Dr. Hrafn Loftsson, lektor, tölvunarfræðideild Háskólans í Reykjavík

Markari og hlutaþáttari fyrir íslenskan texta

Inngangur

Máltækni (tungutækni) er rannsóknar- og þróunarsvið sem hefur það að markmiði að smíða kerfi sem geta unnið með og skilið náttúruleg tungumál og stuðlað að notkun þeirra í samskiptum manns og tölvu. Í skýrslu, sem unnin var fyrir menntamálaráðuneytið á árunum 1998–1999, var lagt til að átak yrði gert til að efla máltækni á Íslandi enda væru Íslendingar að dragast verulega aftur úr öðrum þjóðum á þessu sviði [1]. Sérstaklega var hvatt til að útbúnar yrðu ýmiss konar *máltæknieiningar* fyrir íslenskan texta í þeim tilgangi að auðvelda notkun íslensku í upplýsingatækniþjóðfélaginu. Með máltæknieiningum er t.d. átt við *málheildir* (safn fjölbreyttra texta sem geymdir eru á stöðluðu sniði á rafrænu formi) og tól sem greina texta af ýmsu tagi. Safn máltæknieininga, þ.e. BLARK (e. Basic Language Resource Kit), er nauðsynlegur grunnur undir frekari rannsóknir og þróun í máltækni fyrir sérhvert tungumál [2].

Í þessari grein er tveimur máltæknieiningum lýst stuttlega, annars vegar *markara* og hins vegar *hlutaþáttara*. Sérstaklega er fjallað um þær einingar sem þróaðar voru í doktorsverkefni höfundar. Báðar einingarnar eru nú hluti af BLARK fyrir íslensku.

Markari

Hlutverk markara (e. tagger) er að greina (marka) sérhvert orð í texta í orðflokk og beygingarleg einkenni. Greiningarstrengurinn sem notaður er nefnist *mark* og mengi mögulegra greiningarstrengja nefnist *markamengi*. Möguleg mörk sérhvers orðs eru geymd í sérstöku orðasafni sem markarinn notar. Orð geta því verið *margræð* (e. ambiguous), þ.e. geta átt sér fleira en eitt mark, en aðeins eitt af mörkum markamengis á við sérhvert orð í tilteknu samhengi. Markari eyðir margræðni og framkvæmir því í raun svokallaða *einræðingu* (e. disambiguation).

Íslenska markamengið, sem var búið til samtímis vinnslu textasafns *Íslenskrar orðtíðnibókar* [3], samanstendur af um 700 mögulegum mörkum. Til samanburðar má nefna að eitt helsta markamengið fyrir ensku, *Penn TreeBank Tagset*, samanstendur af aðeins 45 mörkum. Þennan mun má að mestu leyti skýra með því að íslenskan er mun flóknara mál en enska hvað beygingar varðar.

Hér fyrir neðan má sjá markaðan texta fyrir fyrstu setninguna í þessum kafla (markið fyrir sérhvert orð er feitletrað):

Hlutverk **nhen** markara **nkee** er **sfg3en** að **cn** greina **sng** sérhvert **foheo** orð **nheo** í **aþ** texta **nkeþ** í **ao** orðflokk **nkeo** og **c** beygingarleg **lhfosf** einkenni **nhfo**

Sérhvert mark í íslenska markamenginu samanstendur af í mesta lagi sex stöfum sem hver og einn hefur ákveðna merkingu. Fyrsti stafurinn táknar orðflokkinn, t.d. **n**=nafnorð, **s**=sagnorð, **f**=fornafn, **l**=lýsingarorð, **c**=samtenging og **a**=atviksorð/forsetning. Stafir í sætum 2–6 tákna undirflokka og beygingarleg atriði. Lítum t.d. á mörkin **foheo** og **sfg3en.** Í fyrra markinu er **o**= óákveðið fornafn, **h**=hvorugkyn, **e**=eintala og **o**=nefnifall; í seinna markinu er **f**=framsöguháttur, **g**=germynd, **3**=þriðja persóna, **e**=eintala og **n**=nútíð[1].

Mörkurum er gjarnan skipt í tvo flokka. Annars vegar er um að ræða svokallaða *gagnamarkara*, sem læra af fyrirfram markaðri málheild á vélrænan hátt, og hins vegar svokallaða *málfræðilega reglumarkara* sem nota handgerðar reglur til að framkvæma einræðingu. Með tilkomu markaðra málheilda í ýmsum tungumálum hafa gagnamarkarar verið notaðir í ríkum mæli undanfarin 10–15 ár. Gagnamarkarar safna upplýsingum á vélrænan hátt sem síðar eru notaðar við einræðingu á nýjum texta. Upplýsingarnar geta t.d. verið í formi tölfræði eða reglna. Um 90,4% nákvæmni (hlutfall rétt markaðra orða af heildafjölda orða) hefur náðst við mörkun íslensks texta með gagnamörkurum [4, 5][2]. Málfræðilegir reglumarkarar læra ekki vélrænt af fyrirfram mörkuðum málheildum heldur byggja á handgerðum reglum sem búnar eru til af sérfræðingum og þróaðar eru með hliðsjón af mörkuðum texta.

Í þeim tilgangi að reyna að bæta nákvæmni í mörkun íslensks texta þá hefur höfundur þróað málfræðilegan reglumarkara, *IceTagger*. Markarinn byggir á *smækkunaraðferð* (e. reductionist approach), þ.e. mörk sem ekki eiga við í tilteknu staðværu (e. local) samhengi eru fjarlægð í þeirri von að í lokin standi eftir eitt rétt mark fyrir sérhvert orð. Jafnframt því að skoða staðvært samhengi þá notar markarinn *leitaraðferðir* (e. heuristics) sem sjá til þess að beygingarlegt samræmi ríki á milli frumlags og sagnar, á milli frumlags og sagnfyllingar, innan nafnliða og forsetningaliða o.s.frv. [6]. Mikilvægur hluti af *IceTagger* er beygingarlegur greinir, *IceMorphy*, sem giskar á möguleg mörk fyrir óþekkt orð, þ.e. orð sem ekki finnast í orðasafni markarans, og finnur út hvaða mörk fyrir þekkt orð vantar í orðasafnið.

Prófanir hafa sýnt að *IceTagger* nær 91,5% nákvæmni við mörkun sama texta og notaður var við prófanir á gagnamörkurunum og samkvæmt því gerir *IceTagger* 11,5% færri villur en besti gagnamarkarinn. Nákvæmni *IceTagger* við mörkun óþekktra orða er um 75% [5, 7].

Samsetning (e. combination) markara skilar oft meiri nákvæmni en fæst með einstökum mörkurum. Ástæðan er sú að mismunandi markarar hafa tilhneigingu til að gera ólíkar villur og þennan mismun er hægt að nýta til að ná meiri nákvæmni. Ein samsetningaraðferð er *einföld kosning* (e. simple voting). Í henni eru mismunandi markarar látnir greiða atkvæði með marki fyrir sérhvert orð og síðan er það mark valið sem hlýtur flest atkvæði. Með því að setja saman *IceTagger* og fjóra mismunandi gagnamarkara – og beita einfaldri kosningu – hefur tekist að ná um 93,5% nákvæmni við mörkun íslensks texta [7].

Hlutaþáttari

Markmið með vélrænni setningagreiningu eða *þáttun* (e. parsing) er að greina formgerð setninga og tengsl einstakra hluta þeirra. Þáttari er forrit sem framkvæmir setninga-greiningu. Inntak í þáttara er í flestum tilvikum í formi markaðra setninga og úttakið er lýsing á formgerð þeirra og fyrrgreindum tengslum.

Setningagreiningu er oftast skipt í tvo yfirflokka. Annars vegar er um að ræða *fulla þáttun* (e. full parsing), þar sem búið er til fullkomið þáttunartré (e. parse tree) fyrir sérhverja setningu, og hins vegar *hlutaþáttun* (e. shallow parsing) þar sem setningar eru greindar í setningarhluta án þess að krafist sé að sérhver hluti passi inn í fullkomið þáttunartré.

Höfundur hefur þróað svokallaðan *stigvaxandi* (e. incremental) hlutaþáttara, *IceParser*, fyrir íslenskan texta sem byggir á endanlegum stöðuaðferðum[3] [8]. Þáttarinn

samanstendur af röð af *stöðuferjöldum* (e. finite-state transducers) sem er skipt upp í tvær einingar. Sú fyrri sér um greiningu setningarliða og sú síðari um greiningu setningafræðilegra hlutverka. Í setningarliðaeiningunni sér eitt ferjald um greiningu atvikslíða, annað um greiningu lýsingarorðsliða, hið þriðja um greiningu nafnliða o.s.frv. Í seinni einingunni sér eitt ferjald um greiningu frumlaga, annað um greiningu sagnfyllinga, hið þriðja um greiningu andlaga o.s.frv.

Sérhvert stöðuferjald setur merki inn í markaðan textann sem táknar upphaf og lok tiltekinna setningarliða eða setningafræðilegra hlutverka. Ferjöldin leita að hlutstrengjum í inntakstextanum, sem merkja skal, með því að nota safn af setningafræðilegum mynstrum sem skilgreind eru með *reglulegum segðum* (e. regular expressions).

Nákvæma lýsingu á greiningaratriðum hlutaþáttarans má finna í svokölluðu *þáttunarskema* (e. annotation scheme) sem var búið til áður en þáttarinn var þróaður [9]. Lítum t.d. á úttakið úr setningarliðaeiningunni fyrir mörkuðu setninguna úr síðasta kafla:

[NP Hlutverk nhen **NP] [NP** markara nkee **NP] [VPb** er sfg3en **VPb] [VPi** að cn greina sng **VPi] [NP** sérhvert foheo orð nheo **NP] [PP** í aþ **[NP** texta nkeþ **NP] PP]**
[PP í ao **[NPs [NP** orðflokk nkeo **NP] [CP** og c **CP] [NP [AP** beygingarleg lhfosf **AP]** einkenni nhfo **NP] NPs] PP]**

Setningin hefur hér verið bútuð niður í einstaka setningarliði, eins og nafnliði ([NP ... NP]), sagnliði ([VPx ... VPx]), forsetningarliði ([PP ... PP]) og lýsingarorðslið ([AP ... AP]). Úttakið úr þessari einingu er síðan sent sem inntak inn í eininguna sem greinir setningafræðileg hlutverk. Niðurstaðan er:

{*SUBJ> [NP Hlutverk nhen NP] **{*QUAL** [NP markara nkee NP] ***QUAL} *SUBJ>}** [VPb er sfg3en VPb] [VPi að cn greina sng VPi] **{*OBJ<** [NP sérhvert foheo orð nheo NP] ***OBJ<}** [PP í aþ [NP texta nkeþ NP] PP] [PP í ao [NPs [NP orðflokk nkeo NP] [CP og c CP] [NP [AP beygingarleg lhfosf AP] einkenni nhfo NP] NPs] PP]

Þessi greining sýnir *i)* að nafnliðirnir tveir [NP Hlutverk nhen NP] [NP markara nkee NP] eru frumlagið ({*SUBJ> ... *SUBJ>}) í setningunni (örin merkir að tilheyrandi sögn birtist hægra megin við frumlagið); *ii)* að nafnliðurinn [NP markara nkee NP] er eignarfallseinkunn ({*QUAL ... *QUAL}); *iii)* að nafnliðurinn [NP sérhvert foheo orð nheo NP] er andlag ({*OBJ< ... *OBJ<}) sagnliðarins [VPi að cn greina sng VPi].

Árangur í setningagreiningu er oft mældur með tveimur stærðum. Annars vegar með nákvæmni (e. precision) = *fjöldi réttra liða í úttaki þáttara / heildarfjölda liða í úttaki þáttara*, og hins vegar með griphlutfalli (e. recall) = *fjöldi réttra liða í úttaki þáttara / heildarfjölda liða í viðmiðunarmálheild*. Viðmiðunarmálheild (e. gold standard) er málheild sem hefur verið rétt setningagreind. Nákvæmni segir þá til um hversu hátt hlutfall af þeim liðum, sem þáttarinn stingur upp á, er í raun rétt og griphlutfall segir til um hversu hátt hlutfall af liðunum kemur fyrir í viðmiðunarmálheildinni. Í þeim tilgangi að birta aðeins eina stærð fyrir mat á árangri þáttara er jafnframt oft notuð stærðin *F-measure* = 2*nákvæmni*griphlutfall / (nákvæmni + griphlutfall).

Prófanir hafa sýnt að *IceParser* nær 96,7% *F-measure* fyrir alla setningarliði í heild sinni en t.d. 95,1% fyrir lýsingarorðsliði, 96,8% fyrir nafnliði og 99,2% fyrir sagnliði [8]. Þessar tölur eru sambærilegar við árangur þáttara fyrir skyld tungumál. Hér ber að nefna að tölurnar eru miðaðar við að inntakið í *IceParser* sé rétt markaður texti. Þegar um ómarkaðan texta er að ræða þá er t.d. hægt að marka hann fyrst með *IceTagger* áður en hann er þáttaður. Við það lækkar *F-measure* hins vegar fyrir alla setningarliði úr 96,7% í 91,9%.

Lokaorð

Í þessari grein hefur verið fjallað um markara og hlutaþáttara fyrir íslenskan texta. Báðar einingarnar eru grunneiningar fyrir ýmiss konar máltæknikerfi og þess vegna er mikilvægt að halda áfram að þróa þær og bæta. Sú þróun á sér m.a. stað í rannsóknarverkefninu „Aukin mörkunarnákvæmni íslensks texta“ sem fékk styrk hjá Rannsóknasjóði Rannís árið 2007.

IceTagger og *IceParser* er hægt að prófa á vefsíðunni: http://nlp.ru.is/icenlp.htm.

Heimildir

[1] Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson (1999). *Tungutækni: Skýrsla starfshóps*. Menntamálaráðuneytið, Reykjavík.

[2] S. Krauwer (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*. Moskva.

[3] Jörgen Pind (ed.), Friðrik Magnússon og Stefán Briem (1991). *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

[4] Sigrún Helgadóttir (2007). Mörkun íslensks texta. *Orð og tunga* 9:75–107.

[5] Hrafn Loftsson (2007). Tagging Icelandic Text using a Linguistic and a Statistical Tagger. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the ACL*. Rochester, NY.

[6] Hrafn Loftsson (2006). Tagging a morphologically complex language using heuristics. In T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala (eds.), *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Proceedings*. Turku.

[7] Hrafn Loftsson (2006). Tagging Icelandic text: An experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, 40(2): 175–181.

[8] Hrafn Loftsson og Eiríkur Rögnvaldsson (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics, NODALIDA-2007*. Tartu.

[9] Hrafn Loftsson og Eiríkur Rögnvaldsson (2006) A shallow syntactic annotation scheme for Icelandic text. Technical Report RUTR-SSE06004. Department of Computer Science, Reykjavik University.

[1] Frekari skýringar á markamenginu má finna í [4].

[2] Allar tölur sem hér eru birtar í tengslum við nákvæmni í mörkun eða þáttun íslensks texta byggja á prófunargögnum sem fengin eru úr textasafni *Íslenskrar orðtíðnibókar*.

[3] Hlutaþáttarinn var þróaður í samvinnu við Eirík Rögnvaldsson, prófessor við Háskóla Íslands. Rannsóknarverkefnið bar heitið „Hlutaþáttun íslensks texta“ og var styrkt af Rannsóknasjóði Rannís, 2006.