Maren Albertsdóttir og
Stefán Einar Stefánsson

Beygingar- og málfræðigreinikerfi

Maren Albertsdóttir og Stefán Einar Stefánsson

Beygingar- og málfræðigreinikerfi

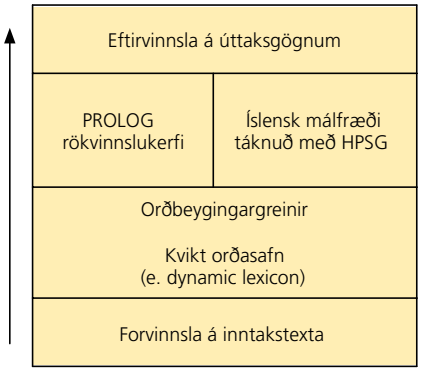
Markmið verkefnisins er að þróa kerfi sem tekur inn setningar á íslensku og skilar upplýsingum um byggingu þeirra og eiginleika einstakra hluta þeirra. Málfræðigreinikerfið nýtir sér beygingargreinikerfi sem tekur inn einstök orð og skilar út upplýsingum um orðflokka og beygingarmyndir. Kerfið skilar því upplýsingum um íslenskar setningar út frá beygingar- og setningafræði.

Verkefnið byggist á þekktum aðferðum sem hafa verið notaðar til að þróa sambærileg kerfi víða um heim fyrir ólík tungumál. Markmiðið er að byggja kerfið eingöngu á beygingar- og setningafræði en geyma merkingarfræði að eins miklu leyti og hægt er. Mörk setninga- og merkingarfræði eru oft óljós en sneitt er framhjá hreinum merkingarfræðilegum atriðum. Kerfið býður samt sem áður upp á að það verði þróað enn frekar út frá merkingarfræði í framtíðinni.

Verkefnið stendur enn yfir og er unnið af Stefáni Einari Stefánssyni og Maren Albertsdóttur ásamt Friðriki Skúlasyni. Eiríkur Rögnvaldsson, umsjónarmaður meistaranáms í tungutækni við Háskóla Íslands, hefur veitt sérfræðiaðstoð sína á verktímanum. Verkefnið hófst formlega í september 2002 og stendur enn. Í lok nóvember 2004 hafði um 36 mannmánuðum verið varið í verkefnið og eru verklok áætluð í maí 2005.

Undirstöðueiningar málfræðigreinikerfisins eru:

- Beygingargreinikerfi eða kvikt orðasafn (e. Dynamic lexicon) sem vinnur með föstu orðasafni
- Íslensk málfræði skilgreind í HPSG (e. Head Driven Phrase Structure Grammar)
- Prolog-rökvinnslukerfi



Undirstöðueiningar

Kvikt orðasafn

Í málfræðigreinikerfinu er stuðst við svokallað kvikt orðasafn (e. dynamic lexicon). Þetta þýðir að í stað þess að geyma langan orðalista með beygingarupplýsingum (auk annarra málfræðiupplýsinga) fyrir hvert orð þá eru þessar upplýsingar sóttar eftir þörfum úr beygingargreininum sem aflar þeirra með reiknifræðilegum hætti. Kostir þessarar aðferðar eru miklir en einn sá helsti er að minnisstærð orðasafnsins minnkar verulega. Þetta á sér í lagi við í íslensku þar sem sama orðið getur haft mjög margar myndir.

Kvika orðasafnið inniheldur um 30.000 einingar sem það getur notað til að mynda ný orð út frá íslenskum beygingar- og orðmyndunarreglum. Orðasafnið er því í raun óendanlegt. Með kvika orðasafninu er notað fast orðasafn til að kljást við undantekningar og sértilfelli.

Orðasöfn sambærilegra kerfa eru oftast byggð upp á annan hátt. Beygingarfræðileg atriði eru oft leyst í málfræðihlutanum ásamt setninga- og merkingarfræði auk þess sem meira er stuðst við föst orðasöfn. Kvika orðasafnið sem notast er við í verkefninu vinnur hins vegar sína vinnu óháð málfræðikerfinu sem slíku. Það er mjög öflugt, nákvæmt og hraðvirkt sem gerir kerfið í heild skilvirkara og auðveldara í viðhaldi og vexti.

HPSG

HPSG (e. Head Driven Phrase Structure Grammar) er málfræði sem auðvelt er að umrita á form sem tölvur skilja. HPSG vinnur með málfræði út frá setninga- og merkingarfræði og sækir ýmislegt í aðrar þekktar málfræðikenningar. Að auki mynda tölvunarfræði, stærðfræði og fleiri skyldar greinar fræðilegan grundvöll undir HPSG.

Sérhvert orð, sem og stærri liðir, fá úthlutað sérstökum ramma í málfræðinni þar sem allar beygingar-, setninga- og merkingarfræðilegar (ef einhverjar eru) upplýsingar koma fram. Heimur málfræðinnar er skilgreindur með nokkurskonar erfðastigveldi sem segir til um af hvaða tagi ákveðinn hlutur er, hvernig hann tengist öðrum hlutum í kerfinu og hvaða skorður hann setur. Hvert orð fær merkimiða frá orðasafninu, einn frá beygingargreinikerfi og annan frá málfræðinni sem segir til um setningafræðilegt hlutverk þess. Reglur, frumsendur og skorður sjá síðan um að binda saman orðin.

Málfræðifyrirbæri í kerfinu er sett fram í römmum sem tákna í raun skorður. Málfræðin samanstendur því af safni af skorðum. Sem dæmi má nefna að sagnorð setja skorður á umhverfi sitt (nærliggjandi orð og/eða liði) þar sem þau vilja t.d. stjórna í hvaða falli frumlag og fyllilíðir eru sem og af hvaða tagi orðin eru, dæmi: ‘Mig (ÞF) langar’ en ekki *’Mér (ÞGF) langar’.

Rökvinnsluvélin

Rökvinnsluvélin reiknar út allar mögulegar greiningar á setningunni sem hún tekur inn samkvæmt þeim skorðum sem málfræðinsetur (e. All-paths parsing). Allar greiningar eru gefnar, sama hversu líklegar þær eru. Útreikningarnir sem vélin framkvæmir eru þar af leiðandi mjög þungir. Kerfið gefur möguleika á að takmarka þær greiningar sem koma til greina t.d. með því að setja inn tölfræðisíu sem velur aðeins „líklegustu" greiningarnar.

Allt þróunarumhverfi hefur verið þróað innan fyrirtækisins.

Afrakstur

Markmiðið með verkefninu var ekki að þróa ákveðið notendakerfi heldur fyrst og fremst að til yrði grunnkerfi sem nýttist í frekari þróunarvinnu á sviði máltækni, gerð kennsluefnis og/eða -forrita eða í hverskyns rannsóknum á íslensku máli. Málfræðigreinikerfið er undirstöðukerfi við þróun í framtíðinni á ýmiskonar tölvubúnaði á sviði máltækni. Kerfið gefur svokallaða djúpgreiningu á texta sem er mjög nákvæm beygingar- og setningafræðileg greining. Aðferðafræði, eins og notuð er í kerfinu, getur nýst við gerð leiðréttingarforrita, talgreina, talgervla og þýðingarvéla.

Verkefnið hefur nú þegar skilað margvíslegum tólum og þekkingu á sviði máltækni. Þróaður hefur verið ýmis hugbúnaður og hugbúnaðareiningar í tengslum við verkefnið ásamt málfræðihugbúnaði (reiknivél). Hluti af íslenskri málfræði hefur verið skilgreindur í HPSG. Byggt hefur verið upp öflugt orðasafn sem byggist á beygingar- og setningafræði og samanstendur af kviku og föstu orðasafni. Í vinnuferlinu hefur skapast þekking á smíði víðtæks málfræðikerfis, ákveðnu verklagi og aðferðafræði sem spilar stórt hlutverk í þróun íslenskra máltæknitóla.

Verkefnið hefur þróast í að verða meira rannsóknarverkefni en búist var við í upphafi. Þess má geta að unnið er að verkefnum víða um heim sem byggjast á sömu aðferðum m.a. fyrir ensku, þýsku og japönsku. Mestöll þróunar- og grunnvinna fer fram í háskólum en er styrkt af opinberum aðilum og fyrirtækjum sem síðan hagnýta hana. Hér á landi hafa hvorki verið stundaðar sambærilegar rannsóknir á sviði máltækni né þeim hluta sem snýr að tölvunarfræði og tæknilegri útfærslu. Að því leytinu blasir allt annað landslag við þeim sem vilja byggja upp máltæknitól fyrir íslensku en víða annarsstaðar. Það er von þátttakenda að sú þekking og reynsla sem hefur skapast og áunnist á þessum tíma skili sér áfram með einhverjum hætti.