Helga Waage

Hjal – gerð íslensks stakorðagreinis

Helga Waage

Hjal – gerð íslensks stakorðagreinis

Hjal-verkefnið nefnist verkefni sem sneri að gerð talgreinis fyrir íslensku. Tilurð verkefnisins var sú að menntamálaráðuneytið ákvað að efna til tungutækniátaks til að styrkja íslenskustuðning í ýmsum tölvukerfum. Um svipað leyti var stofnuð tungutækniskor við Háskóla Íslands, þverfaglegt nám er miðast að því að mennta fólk til rannsókna og starfa á mörkum málvísinda og upplýsingatækni. Nokkur fyrirtæki á hugbúnaðar- og fjarskiptasviðum töldu að skilningur á íslensku mæltu máli væri mikilvægur, bæði fyrir þá þjónustu sem þau vildu geta boðið upp á en einnig væri mikilvægt fyrir íslenska tungu að hún væri jafnsjálfsagt viðmót við tæki og önnur tungumál. Upp úr þessum farvegi varð til samstarfshópur um gerð íslensks stakorðagreinis – Hjalhópurinn.

Að Hjali stóðu Háskóli Íslands, Hex hugbúnaður, Síminn, Nýherji og Grunnur-Gagnalausnir (nú Trackwell). Ákveðið var að leita samstarfs við öflugan framleiðanda á sviði talgreina og varð þýska fyrirtækið Philips (nú í eigu Scansoft) fyrir valinu. Það val helgaðist fyrst og fremst af því að tækni þess var þróuð með það í huga að styðja við mörg tungumál og því var gerð tungumálapakkans vel skilgreind eining innan kerfisins. Íslenska var 48. tungumál talgreinisins og nutum við góðs af yfirgripsmikilli þekkingu Scansoft-manna af gerð talgreina fyrir önnur tungumál.

Hvað er stakorðagreinir?

Stakorðagreinir er talgreinir sem skilur öll orð í einhverju tungumáli – bara ekki öll í einu. Stakorðagreinum er ætlað að skilja alla málhafa en einungis takmarkað orðamengi hverju sinni. Hversu takmarkað orðamengið er fer eftir ýmsu, til dæmis hversu lík orðin eru, hversu einsleitur framburður málhafa er, hversu vel þjálfað orðalíkanið er og hversu öflug tölvan er sem talgreinirinn keyrir á. Stærð orðamengisins getur þannig verið frá örfáum orðum upp í um 2 milljónir. Algeng stærð á orðamengi fyrir þjónustu er af stærðargráðunni 100–1000 orð.

Einnig eru til talgreinar sem skilja samfellt mál – dictational-kerfi. Það eru kerfi sem eru þjálfuð til að skilja samfellt tal hjá einum eða mjög fáum einstaklingum, eða þá að skilja mjög takmarkaðan orðaforða. Einfalt

dæmi um slíkan talgreini er talgreinirinn sem fylgir með Windows-stýrikerfinu og áhugasamir geta dundað sér við að þjálfa upp til að taka niður skjöl sem lesin eru fyrir. Einnig eru til svona talgreinar fyrir mjög takmörkuð orðamengi, til dæmis fyrir skurðlækna. Enn er ekki til talgreinir af þessu tagi fyrir íslensku.

En aftur að stakorðagreininum. Stakorðagreinir vinnur yfirleitt sem viðmót á annað kerfi, til dæmis upplýsingasíma um færð á þjóðvegum. Hann er gjarna notaður sem viðmót í gegnum símkerfi þar sem hann hlustar á viðmælanda sinn og greinir út úr hljóðaflaumnum orð og orð á stangli sem hann sendir síðan til samstarfskerfisins ásamt upplýsingum um það hversu öruggur hann sé um orðagreininguna. Ef um er að ræða upplýsingar um færð á íslenskum hálendisvegum væru orðin sem greinirinn hlustar eftir heiti á borð við *Steingrímsfjarðarheiði, Hellisheiði og Holtavörðuheiði* sem öll eru ólík og greinirinn greinir með miklu öryggi á milli þeirra. Upplýsingaveitur á borð við 118 þurfa hins vegar að greina á milli þúsunda nafna sem hafa þann ágalla að vera ekki einkvæm (margir heita *Guðmundur Jónsson*), nöfn sem hljóma svipað (eins og *Ebba Dóra og Edda Þóra*) svo ekki sé talað um þá sem hringja í 118 ef þá vantar uppskrift að sósu eða eru að leita að góðri hársnyrtistofu. Það er því mun auðveldara að búa til þjónustu sem miðlar sjálfvirkt upplýsingum um færð á hálendisvegum til símnotenda heldur en þjónustu sem veitir upplýsingar um símanúmer þótt bæði kerfin noti sömu grunntæknina.

Framkvæmd Hjalverkefnisins

Sótt var um styrk til verkefnisins í árslok 2002 og er ljóst að ekki hefði orðið af verkefninu ef það hefði ekki verið styrkt á þann máta sem gert var. Undirbúningur verkefnisins hófst síðan í ársbyrjun 2003. Verkefnisstjóri var Sæmundur Þorsteinsson hjá Landssíma Íslands en rekstur verkefnisins var í höndum Helgu Waage hjá Hex. Scansoft lagði línurnar með hvaða verkþætti þyrfti að vinna og hvernig væri best að manna þá. Var ákveðið að ráða þrjá mastersnema í tungutækni til verksins og skyldu þeir vinna meginþorra vinnunnar sumarið 2003. Það þurfti því að tryggja að allur undirbúningur væri með þeim hætti að sá tími sem nemarnir hefðu til umráða myndi nýtast sem best.

Verkáætlun verkefnisins var í stórum drátt-um þessi:

- Málfræðileg forvinna – safna saman þeim gögnum sem þarf til að gera talgreininn
- Safna talsýnum – fá um það bil 2000 Íslendinga til að taka þátt í söfnun á taldæmum
- Skrá talsýni – hlusta á talsýni og skrá hvað hver einstaklingur segir og hvernig það er sagt
- Þjálfa talgreini – gögn send til Þýskaland til að þjálfa tungumálaeininguna

Málfræðileg forvinna

Forsenda þess að vel tækist til með talgreininn var sú að málfræðileg forvinna væri vel úr garði gerð – að öll málhljóð og málhljóðasambönd í íslensku væru þekkt og vitað hvar og hvernig þau kæmu fyrir. Einnig þurfti að útvega lista af staðarheitum, mannanöfnum, fyrirtækjaheitum, algengum fyrirskipunum og töluorðum. Að lokum þurftum við að útbúa eðlilegar setningar samkvæmt niðurstöðum málhljóðagreiningarinnar. Úr þessum gögnum voru útbúin blöð sem send yrðu til þátttakenda í talsýnatöku og fylgir eitt slíkt blað með þessum pistli. Orðalisti með rúmlega 30.000 algengum orðum var undirbúinn til hljóðritunar. Hljóðritaði listinn yrði síðan hafður til hliðsjónar við þjálfun talgreinisins. Að auki voru önnur orð af setningablöðunum sett á listann.

Eiríkur Rögnvaldsson, prófessor við Háskóla Íslands, tók saman megnið af þessum gögnum, sér í lagi var mikilvægt að fá yfirlit yfir öll málhljóð í íslensku.

Við undirbúning gagna kom í ljós að leiðbeiningar Scansoft áttu ekki fyllilega við íslensku. Til dæmis eru töluorð bæði fleiri og flóknari í íslensku en í flestum öðrum tungumálum en mállýskur fáar og ekki mikill munur á þeim frá sjónarhóli þarfa talgreinisins.

Afurðir þessa verkþáttar var Sampa – hljóðritunarstaðall fyrir íslensku og 1000 mismunandi bréf til að nota við talsýnatöku.

Söfnun upptakna

Markmið verkefnisins var að safna 1500-2000 upptökum af fólki 14 ára og eldra af öllu landinu. Sérstaklega var lögð áhersla á að ná málhöfum með norðlenskan framburð svo og konum eldri en 40 ára, sem okkur var sagt að skiluðu sér almennt fremur illa í verkefni af þessu tagi. Söfnunin var tvíþætt – í upphafi var leitað til fjölmiðla um að kynna verkefnið til að fá inn þátttakendur sem hefðu áhuga á verkefninu sem slíku og vildu leggja því lið. Síðan var leitað til Gallup og beðið um aðstoð við að safna því úrtaki sem upp á vantaði. Þátttakandinn fékk sent til sín bréf til að lesa og síðan hringdi hann inn og fylgdi fyrirmælum á blaðinu.

Safnað var talsýnum 2005 einstaklinga, en ríflega 3000 manns skráðu sig til þátttöku. Í 89% tilvika kláraði þátttakandinn símtalið. Yngsti þátttakandinn var 8 ára stúlka (sem því miður varð að sía frá, þar sem hún er of ung til að gagnast þjálfun talgreinisins) en sá elsti var 83 ára karl. Alls hringdu 14 einstaklingar 70 ára og eldri. Flestir sem hringdu voru á aldrinum 18 til 40 ára og voru konur heldur duglegri að hringja inn, eða 55%.

Þátttakendur voru um 1% þjóðarinnar (14 ára og eldri) og þykir það einstakt að svo stór hluti þjóðar hafi tekið þátt í verkefni af þessu tagi.

Úrvinnsla

Úrvinnsla gagna var tvíþætt. Annars vegar hljóðritun orðalista en hins vegar skráning á tali þess sem hringdi inn.

Hljóðrita þurfti öll orð sem komu fyrir í talsýnunum og auk þess um 30.000 algengar orðmyndir. Við gerð þessa orðalista (sem endaði í rúmlega 50.000 orðum) var þá valin sú mynd eða þær myndir sem taldar voru algengastar.

Við skráningu á talsýnum var hlustað á hverja einustu upptöku og skráð nákvæmlega niður hvað viðmælandinn sagði, hljóð sem heyrðust í bakgrunni, hóstar, hik og önnur málhljóð. Einnig voru mállýskur, mismæli og ýmis afbrigði skráð sérstaklega.

Upptökurnar og orðalistarnir voru síðan sendir til Scansoft sem fór yfir þá og notaði síðan gögnin til að þjálfa mállíkan talgreinisins.

Niðurstaða

Fullbúnum talgreini var skilað í byrjun nóvember 2003. Talgreinirinn virkar vel, enda kom í ljós að íslenska er vel fallin til að greina á þennan máta. Hrynjandin í málinu er nokkuð regluleg, það er fremur langt á milli hljóða og framburður er nokkuð einsleitur.

Talgreinirinn hefur nú verið í notkun í rúmt ár og er fyllilega samanburðarhæfur við stakorðagreina fyrir önnur tungumál. Hann hefur verið notaður til að veita ýmiss konar þjónustu í gegnum síma, bæði viðskiptalegs eðlis, til að veita almannaþjónustu og til skemmtunar.

Aðrar afurðir verkefnisins eru hljóðritaður orðalisti sem ekki var áður til fyrir íslensku, upptökur af talmáli 2000 kyn-, aldurs- og búsetugreindra einstaklinga og hljóðritunarstaðall fyrir íslensku.

Dæmi um innhringiblað

Velkomin(n) í hljóðsöfnun Hjals. Hljóðsöfnunin fer þannig fram að fyrst verður þú beðin(n) að segja auðkennistöluna þína. Hana er að finna í bréfinu sem þú fékkst sent. Síðan koma fáeinar spurningar sem við viljum biðja þig um að svara. Að því loknu verður þú beðin(n) að segja setningar og setningabrot úr bréfinu. Kerfið mun leiðbeina þér. Hafðu engar áhyggjur þótt eitthvað komi upp á í upptöku. Haltu áfram eins og ekkert hafi í skorist.

Það væri gott ef þú læsir bréfið einu sinni alveg í gegn áður en við byrjum hljóðsöfnunina.

Í hverri umferð gefur kerfið fyrirmæli og síðan heyrist hljóðmerki. Þú segir svarið eftir að hljóðmerkið heyrist.

Segðu auðkennistölu þína.

Hver er fæðingardagur þinn og ár?
Ertu karl eða kona?

Hvaðan ertu?

Segðu einhverja tölu milli 0 og 10 milljóna.

Hringir þú úr farsíma?

Segðu eftirfarandi staðanöfn:

1. Ólafsvík
2. Víðinesi
3. Héðinsgötu
4. Mjóafirði
5. Hegranesi
6. Holtagerði

Segðu eftirfarandi mannanöfn:

1. Pál Böðvarsson
2. Sólveig Logadóttir
3. Pétur Hreggviðsson
4. Guðrún Sigríður Erlendsdóttir
5. Anna María Ingimundardóttir

Segðu eftirfarandi nöfn stofnana og fyrirtækja:

1. Orkustofnun
2. Forsætisráðuneytið
3. Eimskipafélag Íslands
4. Mjólkursamsalan
5. Flugstöð Leifs Eiríkssonar

Segðu eftirfarandi tölur:

1. annan annar annarri annars ein eina
2. einar fern fernra fjórar fjórði fjórir
3. fjögur fyrsta fyrsti fyrstu níu níundi
4. sjötti tvenn tvennar tvennir tvo tvö
5. þrem þremur þriðji þrjár þrjú öðrum

Segðu eftirfarandi skipanir:

1. ég vildi
2. starta
3. taka frá miða

Segðu eftirfarandi peningaupphæðir:

1. 281 kr.
2. 88.123 kr.
3. 290.284 kr.
4. 1.800.000 kr.
5. 4.000.000 kr.

Segðu eftirfarandi setningar:

1. Jamm, hér er víst hægt að hnjóta um hnullung, sagði hann.
2. Ég gleymdi mér andartak og gerðist montinn.
3. Raunsæismanneskjan hafði rétt fyrir sér.
4. Við Stefán talaði hann lágt og blíðlega og Stefán hváði.
5. Þau horfðu á mig vantrúuð en höfðu hvorki önnur ráð né betri.

Segðu eftirfarandi tímasetningar:

1. Miðvikudagur 5. febrúar
2. Á eftir

Segðu eftirfarandi setningabrot:

1. Mætti ég
2. Svo er nú það

Stafaðu eftirfarandi stafarunur:

1. B A K A R Í I Ð
2. H V E R G I
3. H L M X M Ú X

Kærar þakkir fyrir þátttökuna!