Jón Friðrik Daðason og Kristín Bjarnadóttir

Kvistur: Vélræn stofnhlutagreining samsettra orða

1 Inngangur

Samsetning orða er mjög virk í íslensku og fjöldi nýrra orða er ótakmarkaður. Þetta hefur vandkvæði í för með sér í máltækni þar sem máltæknitól eiga í vandræðum með að greina orð sem ekki er að finna í gagnasöfnunum sem þau nota. Þessi orð eru óþekkt, í máltæknilegu samhengi, og greining þeirra skilar mun verri niðurstöðu en greining þekktra orða. Þetta á við um alla sjálfvirka málfræðigreiningu á textum. Röng greining óþekktra orða getur valdið því að texti verði að hluta gagnslaus eða gagnslítill, t.d. þannig að leitarvélar og önnur tól finni ekki orðmyndir sem eru rangt greindar og þær skili sér því ekki með öðrum dæmum um orðið. Á þennan hátt verður aðgengi að textunum takmarkað. Hægt er að bæta skilvirkni máltæknitóla fyrir íslensku verulega með því að skipta samsettum orðum í þekkta orðhluta með samanburði við stór gagnasöfn um orðaforðann, t.d. *Beygingarlýsingu íslensks nútímamáls* með yfir 5,8 milljónum beygingarmynda (Kristín Bjarnadóttir 2012) og gríðarlega stór textasöfn, t.d. *Íslenskan orðasjóð* (Erla Hallsteinsdóttir o.fl. 2007) sem er málheild með textum af íslenskum vefsíðum árin 2005 og 2010, með yfir hálfum milljarði lesmálsorða.

Í máltækniverkefnum hefur almennt reynst vel að skipta óþekktum samsettum orðum í orðhluta, t.d. í vélrænum þýðingum (Brown 2002, Koehn & Knight 2003, Alfonseca o.fl. 2008), upplýsingaheimt (e. *information retrieval*) (Hedlund o.fl. 2001, Braschler o.fl. 2003) og

talgreiningu (Adda-Decker o.fl. 2000). Sérstaða aðferðarinnar sem beitt er í Kvisti, orðskiptingartólinu sem hér er lýst, er fólgin í því að búa til stofnhlutatré en annars staðar er látið duga að skipta samsettu orðunum í línulega röð af ósamsettum orðum.

Kvistur nýtir stofnhlutagreiningu til þess að finna líklegustu skiptingu orða í orðhluta og er eina tólið sem höfundar þessarar greinar vita um þar sem þeirri aðferð er beitt.¹ Í stofnhlutagreiningunni er gengið út frá tvígreiningu í orðmyndunargreiningunni og samsettu orði er þar með alltaf skipt í fyrri hluta (ákvæði) og síðari hluta, sem er beygingarlegur haus orðsins og oftast einnig merkingarlegur haus (Kristín Bjarnadóttir 2005). Orðmyndunarreglurnar eru endurkvæmar þannig að samsett orð geta einnig verið inntak í þær, bæði sem ákvæði og haus, og liðirnir geta verið úr öllum opnum orðflokkum. Samsett nafnorð þar sem fyrri hlutinn er einnig nafnorð eru algengust og þar eru reglurnar virkastar. Endurkvæmni reglnanna er einnig mest í nafnorðunum og orðhlutarnir þar með flestir og því er þörfin fyrir orðskiptingartól brýnust þar. Samsett nafnorð verða því notuð sem dæmi í greininni en Kvistur reynist þó jafn vel fyrir samsett orð úr öllum orðflokkum.

Kvistur byggist á tölfræðilegu líkani sem metur líkurnar á því að tveir tilteknir orðhlutar geti myndað samsett orð saman. Líkanið er búið til með því að nota stórt safn af fullgreindum íslenskum samsettum orðum sem unnin voru í höndunum (Kristín Bjarnadóttir 1996). Forritið er þjálfað á safninu og síðan notað til þess að finna líklegasta stofnhlutatré (tvíundatré) fyrir samsett orð. Líkurnar á stofnhlutagreiningu orða þar sem einhverjir orðhlutar eru óþekktir eru metnar með því að skoða sambærilegar orðgerðir í öðrum samsettum orðum. Stofnhlutagerðin er notuð til þess að skipta óþekktum samsettum orðum í tvennt og eins langt gengið í skiptingunni og þörf krefur hverju sinni. Oft er nóg að skipta á meginskilum í orði en skipta má orðum í fíngreindari tré ef þörf krefur. Við nefningu og mörkun² er yfirleitt nóg að síðasti orðliðurinn sé rétt greindur þar sem hann er beygingarlegur haus og orðflokkurinn er þar með ljós. Í

¹ Vinna við Kvist tengist öðrum máltækniverkefnum á Stofnun Árna Magnússonar í íslenskum fræðum. Fyrsta gerð Kvists var unnin af Jóni Friðriki Daðasyni undir handleiðslu Eiríks Rögnvaldssonar prófessors í námskeiði í máltækni. Sigrún Helgadóttir og Sven Þ. Sigurðsson veittu ráðgjöf á öllum stigum verksins. Hafi þau öll bestu þakkir fyrir. Ónefndir ritrýnendur *Orðs og tungu* eiga einnig þakkir skildar fyrir mjög góðar athugasemdir.

² *Nefning* (lemmun) er það að finna nefnimynd (uppfléttimynd) lesmálsorðs; *mörkun* er það að greina málfraeðiatriði, s.s. orðflokk og beygingarþætti.

öðrum verkefnum skiptir sjálft orðhlutatréd máli, t.d. við að ákvarða merkingu orðs í vélrænni þýðingu.

Efnisskipan greinarinnar er á þessa leið: Í 2. kafla er stutt lýsing á samsettum orðum í íslensku, m.t.t. þess sem máli skiptir í orðhlutagreiningunni, í 3. kafla er fjallað um aðferðir við vélræna orðskiptingu, í 4. kafla er aðferðafræði við orðskiptinguna í Kvisti lýst og í 5. kafla er aðferðin síðan metin með samanburði við handgreind gögn. Í 6. kafla er fjallað um kostina sem felast í því að geta ákveðið hversu fíngreind skiptingin á að vera, m.t.t. verkefna. Niðurstöður og hugleiðingar um áframhaldandi vinnu eru í 7. kafla.

2 Yfirlit um samsett orð í íslensku

Reglur um samsett orð í íslensku eru gríðarlega virkar og hlutfall samsettra orða er hátt í íslenskum textum. Í *Markaðri íslenskri málheild* (MÍM) (Sigrún Helgadóttir o.fl. 2012) eru t.d. 75,4% allra þekktra orðmynda samsett orð en alls eru u.þ.b. 25 milljónir lesmálsorða í MÍM. Reglur um samsett orð eru endurkvæmar, þannig að í margsamsettum orðum geta bæði ákvæði og hausar verið samsett orð og stofnhlutagreining getur verið tví- eða margræð.³ Í orðinu *skólabókasafn* koma t.d. tvær greiningar til álita, *skóla+bókasafn* og *skólabóka+safn*, en fyrri greiningin ('bókasafn í skóla') er allsráðandi í textum og lesgerð (lexíkaliseruð) í skilningi orðabókarfræða (Bergenholtz o.fl. 1997:167) þó að síðari greiningin ('safn skólabóka') sé einnig rétt samkvæmt orðmyndunarreglum. Fræðilega eru ekki hömlur á fjölda stofnhluta í samsettum orðum en í reynd eru orð úr fleiri en sex liðum ekki algeng (Magnús Snædal 1992). Slík orð koma þó fyrir í gögnunum sem Kvistur hefur verið þjálfður á, t.d. í textum úr Íslenskum orðasjóði. Þar er m.a. að finna orðið *Alþjóðadýraheilbrigðismálastofnun* þar sem liðirnir eru sjö, sjá *Mynd 1*.

³ Reglur um afleiðslu og samsetningu eru einnig endurkvæmar, sbr. samsetta orðið *mal+bik*, afleiddu sögnina *malbik+a*, afleidda nafnorðið *malbik+un* og samsetta orðið *malbikunarvél* (sjá Kristínu Bjarnadóttu 2005). Kvistur greinir ekki afleiðslu enn sem komið er. Í þessari grein á orðið *stofnhluti* aðeins við um samsetningarliði í samsettum orðum, ekki viðskeyti.

Mynd 1. Dæmi um tvíundatré fyrir margsamsett orð.

Hlutfall ofurlangra orða af þessu tagi er lágt í *Íslenskum orðasjóði* en í *Töflu 1* eru tölur um fjölda liða í þekktum orðum þar. Í fremri talnadálki er fjöldi einstakra orðmynda en í síðari talnadálki eru samantlagðar tíðnitölur fyrir sömu orðmyndir (lesmálsorð) eða stafastrengi.

Fjöldi liða	Fjöldi orðmynda	Fjöldi lesmálsorða
1 (grunnorð)	146.000	427.188.333
2	509.497	70.816.731
3	137.804	10.149.391
4	18.698	1.130.527
5	1.188	70.129
6	17	92
7	1	10
Óþekkt orð	3.974.811	34.010.620
Alls	4.788.016	543.365.833

Tafla 1. Fjöldi stofnhluta í þekktum orðum í *Íslenskum orðasjóði*.

Í tölum fyrir óþekkt orð eru bókstafarunur sem geta verið með einu bandstriki en öllum strengjum með tölustöfum og öðrum táknum er sleppt. Til óþekkta orðanna í *Íslenskum orðasjóði* teljast eiginlegar óþekktar orðmyndir (t.d. *jólaórói*, *kolvetnissgas*, *koltvísýringslosun*, *kolmunnavertíð*), auk skammstafana (t.d. *kl.*, *m.a.*, *t.d.*, *NATO*), útlendra orða (t.d. *the*, *and*, *york*, *quixotic*), villna (t.d. *mart*, *kanski*, *mannn*) o.fl. Takmarkið er að Kvistur fækki óþekktum orðum með því að greina samsett orð í þekkta orðhluta þannig að orðið sem heild sé ekki lengur órátt. Möguleikar á margræðni í stofnhlutagreiningu (og þar með

merkingu) aukast eftir því sem liðum fjölgar en þar kemur samanburður við önnur samsett orð, þar sem sömu liðir koma fyrir, að gagni við að ákvarða stofnhlutagreininguna (sjá *Mynd 2* í kafla 4.2 hér á eftir).

Fjölbreytileiki í samsetningarmyndum orða veldur vandkvæðum í vélrænni greiningu á samsettum orðum í íslensku. Til dæmis birtast nafnorð sem ákvæði í samsettum orðum oftast sem stofn eða eignarfall eintölu eða fleirtölu en bandstafir og þágufallsmyndir⁴ finnast líka þótt það sé mun sjaldgæfara (Kristín Bjarnadóttir 2002). Nokkur fjölbreytileiki er einnig í samsetningarmyndum lýsingarorða sem ákvæða, sbr. *góðmenni* (stofn), *lítilsvirðing* (eignarfall), *litliputti* (fallbeygður fyrri hluti) og *betrstofa* (miðstig). Þá eru aðrir orðflokkar ónefndir, auk setningarliða sem einnig geta verið fyrri hlutar í samsettum orðum, t.d. í orðunum *millipínganefnd* og *Góðrarvonarhöfði*.

Föst samsetning	Stofn	<i>handbolti, fótbolti</i>
Laus samsetning	Eignarfall eintölu	<i>handarbak, fótarbein</i>
	Eignarfall fleirtölu	<i>handapat, fótatak</i>
	Þágufall eintölu	<i>hendiveifa</i>
	Þágufall fleirtölu	<i>fótumtroðsla</i>
Bandstafir		<i>labbitúr, ráðunautur</i>

Tafla 2. Form nafnorðsákvæða í samsettum orðum.

Valið á milli fastrar og lausrar samsetningar og milli eignarfalls eintölu og fleirtölu nafnorða í fyrri hluta samsettra orða virðist vera tilviljun eins og sjá má í *Töflu 2*. Valið er þó ekki frjálst (Kristín Bjarnadóttir 1995). E.t.v. má segja að valið sé leskert, þannig að það er orðbundið hver háttur er hafður á hverju sinni, sjá *Töflu 3*.⁵

⁴ Nafnorð með þágufallsákvæði eru mjög fá í *Ritmálssafni Orðabókar Háskólans* en þar eru dálítið fleiri lýsingarorð og lýsingarhættir þátíðar, auk örfárra staðfesta dæma um sagnir í germynd (Kristín Bjarnadóttir 2000a; Kristín Bjarnadóttir 2000b). Um nafnorðið *hendiveifa* eru 9 dæmi í *Ritmálssafni*, t.d. *Konan sendist sem hendiveifa upp stigann*

⁵ Samkvæmt orðmyndunarreglum ættu allar orðmyndir í *Töflu 3* að vera tækar en orðmyndirnar sjálfar virðast vera lesgerðar (lexikalíseraðar) eins og sést af tíðnitölum, t.d. úr *Íslenskum orðasjóði*: *bókbúð* (1), *bókarbúð* (0), *bókabúð* (825). Fleiri dæmi finnast um *bókabúð* (165 þúsund) og *bókbúð* (63) í Google en mörg dæmin um *bókbúð* þar eru augljóslega villur, t.d. *Bókbúð Máls og menningar* og *Bókbúð Braga*, enda eru leitarniðurstöður þaðan ekki mjög traust heimild um nákvæma tíðni þótt þær geti gefið góðar viðbendingar. Niðurstaðan er því að stjörnumerktu orðin séu rétt mynduð; þau eru bara ekki notuð.

Stofn	Ef.et.	Ef.ft.
<i>bóksala</i>	<i>*bókarsala</i>	<i>*bókasala</i>
<i>*bókkápa</i>	<i>bókarkápa</i>	<i>*bókakápa</i>
<i>*bókbúð</i>	<i>*bókarbúð</i>	<i>bókabúð</i>

Tafla 3. Orðbundið form ákvæðis í samsettum orðum.

Parna virðist einfaldlega hafa orðið til hefð sem ákveður hvort fyrri hluti er stofn, eignarfall eintölu eða fleirtölu, og tengsl á milli forms og merkingar fyrri hlutans eru ekki endilega rökrétt. Þetta má sjá í samsettu orðunum *barnsmeðlag* og *barnslífeyrir* en þar helst talan á fyrri hlutanum hvort sem greitt er með einu eða fleiri börnum.⁶ Í sumum samsettum orðum er tala ákvæðisins merkingarbær, t.d. í orðunum *bróðursonur*, *bróðursynir*, *bræðrasynir*. (Orðmyndin *bræðrasonur* finnst ekki.) Þessi merkingarlegi greinarmunur á eintölu og fleirtölu er samt sem áður fremur sjaldgæfur eftir því sem best er vitað en þarna er mikið efni órannsakað. Með tíðnirannsóknum á stóru gagnasafni ætti þetta að skýrast; með því að skoða orð eins og *mannsnafn*, *mannanöfn*, *bókarkápa*, *bókakápur* o.s.frv. og þar getur Kvistur gegnt lykilhlutverki. Þessi fjölbreytni í formi ákvæða getur orðið til þess að tengsl á milli samsetningarmynda einstakra orða glatist vegna þess að hvert afbrigði er meðhöndlað sem sjálfstæð eining, án tengsla við nefnimyndina. Sama á reyndar við í öllu beygingarkerfinu enda er það tiltölulega flókið og markamengið stórt en mörk í íslensku eru á sjöunda hundrað (Hrafn Loftsson o.fl. 2009). Það liggur í hlutarins eðli að dæmi um hverja beygingarmynd verða færri eftir því sem beygingarkerfið verður flóknara og um sumar hugsanlegar beygingarmyndir finnast engin dæmi. Fjölbreytni af þessu tagi getur valdið gagnaskorti þegar dæmi um einstakar orðmyndir verða svo fá að tíðnitölur verða illa marktækar. Þetta er vel þekkt vandamál í máltækni og á einnig við um mörkun og nefningu á orðmyndum í íslenskum textum. Nefning, þ.e. tenging allra beygingarmynda orðs við nefnimynd sína, er því lykilatriði í greiningu á íslenskum orðum og það gildir einnig um samsett orð og orðhluta í þeim.

Beygingarmyndir í íslensku eru ekki aðeins margar heldur eru þær einnig talsvert margræðar þegar þær eru skoðaðar án tengsla við önnur orð (Kristín Bjarnadóttir 2012). Þetta á einnig við um beygingarmyndir sem samsetningarliði, sbr. dæmið *andahyggjumaður* en

⁶ Dæmi um orðið *barnsmeðlag* finnast líka, um greiðslu með einu eða fleiri börnum, en þau eru mun færri en dæmi um *barnsmeðlag*. Höfundar hafa ekki fundið dæmi um orðmyndina *barnslífeyrir*.

fyrsti stofnhlutinn, *anda*, er annaðhvort eignarfall fleirtölu af *önd* eða *andi*. Hinir liðirnir tveir í orðinu eru einræðir, a.m.k. í nefningu (lemmun), þ.e. *hyggja* og *maður*. Meginskipting orðsins er *anda-hyggju+maður* eins og sést af samanburði við önnur orð í gagnasöfn-um. Í *Ritmálssafni Orðabókar Háskólans* eru t.d. 63 samsett orð með hausinn *maður* þar sem ákvæðið er samsett orð með hausinn *hyggja*.⁷ Í langflestum þessara orða er samsetta ákvæðið einnig til sem upp-flettiorð í *Ritmálssafni* (*efnishyggja*, *rökhyggja*, *andahyggja* o.s.frv.) en orðið *hyggjumaður* er aðeins að finna í einu dæmi frá 16. öld: *Hliodur* er *hyggiu* *madr*.

efnishyggjumaður	44	efnishyggja	68
einstaklingshyggjumaður	21	einstaklingshyggja	40
undirhyggjumaður	18	undirhyggja	15
hughyggjumaður	19	hughyggja	30
rökhyggjumaður	17	rökhyggja	7
djúphyggjumaður	18	djúphyggja	6
vélhyggjumaður	1	vélhyggja	10
andahyggjumaður	1	andahyggja	6
fortíðarhyggjumaður	1	fortíðarhyggja	0
þrönghyggjumaður	1	þrönghyggja	0
afstæðishyggjumaður	0	afstæðishyggja	8

Tafla 4. Fjöldi dæma um nokkur orðapör í *Ritmálssafni* með *hyggja* sem ákvæðishaus og haus.

Það eru einmitt vensl af sama tagi og þau sem koma fram í *Töflu 4* sem Kvistur notar til þess að finna rétt orðhlutatré orða með óþekktum orðhlutum, sjá kafla 4.2 hér á eftir. Tenging orðhlutanna við nefni-myndir sínar er hins vegar verkefni sem bíður betri tíma en gagna-safnið, sem nú er notað, ætti að duga til þeirrar vinnu líka.

3 Aðferðafræði við orðskiptingu

Kvistur er eina orðskiptingartólið sem höfundar vita um þar sem stofnhlutagreining er nýtt til þess að finna líklegustu skiptingu orðs

⁷ Í *Töflu 4* eru 11 af 63 samsettum orðum sem enda á *-hyggjumaður* í *Ritmálssafni*. Leitað var bæði í uppflöttiorðum og orðum í dæmum. Samsett nafnorð, sem enda á *hyggja*, eru 229 í *Ritmálssafni*. Nokkur hending er hvort bæði orðin í hverju pari skila sér í *Ritmálssafni* og stundum er auðvelt að finna dæmi á öðrum stöðum til þess að fylla í götin. Orðið *þrönghyggja* finnst t.d. 24 sinnum á *timarit.is* en ekki í *Ritmálssafni*.

í orðhluta. Tæknilega eru ýmsar aðferðir viðhafðar við orðskiptingu en þeim má skipta í tvo flokka.

Sá fyrri samanstendur af aðferðum sem styðjast við þýðingarorðabækur á milli viðfangsmálsins (þar sem reglur um samsett orð eru virkar) og markmáls þar sem orðhlutaskil samsettra orða eru ótvíræð, t.d. í ensku þar sem samsett orð eru gjarnan rituð með orðabilum (t.d. *team spirit*, *lawn tennis*). Slíkar orðabækur má t.d. vinna upp úr samhliða málheildum eins og Brown (2002) gerir fyrir þýsku og ensku. Skipting þar sem þýðingin á orðhlutunum stemmir við þýðinguna á samsetta orðinu sjálfu verður þarna fyrir valinu. Í tilbúnu dæmi úr íslensku er valið á milli *lið*_(no.hk.stofn) + *sandi*_(no.kk.) og *liðs*_(no.hk.ef.et.) + *andi*_(no.kk.) en þar yrði seinni skiptingin fyrir valinu, sbr. ensku þýðinguna 'team spirit'. Gögn af þessu tagi eru ekki tiltæk fyrir íslensku og aðferðin er því ekki nothæf.

Aðrar aðferðir skipta samsettum orðum í þekkta orðhluta og meta líkindi hvernar skiptingar út frá tíðni orðhlutanna. Koehn og Knight (2003) finna allar runur af þekktum sjálfstæðum orðum sem saman geta myndað tiltekið samsett orð og meta líkurnar á hverri skiptingu út frá margfeldismeðaltali (e. *geometric mean*) orðtíðni hlutanna eins og sjá má í *Töflu 5* fyrir orðið *þingstörfunum*.

Ákvæði	Orðtíðni	Haus	Orðtíðni	Margfeldismeðaltal
þing _(no.hk.stofn)	28.846	störfunum _(no.hk.pg.ft.)	297	$(28846 \cdot 297)^{1/2} = 2927,0$
þings _(no.hk.ef.et.)	4.688	törfunum _(no.kk.pg.ft.)	18	$(4688 \cdot 18)^{1/2} = 290,5$

Tafla 5. Dæmi um tíðni orðhluta í samsettu orði ásamt margfeldismeðaltali.

Með samanburði á tíðni í textasafni er ljóst að *störf* eru algengari á *þingi* en *tarfar* og því er fyrri greiningin valin.

Schiller (2005) útfærir svipaða aðferð en brýtur samsett orð niður í runu ákvæða og hausa sem koma fyrir í stóru þjálfunarsafni af fullgreindum þýskum samsettum orðum. Líkurnar gagnvart hverjum orðhluta eru metnar út frá þjálfunarsafninu eftir því hversu oft hann kemur fyrir sem ákvæði eða haus, eftir því sem við á.

Þessar aðferðir eiga það sameiginlegt að þær skipta samsettum orðum aðeins í runur af orðhlutum. Kosturinn við stofnhlutatré samsetts orðs er að þar er hægt að nýta eins fíngreinda skiptingu á orðinu og þörf krefur. Skiptingin *fjármála+ráðherra* myndi henta vel fyrir vélræna þýðingu (að því gefnu að báðir hlutarnir séu þekktir) en með tilliti til beygingar er hentugra að greina beygingarlegan haus (*herra*) sem nægir til þess að beygingarflokkur o.þ.h. sé ljós.

4 Aðferðin

Aðferðin í Kvisti skiptist í eftirfarandi skref:

1. Samsettu orði er skipt upp í stystu mögulegu röð af grunnorðum, kvisti:
 - a. fjármálaráðherra → fjár+mála+ráð+herra
2. Öll möguleg grannorðapör eru skoðuð:
 - a. fjár+mála
 - b. mála+ráð
 - c. ráð+herra
3. Grannorðaparið með hæstu líkurnar er sett saman sem kvistur í tvíundatré:
 - a. fjármála
4. Skref 2 og 3 eru endurtekin þar til aðeins einn liður er eftir fyrir hverja mögulega skiptingu í skrefi 1. Tvíundartréð með hæstar líkur sem heildartré er valið (sjá Myndir 3–5; niðurstaðan er [fjár+mála]+[ráð+herra]).

4.1 Orðskiptingin

Í óbirtum vinnugögnum fyrir *Beygingarlýsingu íslensks nútímamáls* (BÍN) er orðhlutaskipting fyrir u.þ.b. 240 þúsund samsett orð, auk upplýsinga um orðflokk, beygingarflokk og alla þætti beygingargreiningar fyrir allar beygingarmyndir, bæði fyrir grunnorð og samsett orð (Kristín Bjarnadóttir 1996). Kvistur nýtir allar þessar upplýsingar í orðskiptingunni og með þeim er hægt að búa til tvígreind stofnhlutatré fyrir öll samsett orð í BÍN. Alls eru 169 þúsund beygingarmyndir grunnorða í samsettu orðunum í BÍN og u.þ.b. 40 þúsund þeirra koma fram sem fyrri hlutar (ákvæði) samsettra orða og 146 þúsund koma fyrir sem síðari hlutar (hausar). Hausarnir eru fleiri en ákvæðin þar sem formið er fjölbreytilegra þar vegna þess að samsettu orðin geta (hugsanlega) komið fyrir í öllum beygingarmyndum. Hausarnir beygjast m.ö.o. en það gera ákvæðin ekki.⁸ Haus í samsettu

⁸ Undantekningin eru beygð lýsingarorð sem ákvæði, t.d. *litliputti* og *hæstiréttur*. Fyrri hlutar í samsettum mannanöfnum beygjast líka í máli manna, t.d. *Annalísa* (þf. *Ömmulísu*), þó að skoðanir séu skiptar um það hvort svo eigi að vera (Mannanafnanefnd 2005).

orði í íslensku hefur (nánast alltaf) sömu beygingarþætti og samsetta orðið (t.d. *bók*_(no.kvk.) og *orðabók*_(no.kvk.)) og haus í samsettu orði (t.d. *völlur* í *flugvöllur*) er hugsanlegt ákvæði (t.d. í orðinu *vallarsvæði*) ef samsetta orðið sjálft (*flugvöllur*) er ákvæði (t.d. í orðinu *flugvallarsvæði*). Skipting samsetts orðs getur því verið runa af hvaða orðum sem er (sem geta verið ákvæði) og einn þekktur haus.

4.2 Grannorðatenging

Líkur á því að tvö grannorð geti myndað samsettan stofnhluta eru metnar með samanburði við greinda stofnhluta sem koma fyrir í gögnunum um samsett orð í BÍN og tíðni grannorðanna í *Íslenskum orðasjóði* en þar eru u.þ.b. 550 milljónir lesmálsorða af íslenskum vefsíðum. Líkurnar á þekktum pörum eru reiknaðar svona:

$$P(\text{ákvæði} + \text{haus}) = \frac{\text{tíðni}(\text{ákvæði} + \text{haus})}{N}$$

þar sem *tíðni(ákvæði+haus)* er fjöldi samsettra orða í málheild þar sem ákvæðið og hausinn mynda orðhluta saman og *N* er heildarfjöldi samsettra orða í málheildinni.

Líkurnar á því að tvö tiltekin grannorð myndi saman óþekkt samsett orð eru metnar út frá líkindum hvors orðhlutans fyrir sig. Ef *heilsu+vara* væri óþekkt samsetning væru líkurnar metnar með því að margfalda líkurnar á ákvæðinu *heilsu* (með hvaða haus sem er) og líkurnar á hausnum *vara* (með hvaða ákvæði sem er), þ.e.

$$P(\text{heilsu} + \text{vara}) = \frac{\text{tíðni}(\text{heilsu} + *)}{N} * \frac{\text{tíðni}(* + \text{vara})}{N}$$

þar sem *** táknar hvaða orðhluta sem er og *tíðni(heilsu+*)* er fjöldi samsettra orða þar sem *heilsu* kemur fyrir sem ákvæði. Strenginn *heilsu+** mætti kalla *snið* fyrir hvaða samsett orð sem er með ákvæðið *heilsu*, og **+vara* *snið* fyrir öll samsett orð með hausinn *vara*.

Flóknari *snið* eru búin til úr samstæðum samsettum orðum, þ.e. orðum með hliðstæða stofngerð. Með því að þjálfá Kvist á stóru safni af stofnhlutagreindum orðum er hægt að draga ályktanir um skyldleikann. Orðin *fjármálaráðherra*, *dómsmálaráðherra* og fjölmörg önnur orð hafa t.d. sams konar stofnhlutagreiningu; hausinn er *ráðherra* og ákvæðin eru samsett orð þar sem hausinn er *mála*, sjá *Mynd 2*.

Mynd 2. Snið fyrir orð eins og fjármálaráðherra og dómsmálaráðherra.

Þegar búa á til stofnhlutatré fyrir samsetta orðið *fjármálaráðherra* er fyrsta skrefið að finna allar mögulegar runur af grunnorðum sem það getur samanstðið af (sbr. kafla 4.1). Í þessu tilviki er eina mögulega runan *fjár+mála+ráð+herra*. Þá koma þrír samsettir stofnhlutar til greina: *fjár+mála* ($P=0,00122$), *mála+ráð* ($P=0,00000$) og *ráð+herra* ($P=0,00291$).

Mynd 3. Þrjár mögulegar samsetningar grunnorðapara í orðinu *fjármálaráðherra*.

Þar sem grunnorðin *ráð* og *herra* hafa hæstu líkurnar er þeim skeytt saman. Þá bætist við ein möguleg samsetning í viðbót, *mála+ráðherra* ($P=0,00000$).

Mynd 4. Tvær mögulegar skiptingar eftir að líklegasta parinu, *ráð* og *herra*, er skeytt saman.

Grannorðunum *fjár* og *mála* er skeytt saman og eftir stendur ein möguleg samsetning: *fjármála+ráðherra* ($P=0,00019$).

Mynd 5. Ein möguleg skipting, *fjármálaráðherra*, eftir að *fjár* og *mála* er skeytt saman.

Þar sem engin önnur samsetning kemur til greina er grannorðunum *fjármála* og *ráðherra* skeytt saman og þar með er búið að mynda stofnhlutatré fyrir orðið *fjármálaráðherra*.

Sniðið fyrir **málaráðherra* (sjá Mynd 2) dugur til þess að finna stofnhlutagreiningu fyrir óþekkt orð þar sem tréð er eins, t.d. fyrir orð eins og *samgöngumálaráðherra* og *siglingamálaráðherra* (sem eru þekkt) og *spillingarmálaráðherra* (sem ekki er viðtekið).⁹

5 Mat á niðurstöðum

Niðurstöður úr prófun á Kvisti sýna ágætan árangur. Matið var gert með því að keyra Kvist á 6.098 orð úr íslenskum Wikipedia-greinum en þar af voru 3.319 samsett orð. Orðin voru jafnframt handgreind til samanburðar. Prófunin fór fram í tveimur hlutum. Fyrst var nákvæmni Kvists í greiningu á gerð samsettra orða metin en síðan var niðurstaða fyrir öll orðin í safninu metin.

Fjöldi liða	Orðafjöldi	Eindir	Tvígreining	Tré
2	2.709	99,5%	99,6%	99,5%
3	513	93,2%	97,9%	91,8%
4+	97	91,8%	96,9%	88,7%
Alls	3.319	98,3%	99,2%	98,0%

Tafla 6. Mat á greiningu samsettra orða.

⁹ Orðið *spillingarmálaráðherra* var uppástunga Google um leiðréttingu á orðinu *siglingamálaráðherra* sem leitarvélín þekkti ekki.

Í *Töflu 6* sést hlutfall réttra skiptinga í grunnorð í 3. dálki (eindir), hlutfall réttra meginskila í 4. dálki (tvígreining) og hlutfall réttra tvíundatrjáa í 5. dálki (tré).

Heildarniðurstaða úr Kvisti er metin með tilliti til nákvæmni (e. *precision*), heimtar (e. *recall*) og hittni (e. *accuracy*). Nákvæmni er hlutfall allra orða sem Kvistur hefur skipt sem skila sér sem rétt tvíundatré fyrir samsett orð. Heimt er hlutfall allra samsettra orða í prófunargögnunum sem hafa verið rétt greind í tvíundatré. Hittni er hlutfall rétt greindra orða, bæði óskiptra orða (grunnorða) og samsettra orða í tvíundatré. Þegar Kvistur er metinn á þennan hátt á öllum prófunar-textanum úr Wikipediu-greinunum er nákvæmnin 97,6%, heimtin 98,0% og hittnin 99,2%.

6 Stofnhlutatré og skipting eftir þörfum

Upplýsingar þær sem felast í sjálfum stofnhlutatrjánum eru mjög mikilvægur afrakstur Kvists. Oft kemur skipting á meginskilum orða að betri notum í máltækniverkefnum en skipting samsettra orða í línulega röð af stofnhlutum. Þetta á bæði við um þekkt orð og óþekkt. Orðið *aðstoðardagskrángerðarmaður* er t.d. þekkt orð (það er í BÍN) en skiptingin $[[[að]+[stoðar]]+[[[dag]+[skrár]]+[gerðar]]+[maður]]$ er mun gagnlegri en *að+stoðar+dag+skrár+gerðar+maður*. Þarna er það stofnhlutagreiningin sjálf sem skiptir máli við leit og gagnanám (e. *data mining*) hvers konar, við þýðingar, við áherslumynstur í talgervingu o.m.fl., þ.e. *dagskrá* og *dagskrángerð*, *dagskrángerðarmaður* og síðan *aðstoðarmaður*.

Í eftirfarandi lista eru nokkur verkefni þar sem aðgangur að öllu orðgerðartrénu kemur að gagni en hér mætti tína margt fleira til.

1. Leit: Öll skil í samsettum orðum geta komið að gagni við leit og gagnanám. Í málfræðirannsóknnum skiptir leit að hausum í orðum máli, t.d. til þess að safna dæmum um grunnorð, ásamt öllum samsetningum af þeim. Án orðhlutaskila er erfitt að gera greinarmun á orðunum *ending* og *sending* og það getur verið hjálplegt að fá einnig niðurstöður fyrir orð eins og *beygingarending* og *flirtöluending* (þ.e. að leita að orðhlutanum *ending* í stofnhlutatrjánum) án þess að með fljóti orð með sama stafastreng þar sem skilin eru önnur, t.d. *heimsending*. Leit að ákvæðum getur einnig skilað viðbótarefni um tiltekin orð sem getur orðið notadrjúgt í gagnaskorti. Ákvæðið *afburða* er t.d. ýmist ritað sem sjálfstætt

orð eða orðhluti: *leikmennirnir eru afburða góðir; afburðagóð veiði í Eystri-Rangá* (dæmi úr MÍM).

2. Nefning og mörkun: Beygingarmyndir grunnorða eru mun oftast tvíráðar en beygingarmyndir samsettra orða og því er erfiðara að finna mörk þeirra og uppflettimyndir. Sama á við um hausa í samsettum orðum. Möguleg mörk fyrir beygingarmyndina *atvinnutækifærum* eru t.d. níu ef beygingarlegi hausinn *færum* er skoðaður en hann getur verið beygingarmynd af nafnorði, lýsingarorði og tveimur sögnum. Orðhlutinn *tækifærum* er hins vegar einræður, þ.e. þágufall fleirtölu af nafnorði. Meginskilin sýna því að beygingarlegi hausinn *færum* er einræður í þessu orði.
3. Þýðingar: Þýðingarforrit geta notað orðskiptingartól til að þýða óþekkt samsett orð. Þá er æskilegast að skipta orðinu í eins fáa þekkta hluta og mögulegt er, t.d. *miðvikudags+morgunn* ‘Wednesday morning’ og athuga hvort orðhlutarnir finnast í tvímálaorðabókinni sem notuð er. Skiptingin *mið+viku+dags+morgunn* ‘mid week day morning’ skilar ekki góðri niðurstöðu þarna.
4. Talgerving: Til að finna áherslumynstur óþekkttra orða er nauðsynlegt að geta greint meginskil samsettra orða. Áhersla á að vera á fyrsta atkvæði og síðan á fyrsta atkvæði á eftir meginskilum. Orðið *skólabókasafn* í merkingunni ‘bókasafn í skóla’ fær þá aukaáherslu á þriðja atkvæði (*skóla+bókasafn*) en í merkingunni ‘safn skólabóka’ væri aukaáherslan á fimmta atkvæði (*skólabóka+safn*).

7 Lokaorð

Kvistur hefur þá sérstöðu meðal orðskiptingartóla að skila tvíundatrjám um samsett orð í stað þess að greina orðin aðeins í orðhluta, án orðgerðar. Upprunalegt markmið með gerð Kvists var að fækka óþekktum orðum í textum og bæta með því árangur í íslenskum máltækniverkefnum. Niðurstaða af prófunum sýnir að gera má ráð fyrir allt að 41,1% fækkun á óþekktum orðum en til glöggvunar á umfanginu eru óþekkt lesmálsorð í *Íslenskum orðasjóði* 34.010.620 af 543.365.833 lesmálsorðum alls (6,3%, sjá *Töflu 1*). Þetta hlutfall virðist e.t.v. ekki mjög hátt en villufjöldinn, sem það getur leitt af sér, er þó verulegur þar sem villa í greiningu uppflettimynda (nefningu) getur t.d. haft keðjuverkandi áhrif. Hugmyndin er að nota Kvist sem hluta af ýmsum máltækniverkefnum, t.d. við nefningu, orðtöku og

villuleiðréttingu og þess er rétt að geta að hann er mjög mikilvægur hluti Skramba, leiðréttingarforrits Jóns Friðriks Daðasonar, sem m.a. er notað til þess að leiðrétta nútímamálstexta, ljóslesna texta og til þess að færa texta frá ýmsum tímum til nútímastafsetningar (Jón Friðrik Daðason o.fl. 2014). Ætlunin er að bæta Kvist til fullvinnslu á formlegri orðgreiningu á lesmálsorðum í texta. Með því að nýta orðhlutana til áframhaldandi vinnslu verður gagnaskortur vonandi minna vanda-mál en sum samsett orð eru mun algengari sem hluti af samsetningum en sem sjálfstæð orð. Í öllu þessu styðja máltæknaverkefnið hvert annað. Kvistur skilar mikilvægu viðbótarefni í BÍN, nefningin batnar og óþekktu orðunum fækkar.

Framtíðin í þróun Kvists felst í möguleikum á tengslum við merkingargreind orðasöfn, t.d. við *Orðanet* Jóns Hilmars Jónssonar og Þórdísar Úlfarsdóttur (Jón Hilmar Jónsson 2012) og *Merkingarbrunn* Önnu Bjarkar Nikulásdóttur (Anna Björk Nikulásdóttir og Matthew Whelpton 2010). Með því að nýta merkingargreiningu og merkingarflokkun sem þar er að finna er t.d. hægt að leggja til atlögu við rannsókn á merkingarvenslum í samsettum orðum. Þá yrði t.d. auðveldara að greina einstaka liði í orðum, svo sem *önd* og *andi* í *anda-hyggjumaður*. Þar opnast nýr heimur sem ekki verður útlistaður hér en til skýringar má hugsa sér að merkingarflokkun liða myndi leiða líkur að því að orðið *alþingismaur* væri ritvilla fyrir *alþingismaður* og skiptingin *úran+auðgun* væri betri en *úra+nauðgun*. Þetta liggur í augum uppi fyrir fólk en tölvur þurfa líka að kunna á þessu skil.

Heimildir

- Adda-Decker, M., G. Adda og L. Lamel. 2000. Investigating text normalization and pronunciation variants for German broadcast transcription. *Proceedings of ICSLP*, bls. 66–269.
- Alfonseca, E., S. Bilac og S. Pharies. 2008. German Decompounding in a Difficult Corpus. Í: A. Gelbukh (ritstj.). *Computational Linguistics and Intelligent Text Processing*. Vol. 4919, bls. 128–129. Berlín/Heidelberg: Springer.
- Anna Björk Nikulásdóttir og Matthew Whelpton. 2010. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (Workshop 22 of 7th Language Resources and Evaluation Conference)*, bls. 33–39. Valletta.
- Beygingarlýsing íslensks nútímamáls (BÍN). Ritstj. Kristín Bjarnadóttir. bin.arnastofnun.is.

- Bergenholtz, Henning, Ilse Cantell, Ruth Vatvedt Fjeld, Dag Gundersen, Jón Hilmar Jónsson og Bo Svensén. 1997. *Nordisk leksikografisk ordbok*. Nordisk forening for leksikografi. Skrift nr. 4. Ósló: Universitetsforlaget AS.
- Braschler, M., Al Göhring og P. Schäuble. 2003. Eurospider at CLEF 2002. Í: C. Peters, M. Braschler og J. Gonzalo (ritstj.). *Advances in Cross Language Information Retrieval*. Vol. 2785, bls. 164–174. Berlín/Heidelberg: Springer.
- Brown, R. D. 2002. *Corpus-Driven Splitting of Compound Words*. *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translations (TMI-2002)*.
- Erla Hallsteinsdóttir, T. Eckart, C. Biemann, U. Quasthoff og M. Richter. 2007. *Íslenskur Orðasjóður – Building a Large Icelandic Corpus*. *Proceedings of NODALIDA-07*. Tartu.
- Hedlund T., H. Keskustalo, A. Pirkola, E. Airio og K. Järvelin. 2001. Utaclir @ CLEF 2001 – effects of compound splitting and n-gram techniques. *Second Workshop of the Cross-Language Evaluation Forum (CLEF)*, Revised Papers.
- Hrafn Loftsson, Ida Kramarczyk, Sigrún Helgadóttir og Eiríkur Rögnvaldsson. 2009. Improving the PoS tagging accuracy of Icelandic text. Í: Kristiina Jokinen og Eckhard Bick (ritstj.). *NODALIDA 2009 Conference Proceedings*, bls. 103–110.
- Íslenskur orðasjóður*. Universitát Leipzig. http://wortschatz.uni-leipzig.de/ws_isl/.
- Jón Friðrik Daðason, Kristín Bjarnadóttir og Kristján Rúnarsson. 2014. The Journal Fjölur for Everyone: The Post-Processing of Historical OCR Texts. *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives (LRT7HDA)*, bls. 56–62. LREC 2014. Reykjavík.
- Jón Hilmar Jónsson. 2012. Að fanga orðaforðann: orðanet í þágu orðabókar. *Orð og tunga* 14:39–65.
- Koehn, P. og K. Knight. 2003. Empirical methods for compound splitting. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Volume 1. Búdapest.
- Kristín Bjarnadóttir. 1995. Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base. Í: Ásta Svavarsdóttir, Guðrún Kvaran og Jón Hilmar Jónsson (ritstj.). *Nordiske studier i leksikografi* 3:255–263.
- Kristín Bjarnadóttir. 1996. Óbirt gagnasafn um samsett orð (sjá Kristínu Bjarnadóttur 2005).
- Kristín Bjarnadóttir. 2000a. „Rökleysa, lögbrot og hin gyðjumlíka Cameron Diaz.“ *Orðhagi*, afmæliskefja til Jóns Aðalsteins Jónssonar, 12. október 2000, bls. 83–87. Reykjavík.
- Kristín Bjarnadóttir. 2000b. *Þágufallssamsetningar í Ritmálssafni Orðabókar Háskólans*. <http://www.lexis.hi.is/kristinb/datsams.html>.
- Kristín Bjarnadóttir. 2002. *A Short Description of Icelandic Compounds*. <http://www.lexis.hi.is/kristinb/comp-short.pdf>.

- Kristín Bjarnadóttir. 2005. *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum*. Reykjavík: Orðabók Háskólans.
- Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 - AfLaT, LREC 2012*, bls. 13–18. Ístanbul.
- Magnús Snædal. 1992. Hve langt má orðið vera? *Íslenskt mál* 14:173–207.
- Mannanafnanefnd. 2005. *Úrskurður nr. 59/2005*. <http://www.urskurdir.is/DomsOgKirkjumala/Mannanafnanefnd/2005/06>.
- Mörkuð íslensk málheild (MÍM). Ritstj. Sigrún Helgadóttir. mim.arnastofnun.is.
- Ritmalssafn Orðabókar Háskólans. http://www.arnastofnun.is/page/gagna-sofn_ritmalssafn.
- Schiller, A. 2005. German Compound Analysis with wfsc. *Proceedings of the Fifth International Workshop of Finite State Methods in Natural Language Processing (FSMNLP)*, bls. 239–246. Helsinki.
- Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir og Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MÍM). *Proceedings of the SaLTMiL-AfLaT Workshop on “Language technology for normalisation of less-resourced languages”, 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Ísbúúl.
- timarit.is*

Lykilorð

máltækni, orðskipting, stofnhlutagreining, samsett orð

Keywords

language technology, decompounding, constituent structure, Icelandic compounds

Abstract

Compounding is extremely productive in Icelandic and multi-word compounds are common. The likelihood of finding previously unseen compounds in texts is thus very high, which makes out-of-vocabulary words a problem in the use of NLP tools. Kvistur, the decompounder described in this paper, splits Icelandic compounds and shows their binary constituent structure. The probability of a constituent in an unknown (or unanalysed) compound forming a combined constituent with either of its neighbours is estimated, with the use of data on the constituent structure of over 240 thousand compounds from the *Database of Modern Icelandic Inflection* (Kristín Bjarnadóttir 2012), and word frequencies from *Íslenskur orðasjóður*, a corpus of approx. 550 million words. Thus, the structure of an unknown compound is derived by comparison with compounds with partially the same constituents and similar structure in the training data. The granularity of the split returned by the decompounder is

important in tasks such as semantic analysis or machine translation, where a flat (non-structured) sequence of constituents is insufficient.

Jón Friðrik Daðason
Stofnun Árna Magnússonar í íslenskum fræðum / Háskóla Íslands
Neshaga 16
107 Reykjavík
jfd1@hi.is

Kristín Bjarnadóttir
Stofnun Árna Magnússonar í íslenskum fræðum / Háskóla Íslands
Neshaga 16
107 Reykjavík
kristinb@hi.is