifað hvað sem er. Þar er því að finna allt annars konar ritað mál en menn áttu kost á að skoða til skamms tíma – mál sem er fullt af ófullkomnum setningum, erlendum slettum, beygingarmyndum sem víkja frá formlegu málsniði, o.s.frv. Þess vegna gefst þarna gullið tækifæri til að afla dæma um málbreytingar sem eru að koma upp en hafa ekki enn ratað inn í ritmálið. Ég efast t.d. um að margir hlustendur hafi heyrt samböndin *mig finnst* eða *mig þykir*, hvað þá séð þau á prenti – sagnirnar *finnast* og *þykja* taka með sér þágufall. Ef ekki þarf annað en gúgla þessi sambönd innan gæsalappa til að finna fjölda dæma um þau, eins og Anton Karl Ingason meistaranemi í málfræði hefur sýnt fram á. Þarna virðist því örla á málbreytingu sem útilokað er að finna í hefbundnum prentuðum textum.

Notkun stórra textasafna til málrannsókna tíðkaðist reyndar fyrir daga tölvutækninnar þótt úrvinnslan yrði þá að fara fram í höndunum og væri því margfalt seinlegri. En viðhorfin til textasafna og megindlegra athugana á tungumáli gerbreyttust svo að segja á einni nóttu fyrir hálfri öld. Það sem olli þessum straumhvörfum var tilkoma

málkunnáttufræðinnar, generatífrar málfræði, en upphaf hennar er rakið til bókarinnar *Syntactic Structures* sem bandarískí málfræðingurinn Noam Chomsky gaf út 1957. Chomsky gaf lítið fyrir gildi textaathugana en byggði málfræði sína þess í stað á máltilfinningu og dómum málnotenda. Áhrif Chomskys voru mjög mikil og næstu áratugina þótti fæstum setningafræðingum ástæða til að leggjast í dæmasöfnun úr textum til að rökstyðja kenningar sínar, heldur bjuggu sjálfir til dæmi sín og dæmdu þau tæk eða ótæk. Þessi aðferð þykir vissulega enn góð og gild, en á seinni árum hafa menn aftur horfið til dæmasöfnunar úr textum og láta aðferðirnar vinna saman og bæta hvora aðra upp.

Ein meginröksemd Chomskys fyrir því að textasöfn væru gagnslítil í málfræðilegri greiningu og röksemdafærslu var sú að þau væru ævinlega og óhjákvæmilega takmörkuð, endanleg, og tilviljanakennd. Auðvelt er t.d. að tilfæra ýmis dæmi um setningar og setningagerðir sem sjaldan eða aldrei finnast í textasöfnum, jafnvel mjög stórum, en málhöfum ber þó saman um að séu tækar. Þetta hefur oft verið notað sem rök fyrir því að málhæfnin sé að verulegu leyti meðfædd; menn geti ekki hafa lært slíkar setningar af öðrum, heldur hljóti að hafa einhverja meðfædda þekkingu á þeim reglum sem um þær gilda.

Skiptar skoðanir um þessi mál leiddu til hálfgerðs stríðs milli málkunnáttufræðinga (generatífista) og þeirra sem fengust við gagnamálfræði (corpus linguistics). Chomsky talaði víða óvirðulega um gagnamálfræði, og í ritum gagnamálfræðinga er að finna mörg og beitt skot á Chomsky og fylgismenn hans. Hér er þó rétt að halda því til haga að þarna er að verulegu leyti um sýndarágreining að ræða – meðvitað eða ómeðvitað. Menn voru nefnilega ekki að tala um sama hlutinn. Chomsky var að tala um málhæfni (competence) en gagnamálfræðingar skoða málbeitingu (performance). Chomsky var sem sé að tala um málfræðina, málkerfið, en gagnamálfræðingar skoða afurð kerfisins – málið sjálft. Þarna á milli er flókin víxlverkun sem ekki hefur verið kortlögð til fulls, en meginatriðið er að báðar aðferðirnar eiga fullan rétt á sér og eru nauðsynlegar – en þær svara mismunandi spurningum.

Textasöfn eru þannig gagnleg til að finna ýmsar setningagerðir og átta sig á þeim. Það er þannig hægt að nota þau, að vissu marki, til að úrskurða tiltekna setningagerð tæka. Það er hins vegar ekki hægt að nota þau til að úrskurða setningagerð ótæka. Þótt hún komi ekki fyrir í þeim textum sem við skoðum getur það verið tilviljun. Eðli málsins samkvæmt getur textasafn okkar aldrei innihaldið allar hugsanlegar setningar. Ef við erum að lýsa málinu (ekki málkerfinu) gerir þetta ekkert til. Textasafnið sem við erum með undir afmarkar þá viðfangsefni okkar, og ef tiltekin setningagerð kemur ekki fyrir í safninu er hún ekki hluti viðfangsefnisins og kemur okkur þess vegna ekki við.

En ef við erum að lýsa málkerfinu sjálfu horfir málið öðruvísi við. Það málkerfi sem við lýsum á að gera okkur kleift að mynda allar málfræðilega tækar setningar en ekki aðrar. Þess vegna nægir okkur ekki að vita hvers konar setningar eru tækar – við þurfum líka að vita hvers konar setningar væru ótækar. Og því svarar textasafnið ekki – það er vitaskuld ekki hægt að takmarka mengið „tækar setningar“ við þær setningar sem fyrir koma í tilteknu safni, hversu stórt sem það er. Í samtímalegri setningafræði er hægt að snúa sig út úr þessum vanda með þeim einfalda hætti að spyrja málnotendur. Þá erum við ekki háð afmörkuðu textamengi, heldur getum búið til texta eftir þörfum, ef svo má segja, og borið þá undir málnotendur og fengið dóma þeirra um hvort tiltekin setning sé tæk eða ekki.

Þeir sem fást við sögulega setningafræði eiga aftur á móti ekki þessa útleið – þeir verða að reiða sig algerlega á textana. Við höfum ekki annað til að miða við en þá texta sem varðveittir eru eða við höfum aðgang að – og þeir eru ekki alltaf miklir. En við þurfum líka að gæta þess að skoða þá alla áður en við fullyrðum nokkuð, gæta þess að ekki komi eitthvað annað til sem geti valdið því að viðkomandi setningagerð finnst ekki á eldra málstigi, o.s.frv.

Hér á undan var sagt að hægt væri – að vissu marki – að nota textasöfn til að úrskurða tiltekna setningagerð tæka. En þar verður líka að hafa fyrirvara. Því fer nefnilega fjarri að allar setningar sem koma fyrir í textasöfnum séu tækar í raun og veru, þ.e. samræmist málkerfi flestra málnotenda. Þótt prentaðir textar séu venjulega yfirlesnir og villur leiðréttar er samt alltaf í þeim eitthvað um setningar sem flestir myndu telja ótækar – setningar þar sem fyrir koma ýmiss konar pennaglöp, mistök í ritvinnslu, prent- og ásláttarvillur, einstaklingsbundið málfar, o.s.frv. Óvíst er að við tökum nokkuð eftir villunum þegar við lesum textann, en ef við gerum það þá leiðréttum við þær oftast með sjálfum okkur þegjandi og hljóðalaust, í samræmi við málkennd okkar.

En hvenær getum við leyft okkur það? Hvað með þá sem fást við eldri málstig og geta ekki beitt eigin málkennd á textana? Verðum við að líta á allar setningar sem finnast í textum sem jafnréttháar? Ekki gera menn það alltaf í raun; oft leyfa menn sér að telja að afbrigðileg eða óvenjuleg setning eða orðmynd í fornum texta sé pennaglöp af einhverju tagi, en endurspegli ekki málfar skrifarans. En þarna eru menn vissulega á hálum ís, og oft getur verið freisting að láta fræðikenningar taka af sér ráðin; hafna setningum sem koma fyrir ef þær falla ekki að þeirri kenningu sem maður vinnur með, en gera ráð fyrir öðrum sem ekki finnast dæmi um, vegna þess að kenningin segir að þær ættu að geta komið fyrir.

Viðhorf margra málfræðinga til texta og textasafna hefur breyst á seinni árum. Nú þykir miklu eðlilegra en fyrir fáum árum að rökstyðja setningafræðilegar greiningar með dæmum úr töluðu eða rituðu máli. Hrint hefur verið af stað viðamiklum rannsóknarverkefnum til að kanna setningafræðilegan mállýskumun og safna setningafræðilegum dæmum. Að hluta til má skýra þessa þróun með því að máldæmi, einkum ritmálsdæmi, eru orðin mun auðfengnari en áður. Með tilkomu sístækkandi rafrænna textasafna er nú orðið auðvelt að safna fjölbreyttum textadæmum af ýmsu tagi. Vefurinn hefur svo gert mönnum kleift að komast í margs konar texta sem áður voru óaðgengilegir og einnig hafa þar orðið til nýjar textategundir sem margar hverjar standa nær talmálinu en hefðbundnu ritmáli.

En textasöfn eru ekki bara gagnleg til að finna dæmi um orð og setningagerðir; þau má ekki síst nota til að kanna tíðni orða og orðasambanda. Fyrsta íslenska rannsóknin af slíku tagi var reyndar gerð löngu fyrir daga tölvutækninnar. Þá rannsókn gerði Ársæll Sigurðsson skólastjóri og birtust niðurstöður hans í *Menntamálum* árið 1940. Tilgangur rannsóknarinnar var hagnýtur; „að finna leið til að gera stafsetningarkennsluna aðgengilegri og raunhæfari en áður, en þó vænlegri til betri árangurs“, eins og Ársæll segir. Textar hans voru úr stílum barna, sendibréfum fullorðinna, lesbókum, náttúrufræði, sögu og landafræði. Textarnir voru alls um 100 þúsund lesmálsorð, en mismunandi orðmyndir voru 13.636. Algengasta orðmyndin í íslensku samkvæmt þessari könnun var *og*, þá kom *að*, og í þriðja og fjórða sæti forsetningarnar *í* og *á*.

Langsamlega ítarlegasta rannsókn sem til er á tíðni íslenskra orða var gerð hjá Orðabók Háskólans fyrir um 20 árum, laust fyrir 1990, en niðurstöður hennar birtust í *Íslenskri orðtíðnibók* árið 1991. Höfundar þeirrar bókar eru Jörgen Pind, Friðrik Magnússon og Stefán Briem. Hráefnið í þá könnun var valið þannig að það gæfi mynd af helstu textategundum þótt vissulega vanti þar ýmislegt. Að baki könnuninni liggja rúm 500 þúsund lesmálsorð úr 100 mismunandi textum af fimm efnisflokkum, 20 úr hverjum flokki. Þótt þessi könnun gerð hálfri öld síðar en könnun Ársæls, og textarnir sem liggja til grundvallar séu af ólíkum toga, eru fjórar algengustu orðmyndirnar samt þær sömu, og í sömu röð – *og, að, í, á*. Þar á eftir koma svo *sem, var, hann, við, það, en*.

Fram til þessa hefur vantað stórt og aðgengilegt textasafn sem unnt væri að nota við margvíslegar rannsóknir á íslensku nútímamáli, en nú er að verða breyting þar á. Undanfarin ár hefur verið unnið að því undir stjórn Sigrúnar Helgadóttur á Stofnun Árna Magnússonar í íslenskum fræðum að koma upp stóru safni fjölbreyttra texta sem valdir eru eftir ákveðnum reglum. Um slíkt safn hefur verið notað nýyrðið *málheild*. Safnið verður samtals 25 milljónir lesmálsorða af ýmsum tegundum, úr textum frá síðustu 10 árum, þ.e. árunum 2000-2009. Stærstu textaflokkarnir eru dagblaðaefni og skáldverk, en einnig eru þarna bloggtextar, ýmiss konar fræðitextar af Vísindavefnum, textar af vefsetrum stofnana og fyrirtækja, lagatextar og ýmislegt fleira, þar á meðal dálítið af talmáli. Textarnir verða markaðir, það er að segja greindir í orðflokka og öll hefðbundin málfræðileg greiningaratriði s.s. kyn, tölu, fall, hátt, tíð, stig o.s.frv. Safnið er því nefnt *mörkuð íslensk málheild*. Það hefur að miklu leyti verið kostað af styrk frá tungutækniverkefni menntamálaráðuneytisins.

Safn af þessu tagi er mjög mikilvægt í máltækni vegna þess að úr því má vinna ítarlegar upplýsingar um íslenskt mál sem nauðsynlegar eru við gerð ýmiss konar máltæknibúnaðar, svo sem leiðréttingarforrita, þýðingarforrita, leitarforrita o.fl. En safnið gagnast einnig þeim sem vinna að rannsóknum á íslensku; þar er hægt að skoða tíðni orða, orðmynda, orðasambanda og setningagerða. Jafnframt á safnið að geta gagnast almenningi vegna þess að unnt verður að fletta upp í því á vefnum og skoða þannig notkun orða og orðasambanda.