

Jón Hilmar Jónsson

Að fanga orðaforðann: orðanet í þágu orðabókar

1 Inngangur

Í þessari grein er fjallað um orðabókarverkefnið *Íslenskt orðanet*. Verkefnið miðar að því að skipa íslenskum orðaforða, jafnt stökum orðum og merkingarbærum orðasamböndum, í samfellda heild, þar sem merkingarskylt orðafar tengist saman og myndar merkingarflokka af mismunandi tagi. Merkingarflokkunin byggist á því að rekja formleg vensl í efnismiklu safni orðasambanda og samsetninga og kalla með því fram merkingarvensl sem þau endurspegla.¹

Efni greinarinnar er skipað þannig að fyrst eru reifaðar þær breyttu aðstæður til orðabókarlýsingar sem rafræn gagnavinnsla og textabirting hefur haft í för með sér. Þá er gerð grein fyrir forsendum verkefnisins og efnviðinum og í framhaldi af því er viðfangsefninu lýst og þeim markmiðum sem að er stefnt. Endurmótun flettulistans í átt að einræðum (e. *monosemantic*) flettum, einræðingin sjálf (e. *disambiguation*) og myndun fleirytra flettna er síðan í brennidepli og þar á eftir er fjallað um gagnaefnið, ólíkar gagnategundir og merkingargreininguna sérstaklega. Þar beinist athyglin einkum að orðapörum og gildi þeirra við greiningu á merkingarlegum skyldleika og mat á skyldleikastigi. Loks er gerð stutt grein fyrir birtingarformi orðanetsins og vefsíðunni *ordanet.is*.

1 Verkefnið *Íslenskt orðanet* er samstarfsverkefni greinarhöfundar og Þórdísar Úlfarsdóttur. Ragnar Hafstað forritari hefur annast gerð gagnagrunna og forritun, og saman hafa Þórdís og Ragnar séð um gagnavinnslu. Verkefnið naut styrks úr Rannsóknarsjóði Rannís um þriggja ára bil.

2 Áskorun um endurmat orðabókarlýsingar

Rafræn birting orðabókargagna og sú gagnavinnsla sem þar er að baki hefur á margvíslegan hátt breytt forsendum og aðstæðum orðabókargerðar og losað um ýmsar hömlur við mótun og framsetningu orðabókarlýsingar. Frá sjónarmiði orðabókarnotenda virðist augljóst hagræði að því að geta farið beinustu leið að því orði sem athuga skal, tilgreina einfaldlega ritmynd orðs án þess að huga að stöðu þess í stafrófsröð með öðrum orðum eins og löngum hefur þurft. Frá sjónarmiði orðabókarhöfundar fela þær aðstæður líka í sér verulegt hagræði. Þar skiptir mestu að orðabókarlýsingin er ekki bundin endanlegum búningi frá upphafi, bæta má við nýjum orðum og efnisþáttum og endurskoða og lagfæra lýsinguna á öllum stigum verksins.

Í samanburði við orðabækur í prentuðum búningi eru meginumskiptin þau að viðhafa má einfaldari og frjálsari efnisskipan, þar sem einingar orðabókartextans eru í minna mæli háðar og undirskipaðar öðrum einingum. Um leið gefst kostur á margs konar tengingum og vísunum jafnt innan textans sem út fyrir hann, sem dýpka lýsinguna og bregða upp nýjum sjónarhornum (sjá nánari umfjöllun hjá Trap-Jensen 2008). Það gefur m.a. færi á að skoða samstæðar heildir innan orðaforðans betur og nánar en áður og draga fram heillega mynd af efnisþáttum sem löngum hafa birst á brotakenndan hátt í lýsingu einstakra orða.

Svo að þetta sjónarmið fái notið sín verður að endurskoða og umbreyta þeirri efnisskipan sem ráðið hefur ferðinni í prentuðum orðabókum. Við almennustu og víðtækustu lýsingu orðaforðans hafa formbundnir eiginleikar orðanna (flettiorðanna) jafnan verið lagðir til grundvallar þar sem merkingarlegur breytileiki og margræðni er með í för, með viðeigandi skiptingu í merkingarliði o.s.frv. Sú efnisskipan á ekki við þegar innbyrðis vensl og samstætt orðafar er til athugunar. Í stað þess þarf lýsingin í auknum mæli að beinast að merkingarbundnum eiginleikum orðanna, þar sem hver fletta er merkingarlega einræð. Með því móti verður merkingarþátturinn virkur til flokkunar orðaforðans og getur kallast á við málfræðileg einkenni orðanna (t.d. orðflokk) eða formbúning þeirra (sjá einnig Jón Hilmar Jónsson 2009a: 259–262).

Merkingarlegt sjálfstæði snýr ekki aðeins að stökum orðum. Þar eiga merkingarbær orðasambönd einnig hlut að máli. Í almennum orðabókum hefur lýsingu orðasambanda á ýmsan hátt verið þröngur

stakkur skorinn, þar sem þau eru jafnan undirskipaðar einingar sem erfitt getur verið að rata að og fá yfirsýn um, hvað þá að tengja saman út frá sameiginlegum einkennum.² Í orðabókarlýsingu sem miðast við merkingarbundna eiginleika flettnanna er ekkert því til fyrirstöðu að orðasambönd fái sjálfstæða stöðu, komi fram óháð þeim orðum sem þau eru mynduð úr og geti tengst stökum orðum sem jafngildar einingar.

Orðanetum er það sameiginlegt að þar er fengist við greiningu á merkingarvenslum orða í tilteknu tungumáli. Aðferðirnar eru hins vegar mismunandi og hlutverkið er sömuleiðis breytilegt (sjá Whelpton 2012 (þetta hefti)). Langþekktasta orðanetið er WordNet, sem upphaflega lýsir ensku með áherslu á samheitavensl stakra orða, en hefur síðan orðið fyrirmynd sams konar lýsingar á orðaforða fjölmargra tungumála. Því er ætlað að endurspegla innri skipan orðaforðans í huga málnotenda en um leið miðast greiningin mjög við máltæknileg not og vélræna meðferð.

Orðanetið sem hér verður lýst stendur nær almennri orðabókarlýsingu, greiningin byggist á orðavenslum eins og þau koma fram í textasamhengi og hún tekur jafnt til stakra orða og merkingarbærra orðasambanda. Hér eru innbyrðis vensl merkingarbærra orðasafns-eininga (orða jafnt sem orðasambanda) í brennidepli og leggja grunn að margháttadri flokkun orðaforðans. Öll úrvinnsla og greining efnisins miðast við rafræna birtingu. Lýsingunni eru ekki sett nein stærðarmörk að því er varðar fjölda flettna og gildi flettnanna ræðst fyrst og fremst af því hversu virkar þær eru í tengslum við aðrar flettur.

3 Forsendur og efniviður

Hvatinn að því að ráðast í gerð íslenskrar orðabókarlýsingar með ofangreind sjónarmið í huga er að miklu leyti sóttur til þeirrar reynslu sem fengist hafði við samningu þriggja efnislega samstæðra orðabóka um íslensk orðasambönd, *Orðastaðar*, *Orðaheims* og *Stóru orðabókarinnar um íslenska málnotkun* (sjá Jón Hilmar Jónsson 2001, 2002 og 2005).

2 Hugtakið *orðasamband* hefur hér afar rúma merkingu og tekur til sambanda tveggja eða fleiri orða sem oft koma fram sem meira eða minna samstæð setningarleg heild. Nánari tegundargreining orðasambanda markast m.a. af því hversu fastbundin þau eru og hvernig merking þeirra tengist merkingu orðanna sem þau eru mynduð af.

Þar var viðfangsefnið þess eðlis að bókarformið þrengdi mjög að lýsingunni. Í tveimur síðari bókunum er bæði um að ræða lýsingu á orðum og hugtökum, og textinn skiptist í aðgreinda bókarhluta, þar sem sérstök stafrófsröðuð orðaskrá vísar leiðina að orðasamböndum í aðalflettulistanum. Í orðaskránni eru vensl orða og orðasambanda, jafnt raðvensl (e. *syntagmatic relations*) sem staðvensl (e. *paradigmatic relations*), víða enn betur sýnileg en í megintextanum og við blasir hvernig formlegar samstæður geta endurspeglað merkingarlegan skyldleika. Þá hafði hið merkingarlega sjónarhorn fengið stóraukið vægi með *Orðaheimi* þar sem flettturnar eru hugtakaheiti og sameina merkingarskyld orðasambönd. Loks fylgdi *Stóru orðabókinni um íslenska málnotkun* útgáfa á geisladiski sem sýndi ljóslega fram á kosti og möguleika rafrænnar birtingar á viðamiklum og margbrotnum orðabókargögnum.

Á Orðabók Háskólans hafði um árabíl verið fengist við að semja skrá um orðasambönd sem fram koma í notkunardæmum í ritmáls-safni, þar sem framsetning orðasambandanna er með sama sniði og í orðabókunum fyrrnefndu. Því lá beint við að sameina öll fyrrnefnd orðabókargögn og mynda með því stofn í stærra orðabókarverki sem miðaðist við rafræna birtingu eingöngu. Verkinu var valið heitið *Íslenskt orðanet* út frá þeirri áherslu sem lögð er á greiningu á því margþætta venslaneti sem gögnin mynda, auk þess sem haft er í huga að verkið er ekki bundið sérstökum stærðar- eða efnismörkum fyrir fram.

Saman myndar þetta orðabókarefni tvíþættan stofn í gagnagrunni orðanetsins. Í flettulistanum koma saman þau lykilorð (einkum nafn-orð, lýsingarorð og sagnir) sem orðasambönd hafa verið færð undir en við þann lista bætast allar samsetningar (og stofnhlutar þeirra) sem fram koma undir flettiorðum í *Orðastað*. Samtals er hér um að ræða rösklega 160.000 einyrtar flettur. Hinn meginhluti stofnsins eru orðasamböndin sjálf, upphaflega um 180.000 sambönd.

Ljóst er að merkingargreining og merkingarflokkun á þessum gögnum getur haft stoð af öðrum tiltækum gögnum um merkingarlegan skyldleika orða. Í talmálsskrá Orðabókar Háskólans eru rakin samheiti (og að nokkru yfirheiti) sem heimildarmenn Orðabókarinnar tilgreina í umsögnum sínum um einstök orð. Þessi gögn voru felld inn í gagnagrunn orðanetsins en þau snerta alls um 17.000 orð.

Meðal jafnheita (þýðingarorða, e. *equivalents*) í tvímála orðabókum er samheitakennt orðafar að vonum víða áberandi. Því þótti ástæða til að afla gagna úr þeirri átt og fékkst leyfi útgefanda til að nýta efni úr

tveimur erlend-íslenskum orðabókum, *Dansk-íslenskri orðabók* (1992) og *Ensk-íslensku orðabókinni* (2006). Alls er hér um að ræða um 26.000 jafnheitasamstæður (með tveimur eða fleiri jafnheitum). Jafnt þetta efni sem efni talmálsskrárinnar veitir víða mikilvægan vitnisburð sem hingað til hefur aðeins verið nýttur að litlu leyti. Í því sambandi er rétt að hafa í huga að hér er að nokkru á ferðinni annars konar orðafar en í megingögnunum. Í talmálsefninu fer mikið fyrir sjaldgæfum og oft staðbundnum orðum sem lítið ber á í rituðum heimildum. Meðal jafnheitanna hefur andspænið við erlend orð aftur á móti sín áhrif og þar ber meira á nýyrðum og lausmótuðu orðafari en í megingögnunum.

4 Viðfangsefni og markmið

Eins og fram hefur komið er markmiðið með orðanetinu að rekja og greina merkingarvensl innan orðaforðans með áherslu á samheiti og hugtakavensl og láta setningarleg og orðmyndunarleg vensl vísa veginn í því efni. Viðfangsefnið er m.a. í því fólgið að móta viðeigandi efnisskipan og finna orðanetinu umgjörð og birtingarmynd við hæfi.

Gagnaefnið er vitaskuld margþættara en svo að það eigi allt beint erindi við almenna notendur, og vinnugrunnur orðanetsins, sem gefið var nafnið *Þesárus*, snýr fyrst og fremst að þeim sem vinna við greininguna og aðra þætti verksins. En til þess að koma verkinu á framfæri og opna aðgang að greiningunni var ákveðið að orðanetið ætti sérstaka vefsíðu, *ordanet.is*, þar sem það kæmi fram í orðabókarbúningi og auðvelt væri að leita og svipast um. Um vefsíðuna og innihald hennar verður fjallað nánar í 7. kafla en afmörkun þess sem þar birtist getur verið rúm eða þröng eftir atvikum. Ákvörðun í því efni snýst ekki síst um það hvort eða að hvaða marki orðasambandagögnin sem liggja greiningunni til grundvallar eigi að vera sýnileg á yfirborðinu eða hvort merkingarþátturinn og merkingarvenslin séu höfð í fyrirrúmi. Hér var síðari leiðin valin en tekið skal fram að birtingarmyndin er enn í mótun og getur breyst í samræmi við framvindu greiningarinnar.

Sé viðfangsefni orðanetsins skilgreint með hliðsjón af hefðbundnum orðabókartegundum má segja að það spanni þau hlutverk sem samheita- og hugtakaorðabókum er ætlað að gegna. Sá munur er þó á að hér eru þessi hlutverk samofin og markmiðið er að varpa skýrara og breytilegra ljósi á merkingarlega stöðu einstakra flettna í

hópi merkingarskyldra flettna með því að draga fram þau vensl sem orðabókargögnin leiða í ljós. Í stað þess að láta nægja að tengja saman tiltekin orð sem samheiti og byggja þar einkum á huglægu mati er hugsunin sú að birta margbrotnari mynd þar sem venslasamhengið við nálæg orð talar sínu máli. Meginmarkmiðið er að geta, á grundvelli orðanotkunar og greiningar gagnananna, gefið vísbendingar um hversu nán merkingarvenslin eru við einstakar flettur. Slík greining styður og kallast á við samheitamatið og leggur um leið grunninn að viðtækari merkingarflokkun.

Við samanburð á orðabókartegundum er orðabókartextinn ekki einn til vitnis um takmörkuð tengsl milli algengra og mikilvægra tegundarflokka heldur býr hver tegund um sig að mestu að sínu ef svo má segja, sínum sérstaka efniviði og þeim efnistöfum sem mótast við orðabókargerðina. Þess eru vissulega dæmi að orðabækur sameini þessar ólíku tegundir, t.d. þannig að almenn skýringaorðabók feli í sér samheitapátt sem eins konar viðauka við flettiorðin. Hins vegar hefur síður verið tekist á við að tengja þessa ólíku tegundarflokka saman á heildstæðan hátt út frá sameiginlegum efniviði.

Forsendan fyrir því að orðanetið geti spannað svo víðtækt hlutverk sem því er ætlað og birt heildstæða orðabókarlýsingu er sú að lýsingarþættirnir séu sprottnir af sameiginlegum efniviði og birting þeirra samofin. Gögnin sjálf og greining þeirra er einnig látin tala sínu máli í ríkara mæli en venja er til notendum til glöggvunar.

Staða og gildi orðasambanda sem undirstöðugagna í orðanetinu gefur færi á að greina og marka merkingarbær orðasambönd eftir formeinkennum og ná með því fram virku samspili merkingarlegrar og formbundinnar flokkunar. Slík greining er hliðstæð orðflokka-greiningu stakra orða og þannig geta orð og orðasambönd tengst saman út frá málfræðilegum einkennum. Sem flettur verða orðasamböndin að eiga sér staðlaðar myndir sem að sínu leyti gefa færi á innbyrðis flokkun eftir setningargerð. Þannig eru þrenns konar flokkunarþættir fyrir hendi meðal flettnanna sem tengja má á mismunandi vegu: setningarlegir, orðbundnir (lexíkalskir) og merkingarlegrir. Orðasambandið *drekka sig fullan* er að setningargerð í flokki með samböndum eins og *rífa sig lausan* og *spenna sig fastan*. Merkingarlega tengist það hins vegar samböndum eins og *skvetta í sig* og *detta í það*. Orðið *drekka* tengir það svo við önnur sambönd með þeirri sögn: *drekka af stút*, *drekka í botn* o.s.frv.

Við þær aðstæður sem hér er lýst er uppbygging og mótun flettu-listans samofin greiningu efnisins. Eins og áður kom fram er upp-

haflegur flettuforði að meginhluta bundinn setningarlegum og orð-myndunarlegum venslum og takmarkast ekki af notkunartíðni né öðrum hefðbundnum þáttum. Efniviðurinn og greining hans leiðir svo í ljós hversu virka stöðu fletturnar fá í orðabókarlýsingunni. Flettulistinn er áfram opinn á öllum stigum verksins og nýjar flettur bætast við eftir því sem greiningin gefur tilefni til.

Orðabókarlýsing sem verður til á þennan hátt og ætlað er svo við-tækt hlutverk er ekki kyrrstætt fyrirbæri heldur einkennist af stöð-ugum umbreytingum og endurbótum og í rauninni á hún sér engin skýr endimörk. Þannig kemur hún notendum fyrir sjónir og með það í huga verða þeir að nýta hana og meta. En eftir því sem greiningunni vindur fram verður lýsingin heillegri og þéttari og getur betur svarað fjölþættum kröfum notenda.

5 Endurmótun flettulistans, einræðing og fleiryrtar flettur

Upphaflegur flettulisti orðanetsins, eins og hann er sóttur til orðabók-argagnanna sem það byggist á, samanstendur af einyrtum flettiorð-um. Mörg þessara flettiorða eru bundin merkingarlegum breytileika og koma fram í ólíkum merkingarbrigðum og sá breytileiki er að nokkru leyti háður formlegu og setningarlegu samhengi. Þessi einkenni eru fyrirferðarmest meðal sagna þar sem margs konar sagnasambönd hafa sjálfstæða merkingu og heilir sagnliðir taka myndhverfingu sem rýfur merkingartengslin við einstök orð innan liðarins. Ef slík sambönd og önnur merkingarbrigði eiga að fá virka aðild að lýsingunni verður að gera hvert og eitt þeirra að sjálfstæðri flettu sem kallast á við aðrar flettur innan flettulistans (sjá einnig Jón Hilmar Jónsson 2009b).

Til þess að koma því í kring þarf að umbreyta flettulistanum á tvo vegu, ná fram merkingarlegri einræðingu og flettugera merkingarbær orðasambönd (þ.e. gera slík orðasambönd að fullgildum flettum) á samræmdan hátt. Ákvörðun um einræðingu ræðst eðli málsins samkvæmt af vitnisburði orðabókarefnisins og fylgir þar með framvind-unni í greiningu þess. Myndun fleiryrttra flettna getur hins vegar að verulegu leyti farið fram á samfelldari hátt með hliðsjón af stöðluðum myndum orðasambandanna. Sú aðgerð hefur langmest áhrif á sagnir og sagnasambönd svo að fjöldi sagna flettna í orðanetinu hefur allt að

Því tífaldast. Fleiryrtir nafnliðir koma einnig til sögunnar, og atviksliðir koma fram á sjónarsviðið sem virkar flettur. Meðferð lýsingarorða í þessu samhengi er hins vegar óskýrari, þar sem merking þeirra er gjarna afar kvik, en einræðing og fleiryrtar flettur hafa þar einnig hlutverki að gegna eins og vikið verður að í kafla 5.4.

Í flettulistanum kemur einræðingin fram í töluröðuðum lista þar sem lýsandi skýring eða skýringarorð auðkennir hverja flettu fyrir sig:

- (1) 1 fýla no kvk (vond lykt)
- 2 fýla no kvk (ólund)
- 3 fýla no kvk (þoka, fúlviðri)

Slík aðgreining einyrtra flettna er mest áberandi meðal nafnorða enda eru merkingarskilin þar hvað greinilegust. Meðal fleiryrttra flettna er hún hins vegar fátíð enda nær einræðingin þar oftast að endurspeglast í sjálfri flettumyndinni.

5.1 Fleiryrtar sagnaflettur

Sagnaflettur í orðanetinu eru settar fram sem heilir sagnliðir og rökliðagerðin á þann hátt gerð sýnileg. Framsetningin er að fyrirmynd orðasambanda í *Orðastað* og *Orðaheimi*, með afmörkun breytilegra liða innan oddklofa:

- (2) gefa <honum, henni> <mat, hressingu>
renna <bílnum> <inn í stæðið>

Fletturarnar í (2) hefjast á sögn í nafnhætti án þess að á undan fari tilgreint frumlag. Flettur með því sniði eiga við þegar um er að ræða nefnifallsfrumlag í eintölu með vísun til persónu. Annars er frumlagsliðurinn tilgreindur á undan sögninni eins og í (3):

- (3) <dómur> fellur
<honum, henni> leiðist <námið>
<þeir, þær, þau> þérast

Krafan um einræðingu og fleiryrtar flettur gerir það að verkum að af einstökum sögnum geta sprottið fjölmargar sagnaflettur, breytilegar að formi og merkingu. Svo að vísað sé til algengra og fyrirferðarmikilla sagna kemur sögnin *halda* fram í rösklega 500 flettum, í ólíku um-

hverfi orða og setningarliða. Flestar tengjast persónubundnu frumlagi í eintölu eins og sýnd eru dæmi um í (4):

- (4) halda á <bókinni>
 halda áfram <ferðinni>
 halda hlífiskildi yfir <honum, henni>
 halda kjafti
 halda sér <vel>
 halda uppteknum hætti
 halda <sættina>
 halda <starfseminni> í gangi
 halda <þessu> til streitu

Aðrar víkja frá því með tilgreindum frumlagslið og sagnmynd í 3. persónu eins og í (5):

- (5) <flíkin> heldur ekki þræði
 <skipið> heldur sjó
 <þessu> heldur fram <langa hríð>
 <honum, henni> halda engin bönd
 <þeir, þær, þau> halda saman

Þótt sagnmyndin sé greinilegt samkenni þessara sagnaflettna fer merking sagnarinnar *halda* hér mjög á dreif. Hins vegar koma fram sameiginlegir þættir í formi flettnanna sem endurspegla tiltekin einkenni, bæði gagnvart sögninni *halda* sérstaklega og gagnvart sagnaflettunum í heild, þættir eins og fallstjórn og einkenni frumlags.

Með því að marka flettustrengi sagnaflettnanna málfræðilega fæst mikilvæg yfirsýn um setningarleg einkenni þeirra. Jafnframt fá flettunar málfræðilegt auðkenni sambærilegt við orðflokk einyrtra flettna sem í raun er þó mun nákvæmara (greinargerð um mörkunina er að finna hjá Þórdísi Úlfarsdóttur 2006). Mörkunarþættirnir mótast af því gildi sem þeir hafa gagnvart flettunum, þar sem orðflokkur, fall og ákveðni (greinir) og að nokkru leyti tala skipta meginmáli en litið er fram hjá undirskipuðum þáttum eins og kyni nafnorða. Markið (mörkunarstrengurinn) skiptist í liði sem fylgja orðaröð flettunnar, breytilegir liðir afmarkast með oddklofum og þar er aðeins tilgreindur einn markþáttur (t.d. þannig að atviksliður sem heild fær markþáttinn ao).³

3 Orðflokkar eru skammstafaðir á hefðbundinn hátt, nafnorð með no, sagnir með so, lýsingarorð með lo, atviksorð (og atviksliðir innan oddklofa) með ao, fornöfn með fn, forsetningar með fs, samtengingar með st. Óákveðin fornöfn fá viðbótarþáttinn

Svo að aftur sé litið til flettna með sögninni *halda* varpa mörkun-arstrengir þeirra ljósi á setningarleg einkenni sagnarinnar og ýmsar samstæður verða sýnilegar eins og sjá má í (6)1–4:

- (6) 1 so no-dg
 halda þræðinum
 halda munninum
 halda ærunni
- 2 so fs no-a fs <fn-p-d>
 halda í ár með <honum, henni>
 halda í hönd með <honum, henni>
 halda í taum með <honum, henni>
- 3 so <no-a>
 halda <fund, ráðstefnu>
 halda <jól, páska; afmæli>
 halda <vinnumann, ráðskonu>
- 4 <no-ng> so
 <áin; ísinn> heldur
 <reipið> heldur
 <samningurinn> heldur

Þannig getur fengist athyglisverður samanburður á einstökum sögnum, bæði að því er varðar setningarleg einkenni og hversu virkar þær eru í ýmsum sagnasamböndum.

Með tilliti til flokkunar er þó mikilvægara að athuga hvernig tiltek-in mörk koma fram í heild meðal flettnanna, óháð einstökum sögnum. Á mynd 1 koma fram stafrófsraðaðir bútar úr flettulista þriggja marka. Í (1a) er nafnliðurinn breytilegur og þar beinist athyglin að fallstjórn og merkingareinkennum nafnorðins en í (1b) og (1c) er um fast nafnorð að ræða og sambandið myndar skýra merkingarheild, í flestum tilvikum sem fastmótað orðtak.

Myndun fleiryrrtra sagnaflettna dregur hér orðtökin inn í raðir sagnanna, þar sem þau birtast í samhengi við formlega og merkingarlega skyldar flettur en standa ekki einangruð í undirskipuðum liðum nafnorðaflettna eins og venja er til í prentuðum orðabókum.

Málfræðileg mörkun fleiryrrtra flettna, sagnaflettna jafnt sem annarra, hefur verulegt hagnýtt gildi til leiðsagnar um samheitavensl og

–óákv fornafn með vísun til persónu er auðkennt með -p og afturbeygt fornafn fær viðbótarþáttinn -r. Fall nafnorða, lýsingarorða og fornafna er auðkennt með -n fyrir nefnifall, -a fyrir þolfall, -d fyrir þágufall og -g fyrir eignarfall. Nafnorð með greini fær auðkennið g aftan við fallþáttinn.

aðra merkingarflokkun. Að því er varðar samheitavensl er samstæð rökliðagerð mikilvægt skilyrði og sé það uppfyllt (þannig að samheit-in séu umskiptanleg í setningarlegu umhverfi) eru fleiryrtar flettur jafngildar stökum orðum í því samhengi.

244 brjóta <glerið, bollann, spýtuna> so [so <no-ag>] cc
245 brjóta <innsiglið, fjöturinn> so [so <no-ag>] cc
246 brjóta <ljósgelsinn> so [so <no-ag>]
247 brjóta <reglurnar, boð yfirvaldanna> so [so <no-ag>]
248 brjóta <örkin> so [so <no-ag>]
249 brunatryggja <húsið, innbúið> so [so <no-ag>]
250 brúa <lækinn, ána> so [so <no-ag>] cc
251 brúka <hestinn> so [so <no-ag>]
252 brúleggja <ána; stéttina> so [so <no-ag>] cc
253 brúna <steikina; kartöflurnar> so [so <no-ag>] cc
254 brúnelda <steikina> so [so <no-ag>]
255 brydda <skóna> so [so <no-ag>] cc
256 brytja <kjötið, mörinn> so [so <no-ag>] cc
257 brýna <hnífinn, eggina> so [so <no-ag>]
258 bræða <isinn, klakann> so [so <no-ag>] cc

(1a)

61 brjóta brá so [so no-a]
62 brjóta land so [so no-a]
63 brjóta lög so [so no-a]
64 brjóta trúnað so [so no-a]
65 brúka kjaft so [so no-a]
66 brúka munn so [so no-a]
67 byggja loftkastala so [so no-a]
68 byggja skýjaborgir so [so no-a]
69 byrja barn so [so no-a]
70 bæta bót so [so no-a]
71 dekkja borð so [so no-a]

(1b)

40 brjóta heilann so [so no-ag]
41 brjóta ísinn so [so no-ag]
42 brýna gogginn so [so no-ag]
43 brýna klærnar so [so no-ag]
44 brýna kutana so [so no-ag]
45 brýna raustina so [so no-ag]
46 brýna röddina so [so no-ag]
47 burstu tennurnar so [so no-ag]
48 byrgja andlitið so [so no-ag]
49 deyfa eggina so [so no-ag]
50 draga andann so [so no-ag]
51 draga fótinn so [so no-ag]
52 draga fæturna so [so no-ag]
53 draga halann so [so no-ag]
54 draga hlassið so [so no-ag]
55 draga kvíðinn so [so no-ag]
56 draga lappirnar so [so no-ag]

(1c)

Mynd 1: Svipmyndir af mörkunarstrengjum fleiryrttra sagnaflettna. Í (1a) er breytilegur nafnliður afmarkaður með oddklofum. Í (1b) og (1c) mynda samböndin skýra merkingarheild með föstu nafnorði.

5.2 Nafnliðaflettur

Þótt mest beri á sagnliðum meðal merkingarbærra orðasambanda geta heilir nafnliðir einnig verið merkingarbærar einingar sambærilegar við stök nafnorð. Fyrirferð nafnliða sem fleiryrtar flettna hefur aukist verulega í orðanetinu eftir því sem greiningunni hefur miðað áfram og nýtt gagnaefni hefur bæst við. Slíkar flettur fá, eins og fleiryrtar sagnaflettur, málfræðilega mörkun sem nýta má til frekari flokkunar þeirra. Þær hafa sérstaklega komið til sögunnar við myndun orðapara (sjá kafla 6.3), þar sem merkingarbærir fleiryrtir nafnliðir kallast á við stök nafnorð í hliðskipuðum samböndum: *styrjaldir* og *vopnuð átök*, *óbyggðir* og *villt náttúra*, *opinberar heimsóknir* og *veisluhöld*, *félagslíf*

og mannleg samskipti, afbrýðisemi og illt umtal, gróið land og bithagar. Gildi þeirra felst hér að nokkru leyti í því að skerpa merkingarleg einkenni þeirra flettna sem þær tengjast en um leið búa þær yfir merkingareinkennum sem beina þeim í samheitasambönd og stærri merkingarflokka. Aðrir nafnliðir eru sjálfstæðir og merkingarbærir án tillits til samhengis: *ást í meinum, eldri borgarar, samsett orð, dýr merkurinnar, tímans tönn*.

Staða og hlutverk nafnorða í setningarlegu samhengi er breytilegt með tilliti til þess hvort þau eru rökliðir (frumlög og andlög) eða umsagnir (sagnfyllingar). Nafnorð í sagnfyllingarstöðu með sögninni *vera* hafa áþekkt hlutverk og lýsingarorð, og sum nafnorð koma fyrst og fremst fram í þeirri stöðu: *vera ágætismaður, vera drullusokkur, vera plága, vera söngmaður*. Til að draga fram þessa sérstöðu og skerpa merkingartengslin við sambærileg lýsingarorð eru flettumyndir nafnorða af þessu tagi hafðar tvíyrta og eftir atvikum fleiryrtar: *vera hamhleypa, vera hamhleypa til verka, vera hamhleypa að dugnaði, vera funi í skapi, vera hafsjór af fróðleik*. Tengslin við lýsingarorð í fleiryrtum samböndum koma auk þess gjarna fram í því að sambandið (og þar með flettan) rúmar bæði nafnorð og lýsingarorð: *vera besta skinn, vera harður húsbóndi; vera forkur duglegur, vera köttur liðugur*. Þessi tilhögun styrkir stöðu þessara sambanda sem flettna og greiðir fyrir því að flettturnar rati í viðeigandi merkingarflokka. Nafnorðaflettur af þessu tagi, með nafnháttarmynd sagnarinnar *vera* sem upphafslið, eiga við persónubundið (ótilgreint) frumlag.

5.3 Atviksliðaflettur

Merkingarbærir atviksliðir hafa yfirleitt verið illa sýnilegir sem afmarkaðar einingar í orðabókartexta og að því marki sem þeir koma fram eru þeir undirskipaðir einyrtum flettiorðum, oftast nafnorðum eða sögnum.

Í orðanetinu eiga slíkir liðir fullan rétt á sér sem fleiryrtar flettur í virkum form- og merkingartengslum innbyrðis og við stök atviksorð, og við flettur af öðrum orðflokkum innan stærri merkingarheilda. Hér sem annars staðar dregur málfræðileg mörkun fram orðbundin og setningarleg munstur, sem í mörgum tilvikum endurspeglar merkingarleg vensl, eins og sýnd eru dæmi um á mynd 2.

1	frá blautu barnsbeini	ao	[fs lo-d no-d]
2	frá fornheiðnum tíma	ao	[fs lo-d no-d]
3	frá fornu fari	ao	[fs lo-d no-d]
4	frá forsögulegum tíma	ao	[fs lo-d no-d]
5	frá fyrsta fari	ao	[fs lo-d no-d]
6	frá fyrstu gerð	ao	[fs lo-d no-d]
7	frá fyrstu hendi	ao	[fs lo-d no-d]
8	frá fyrstu tíð	ao	[fs lo-d no-d]
9	frá gamalli tíð	ao	[fs lo-d no-d]

Mynd 2: Mörkunarstrengir fleiryrttra atviksliðafllettna draga fram orðbundin og setningarleg munstur, sem oft endurspeгла merkingarleg vensl.

Atviksliðafllettur í orðanetinu eru nú orðnar nær 12.000 talsins og gera má ráð fyrir að þeim fjölgi við nánari úrvinnslu og með nýjum efniviði. Enn sem komið er takmarkast greining þeirra í orðanetinu að mestu við formið (mörkunarstrengi) en merkingarflokkunin er skemmra á veg komin.

5.4 Fleiryrtar lýsingarorðafluttur

Lýsingarorð eru eðli málsins samkvæmt erfið viðureignar þegar greina á í sundur merkingarbrigði og ákveða viðeigandi skipan merkingarliða. Merking þeirra er kvik og í málnotkunarsamhengi litast hún og mótast af nafnorðunum sem þau standa með og einkenna hverju sinni. Því er vandkvæðum bundið að viðhafa einræðingu á sambærilegan hátt og meðal nafnorða. Meðferð lýsingarorða í orðanetinu er enn ekki fullmótuð en að því leyti sem skýr einræðing virðist raunhæf er hún látin koma fram í fleiryrtum flettumyndum.

Í kafla 5.2 var fjallað um meðferð nafnorða í sagnfyllingarstöðu. Farið er með lýsingarorð í sömu stöðu á sambærilegan hátt, flettumyndirnar eru fleiryrtar með *vera* sem upphafslið og miðast við persónubundið (ótilgreint) frumlag í nefnifalli: *vera áflogagjarn*, *vera ánægður með sig*, *vera þurr á manninn*. Með þessari framsetningu og mörkunarstrengjum þar sem sögn er í fremsta sæti dragast lýsingarorðasambönd af þessu tagi og sambærilegir nafnliðir inn í umhverfi sagnafllettna og kallast á við þær. Það endurspeglar stöðu þeirra í málnotkuninni og leggur til mikilvægan aðgreiningar- og einræðingarþátt í flettumyndun lýsingarorða. Flettumyndunin sem slík tryggir raunar einræðingu á langflestum tilvikum en tvíræðar myndir koma þó fyrir eins og sjá má í (7) og (8):

- (7) 1 vera sofandi (sofa) [so lo-n]
 2 vera sofandi (vera sljór/viðutan) [so lo-n]
- (8) 1 vera klár (vera skýr/skarpur) [so lo-n]
 2 vera klár (vera undirbúinn) [so lo-n]

Merking lýsingarorða sem eiga við hluti og fyrirbæri er augljóslega kvikari að þessu leyti og það takmarkar möguleika á sundurgreiningu og myndun fleiryrttra flettna þeirra á meðal. En að því marki sem það á við kemur einræðingin fram í tvískiptum flettumyndum, þar sem aðgreinandi og einkennandi nafnorð, afmarkað með hornklofum í nefnimynd sinni, fer á eftir hefðbundinni flettumynd eins og sýnt er í (9):

- (9) 1 bitur [bragð] [lo [no]]
 2 bitur [frost] [lo [no]]
 3 bitur [háð] [lo [no]]
 4 bitur [hnífur, sverð] [lo [no]]
 5 bitur [reynsla] [lo [no]]

Með því er markað fyrir tengingu við önnur lýsingarorð líkrar merkningar með sömu nafnorðatengslum eins og dæmin í (10) sýna:

- (10) 1 beiskur [háð] [lo [no]]
 2 beittur [háð] [lo [no]]
 3 bitur [háð] [lo [no]]
 4 egghvass [háð] [lo no]
 5 napur [háð] [lo no]

En einkennisþættirnir eru misjafnlega skýrir og nafnorðin sem endurspegla þá eru vitaskuld ekki gefin fyrirfram. Hins vegar getur þessi flettumyndun verið afar hagnýt með tilliti til þess að tengja saman lýsingarorðaflettur með sameiginlegum einkennislið burtseð frá lýsingarorðinu sem á undan fer. Frá því sjónarmiði getur þótt ástæða til að tefla fram slíkum myndum til þess beinlínis að geta rakið lýsingarorðatengsl einstakra nafnorða: *brúnahvass* [fjall], *fannþakinn* [fjall, hlið], *keilulaga* [fjall], *ókleifur* [fjall, fjallstindur, klettur]. Hvað sem því líður halda hinar upphaflegu einyrtu lýsingarorðaflettur jafnframt gildi sínu óbreyttar innan flettulistans með því gagnaefni sem þeim tengist.

6 Gagnaefni, gagnategundir og merkingargreining

Í gagnagrunni orðanetsins koma saman margþætt gögn sem, eins og áður er rakið, eru að miklu leyti af sama stofni og með samstæðu yfirbragði. Uppbygging og endurmótun flettulistans, sem lýst hefur verið hér að framan, er mikilvægur liður í meðferð og greiningu þessara gagna. Í framhaldi af þeirri lýsingu verður hér gerð almennari grein fyrir gögnunum og hvers eðlis þau eru, hvernig þau skiptast í aðgreindar tegundir, að hverju er stefnt með greiningu gagnanna og hvernig greiningin fer fram. Sem veggvísir í þessu efni er valin nafnorðsflettan *hlátur* með því gagnaefni sem henni tilheyrir en þar er alls um að ræða rösklega 400 gagnafærslur (orðasambönd, samsetningar og notkunardæmi).

6.1 Flettan *hlátur*

Eins og algengt er meðal nafnorða skiptast orðasambönd sem vitna um dæmigert notkunarsamhengi orðsins *hlátur* í þrjá hluta. Í fyrsta lagi er um að ræða orðastæður með lýsingarorði: *dynjandi hlátur*, *óstöðvandi hlátur*, *kaldranalegur hlátur*, *tröllslegur hlátur*. Önnur eru sagnasambönd: *bresta í hlátur*, *það setur að <honum, henni> hlátur*, *veltast um af hlátri*. Í þriðja lagi kemur orðið fram í hliðskipuðum orðapörum: *glæðværd og hlátur*, *kæti og hlátur*, *hlátur og fliss*, *hlátur og mas*. Eins og áður er rakið ummyndast sagnasamböndin í fleiryrtar sagnaflettur, og sami kostur er einnig fyrir hendi gagnvart lýsingarorðum: *dynjandi [hlátur]*, *niðurbældur [hlátur]*, *hávær [hlátur]*.

Hin megingagnategundin eru samsetningar orðsins og orðhlutar þeirra: *gleði-hlátur*, *hrossa-hlátur*, *skelli-hlátur*, *hæðnis-hlátur*; *hlátur-gusa*, *hláturs-kast*, *hláturs-kjölt*, *hlátur-mildur*.

Í fyrstu lotu beinist greining gagnanna að merkingarlegum samstæðum meðal lykilorða flettunnar, þ.e. meðal orða og orðasambanda sem hún tengist í gagnagrunninum. Þá aðgerð má takmarka við að mynda samstæður innan flettugreinarinnar og skipa þar saman tilteknum orðasamböndum með samheitum: *óstjórnlegur hlátur*, *óstöðvandi hlátur*; *spjall og hlátur*, *samtal og hlátur*. Slíkar samstæður má svo rekja saman við samstæður í öðrum flettum út frá sameiginlegum orðasambandaliðum og draga þannig saman hóp tengdra orð til frekari

flokkunar. Önnur og beinni aðgerð er að skipa saman skyldum flettum undir tiltekna gagnategund í gagnagrunni orðanetsins. Meðal slíkra gagnategunda eru SAMHEITI, SKYLDHEITI (merkingarskyldar flettur sem ekki ná því að teljast samheiti (sbr. e. *near-synonyms*)), ANDHEITI og MERKING, en síðastnefnda tegundin sameinar heila merkingarflokka undir viðeigandi fyrirsögn. Oft er erfitt að dæma um samheiti því skyldleikastigið er mismunandi og ýmsir aðgreinandi þættir hafa þar áhrif, orðin geta tilheyrt ólíkum tíma og samfélagsgerð, haft ólíkt stílgildi o.s.frv. Huglægt mat og innsæi þess sem greinir ræður því hvernig greint er og þá fer ekki hjá því að nokkurs ósamræmis gæti. Með það í huga eru tvær skyldar gagnategundir í boði þegar meta skal samheitakennt orðafar.

Greining á merkingarvenslum á þann hátt sem hér var lýst tekur jafnt til orðasambanda, samsetninga og orðhluta. Í efni flettunnar *hlátur* má m.a. greina eftirtalin samheiti: *hláturgjarn*, *hláturmildur*; *galsa-hlátur*, *gáskahlátur*; *holur* [*hlátur*], *tómlegur* [*hlátur*]; *vera að springa úr hlátri*, *vera að rifna úr hlátri*, *verða vitlaus í hlátri*; *hviða*, *roka*. Aðrar samstæður bera fremur svip skyldheita: *hálfkæfður* [*hlátur*], *krampakenndur* [*hlátur*]; *dillandi* [*hlátur*], *galopinn* [*hlátur*]; *hlátur*, *fliss*. Þá má einnig greina andheiti: *glæðvær* [*hlátur*] – *gleðisnauður* [*hlátur*]; *hávær* [*hlátur*] – *lágvær* [*hlátur*]. Loks má tengja saman flettur undir fyrirsögnum innan tegundarinnar MERKING: *hlátur*, *hlæja*, *springa af hlátri* (undir **hlátur**); *aðhlátur*, *hafa <hann, hana> að hlátri*, *verða að hlátri* (undir **aðhlátur**); *kuldahlátur*, *hlæja kuldahlátri*, *hlæja kalt* (undir **kuldahlátur**). Greining á efni annarra flettna á svo sinn hlut að því að byggja upp þær samstæður sem hér hafa myndast.

6.2 Merkingarflokkun og samheiti

Merkingarflokkun flettnanna tekur enn sem komið er að mestu til sagnaflettna en þar er greiningin í *Orðaheimi* höfð til fyrirmyndar svo langt sem hún nær. Þorri sagnaflettnanna hefur verið flokkaður og flokkun nafnorðaflettna er komin nokkuð á veg. Í stað þess að byggja upp fast og stigveldisbundið kerfi merkingarflokka eins og sterkust hefð er fyrir í hugtakaorðabókum (sbr. m.a. *Roget's Thesaurus of English Words and Phrases* (1988) og orðabók Dornseiffs, *Der deutsche Wortschatz nach Sachgruppen* (2004)) er farin sú leið að láta tengsl flokkanna endurspeglast í skörun þeirra í milli og viðeigandi millivísunum. Ein-

stakar flettur geta þá birst í ólíku flokkunarsamhengi og eftir atvikum tengst fleiri en einum merkingarflokki. Notendur þurfa ekki endilega að þekkja heiti merkingarflokkanna eða rekast á þau heldur er gert ráð fyrir að þeir komist á viðeigandi slóð út frá flettu(m) sem þeir gera ráð fyrir að varði leiðina þangað.

Í 7. kafla er gerð nánari grein fyrir merkingarflokkuninni eins og hún birtist notendum á vefsíðu orðanetsins. Framgangur hennar í greiningunni ræðst af því hvaða áhersla er lögð á þá þætti sem liggja henni til grundvallar. Þar hafa vissar aðgerðir orðið fyrirferðarmeiri en séð var fyrir í upphafi og gagnagrunnurinn hefur verið aukinn með verulegu viðbótarefni, í því augnamiði að víkka sjónarsviðið við greiningu merkingarvenslanna og styrkja með því flokkunina. Greining samheita (og tilbrigða þeirra) hefur frá upphafi verið fyrirferðarmikill þáttur í uppbyggingu orðanetsins, og sú greining hefur um leið skilað efni til afmörkunar stærri merkingarflokka. En merkingarflokkar og samheiti, þar sem byggt er á mati og innsæi greinandans, ná ekki ein sér að lýsa upp það svið sem merkingarvenslin birtast á. Þar þurfa gögnin sjálf líka að geta talað sínu máli og sveigt til þær skörpu línur sem alltaf myndast þegar skipað er í aðgreinda flokka.

6.3 Orðapör sem gagnategund

Meðal orðasambanda sem fram komu undir flettunni *hlátur* eru orðapör, þ.e. hliðskipuð sambönd með jafnvægum liðum: *pískur og hlátur*, *hlátur og skvaldur*. Í grunngögnum orðanetsins er allmikið af slíkum samböndum, nógu mikið til að gefa skýra vísbendingu um gildi þeirra við merkingargreiningu og merkingarflokkun og til að sannfærast um að orðapör skuli skilgreind sem sjálfstæð gagnategund (PAR) í gagnagrunni orðanetsins. Vensl slíkra hliðskipaðra setningarliða eru sérlega bein og nán að því leyti að liðirnir falla inn í sama umhverfi og þau eru því ekki háð samhenginu við aðra setningarliði (t.d. sameiginlegri sögn eða lýsingarorði þegar um er að ræða nafnorð). Merkingarlegt samband liðanna er hins vegar misjafnlega náðið og af því sprettur fjölbreytnin þegar litið er til einstakra flettna.

Með tilliti til þessa þótti þörf á að afla viðbótarefniviðar. Ákveðið var að leita fanga í hinu umfangsmikla textasafni Landsbókasafns Íslands – Háskólabókasafns *Tímarit.is*, en *Íslenskt textasafn* Stofnunar Árna Magnússonar í íslenskum fræðum hefur einnig verið nýtt í sama

skyni. Með markvissri leit að orðapörum í textum þessara safna hefur fjöldi þeirra stórauðist og gildi gagnategundarinnar aukist að sama skapi í orðanetinu. Til þessa hefur megináherslan verið lögð á nafnorð en orðapörin hafa einnig augljóst gildi gagnvart lýsingarorðum og raunar einnig gagnvart sögnum.

Með því að rekja orðapör með einstökum orðum er greinilega farið inn á slóð samheita og í vænu safni slíkra sambanda er jafnan að finna samstæð orð af því tagi, sumpart orð sem eiga við orðið sem rakið er út frá en ekki síður meðal orða sem það dregur til sín. Orðapör sem rakin eru út frá flettunni *hlátur* skila t.d. samheitunum *gamanssemi* og *spaugsemi*, og *hávaði* og *skarkali* ná saman á sama hátt. Málnotkunarlegt gildi orðapara virðist ekki síst í því fylgið að skerpa og auðga myndina af því sem er til umræðu eða umsagnar og gefa því viðeigandi merkingarblæ. En inntak og eðli þess sem við er átt getur líka verið samsett og þá tjáir orðaparið náðið samband þeirra orða sem í hlut eiga. Þetta skýrir að einhverju leyti þá gífurlegu fjölbreytni og þann mikla fjölda orðapara sem einstök orð reynast eiga hlut að í textadæmum fyrrgreindra safna. Þar eru ekki aðeins á ferðinni samheiti heldur skilar dæmasafnið víða drjúgum skerf og jafnvel nýjum orðum inn í einstaka merkingarflokka. Af þessum sökum hefur flettuforði orðanetsins aukist að mun, m.a. með fleiryrtum nafnliðaflettum eins og áður er getið.

Gildi orðapara sem vitnisburðar um notkunar- og merkingarvensl kemur fram í því að á vefsíðu orðanetsins eru þau tilgreind við einstakar flettur notendum til glöggvunar. Sem gagnategund í greiningunni reynast þau svo hafa undirstöðugildi þegar kemur að því að meta það sem orðanetinu var m.a. ætlað að vitna um, hvernig skyldleikastigi flettnanna er háttað og hvort greina megir á milli samheita með tilliti til þess.

6.4 Mat á merkingarskyldleika

Gerð orðanetsins hvílir eins og áður segir á þeirri meginhugmynd að merkingarvensl megir rekja út frá venslum orða í orðasamböndum og samsetningum. Þá er nærtækt að líta svo á að styrkur merkingarvenslanna, merkingarleg nálægð eða skyldleikastig flettnanna hveirrar gagnvart annarri, komi fram í því hversu oft og víða um er að ræða formvensl við sömu orð. En hafa verður í huga að fletturarnar

eru misjafnlega virkar, og svo miðað sé við orðasambandavensl (samsetningarvensl hafa óskýrara gildi að þessu leyti) getur stundum verið erfitt að greina dæmigerð og einkennandi orðasambönd. En virkar flettur, þar sem um er að ræða drjúgan fjölda tengdra flettna (lykilorða) í gagnagrunninum, má bera saman og athuga hvaða mælikvarðar koma til greina við að meta skyldleikastigið. Slíkt mat á áþreifanlegum og tölulegum grunni hefur bæði fræðilega og hagnýta þýðingu og styður þær beinu flokkunaraðgerðir sem áður er lýst (um aðferðir við mat á merkingarskyldleika má m.a. lesa hjá Önnu Björk Nikulásdóttur og Matthew Whelpton 2010).

Merkingarflokkurinn **ósannindi** sameinar orð og (merkingarbær) orðasambönd, þar sem saman koma nafnorð, lýsingarorð og sagnir: *lygi, uppspuni, skrök, staðlausir stafir; lygari, lygalaupur, ósannindamaður; lyginn, skröksamur, skreytinn; ljúga, búga* <þetta> til, fara frjállega með staðreyndir, segja ósatt. Með yfirsýn um þetta orðafar í gagnagrunni orðanetsins má mynda klasa með samheitum (*lygi, ósannindi, skrök, uppspuni*), tengja saman skyldheiti (*lygi, hálf sannleiki; ljúga, leyna sannleikanum*) og greina andheitapör (*sannleikur – lygi; sannindi – ósannindi*). Þá er byggt á mati greinandans án þess að það sé stutt sýnilegum gögnum. En þar sem um er að ræða flettur með fjölbreyttum lykilorðum getur beinn samanburður á lykilorðatengslum varpað ljósi á merkingarvenslin og verið til vísbendingar um hversu náin þau eru. Á mynd 3 eru nokkrir bútar úr slíku tengslakorti flettnanna *lygi* og *ósannindi* þar sem fram kemur að flettarnar eiga sér 87 sameiginleg lykilorð (í miðdálkinum) en önnur lykilorð eru bundin annarri þeirra í gagnaefninu.

Hér ber mest á gagnategundunum ORÐASAMBAND (O) og ORÐAPAR (P) en SAMSETNINGARLIÐUR (H), ORÐTAK (ot), ANDHEITI (A) og JAFNHEITARUNA úr erlend-íslenskum orðabókum (R) koma einnig við sögu. Með því að bera flettuna *lygi* að fleiri flettum á þennan hátt fæst mynd af því hvaða flettur standa henni næst með tilliti til lykilorðatengsla.

Gallinn við þennan vitnisburð er sá að hér er blandað saman ólíkum gagnategundum sem geta vegið misjafnlega þungt. Með því að binda samanburðinn við tiltekna tegund sem sýnir nægilega virkni má ætla að myndin geti orðið skýrari og ná megi lengra við að greina og meta merkingarlegt skyldleikastig. Þá er vænlegast að líta til orðapara því þar er virknin og fjölbreytnin gjarna mest og þar tengjast saman málfræðilega samstæðar flettur (af sama orðflokki).

lygi 127	bæði 87	ósannindi 156
H hðfuð- forl	O bláber lo O	ó- forl H
H Loka- forl	O einber lo O	beinn lo O
H sjálfs- forl	O helber lo O	blygðunarlaus lo O
O andskotans lo	O himinhrópandi lo O	botnlaus lo O
O argasti lo	O hreinn lo O	fáheyrður lo O
O blákaldur lo	O hæfulaus lo O	freklegur lo O
O bölvður lo	O rakalaus lo O	hæfulífill lo O
O hatramlegur lo	O tilhæfulaus lo O	illgirnislegur lo O
O helvítis lo	O vísvitandi lo O	marghrakinn lo O
O hróplegur lo	P aðdróttun no kvk P	marklaus lo O
O hvítur lo	P H áróður no kk P	staflaus lo O
O illkvittinn lo	P bakbit no hvk P	svívírdilegur lo O
H -legur lo	P baknag no hvk P	uppspunninn lo O
O líkast lo	P baktal no hvk P	vanæmandi lo O
O líkindalegur lo	P blekking no kvk P	hreinn og beinn losamb O
O líkur lo	P blekkingaleikur no kk P	hreinn og klár losamb O
O óáheyrilegur lo	P bull no hvk P	afbökun no kvk P
O purkunarlaus lo	P dylgjur no kvk flt P	afflutningur no kk P
O púra lo	P fals no hvk P	alhæfing no kvk P
O rakinn lo	P fjarstæða no kvk P	atvinnurógur no kk P
O skefjalaus lo	P fleipur no hvk P	atyrðing no kvk P
H stór lo	P fyrirláttur no kk P	áburður no kk P
O svartasti lo	P fölsun no kvk P	álas no hvk P
O svartur lo	P getsakir no kvk flt P	álygar no kvk flt P
P bakmælgí no kvk	P gort no hvk P	bellibrögð no hvk flt P
H blað no hvk	P grobb no hvk P	blaður no hvk P
P blöff no hvk	P gróusaga no kvk P	blekkingaáróður no kk P
H kjaftur no kk	P raup no hvk P	illkvittni no kvk P
P kjaftæði no hvk	P rógburður no kk P	kjaftasaga no kvk P
H kvendi no hvk	P H rógur no kk P	landráðabrigsl no hvk flt P
P lastmæli no hvk flt	A sannleiki no kk A	lastmælgí no kvk P
P launmorð no hvk	A sannleikur no kk A	lausafrétt no kvk P
H laupur no kk	P sjálfsblekking no kvk P	látalæti no hvk flt P
O P lausung no kvk	P sjálfshól no hvk P	lokleysa no kvk P
P leikaraskapur no kk	P skítkast no hvk P	lygar no kvk flt P
H líf no hvk	P skrum no hvk P	lygi no kvk R P
P lífsblekking no kvk	P skrök no hvk R P	lögvilla no kvk P
P lífsflótti no kk	P slúður no hvk P	mannorðsmeiðing no kvk P
P lym ska no kvk	P svik no hvk flt P	mannorðsspell no hvk P
P mannhatu no hvk	P svindl no hvk P	mannskemmdir no kvk flt P
P mannlast no hvk	P sögu fölsun no kvk P	meiningarleysa no kvk P
H mál no hvk	P tilbúningur no kk P	mishermi no hvk P
P meiðyrði no hvk flt	P tvöfeldni no kvk P	moldviðri no hvk P
H munnur no kk	P undirmál no hvk flt P	mont no hvk P
H múr no kk	P uppspuni no kk P	mótsögn no kvk P
H mælir no kk	P útúr snúningur no kk flt	niðurrif no hvk P

Mynd 3: Brot úr samanburðarkorti lykilorða með flettunum lygi og ósannindi. Teiknið framan við lykilorðið vísar til þeirrar gagnategundar sem birtir viðkomandi vensl (m.a. A fyrir andheiti, H fyrir samsetningarlið, O fyrir orðasamband, P fyrir orðapar). Lykilorðin í miðdálkinum eru sameiginleg (teiknið tvítekið), hin koma fram með annarri flettunni.

Samanburður út frá orðapörum fæst ekki svo mark sé að nema byggt sé á gríðarstóru textasafni. Þar hefur *Tímarit.is* verið undirstaðan og skilað miklum og fjölbreyttum gögnum og aukið verulega við flettuforða orðanetsins.

Raunhæfasti mælikvarði á merkingarskyldleika sem orðapör hafa að bjóða er fólgin í því að greina fjölda sameiginlegra fylgdarorða (meðorða). Við slíka greiningu kemur til dæmis fram að orðin *lygi* og *ósannindi* tengjast bæði orðinu *óheilindi* (*lygi* og *óheilindi*, *óheilindi* og *ósannindi*) og *óheilindi* tengist jafnframt sumum þeirra orða sem mynda önnur orðapör með orðunum *lygi* og *ósannindi*. Þar sem hvert orðapar er aðeins skráð einu sinni burt séð frá textatíðni gætir lítt séreinkenna einstakra texta og virkustu fylgdarorðin skera sig glögglega úr í stóru safni.

Í gagnagrunni orðanetsins koma fram 132 orðapör með flettunni *lygi*. Samanburður á orðapörum leiðir í ljós að eftirtaldar flettur standa henni næst í þessari röð (fjöldi sameiginlegra fylgdarorða er í svigum): *ósannindi* (69), *rógur* (50), *rógburður* (44), *blekking* (36), *ósannsögli* (35), *uppspuni* (34), *fleipur* (32). Myndina má svo skerpa enn frekar með því að stilla saman tveimur flettum eins og gert er í töflu 1, þar sem fram kemur að virkustu tengsl flettnanna *lygi* og *ósannindi* eru að miklu leyti bundin sömu orðum. Töludálkarnir sýna fjölda orðapara, sá fremsti á við flettuna sem er yfirskrift samanburðarins, sá í miðið við flettuna framan við og í þeim aftasta er fjöldi orðapara þar sem fylgdarorðið er sameiginlegt þeim báðum. Þannig kemur orðið *lygi* fram í pörum eins og *lygi* og *skrök* og *tilbúningur* og *lygi*. Fylgdarorðin *skrök* og *tilbúningur* reynast svo að sínu leyti mynda pör með mörgum þeirra orða sem tengjast orðinu *lygi* á sama hátt, t.d. *fals* og *skrök* (sbr. *lygi* og *fals*), *tilbúningur* og *ímyndun* (sbr. *lygi* og *ímyndun*). Við þennan samanburð ber ekki aðeins að líta á fjöldann því hlutfallið skiptir einnig máli og flettur neðar á listanum (með færri orðapörum) geta sýnt hlutfallslega mikil tengsl.

lygi	ósannindi
ósannindi no hvk flt 132 209 69	lygar no kvk flt 209 204 96
rógur no kk 132 187 50	rógur no kk 209 187 76
rógburður no kk 132 117 44	lygi no kvk 209 132 69
blekking no kvk 132 87 36	rógburður no kk 209 117 62
ósannsögli no kvk 132 63 35	uppspuni no kk 209 80 45
uppspuni no kk 132 80 34	tilbúningur no kk 209 50 31
fleipur no hvk 132 76 32	skrök no hvk 209 41 30

tilbúningur no kk 132 50 29	dylgjur no kvk flt 209 38 22
spilling no kvk 132 395 25	raup no hvk 209 54 19
bull no hvk 132 73 22	skröksaga no kvk 209 31 16
fals no hvk 132 34 22	vitleysa no kvk 209 59 15
skrök no hvk 132 41 19	þvættingur no kk 209 20 14
ímyndun no kvk 132 80 18	skreytni no kvk 209 17 13
slægð no kvk 132 120 17	óheilindi no hvk flt 209 26 12
dylgjur no kvk flt 132 38 17	
hjátrú no kvk 132 145 16	
undirferli no hvk flt 132 50 16	
svik no hvk flt 132 65 15	
óheilindi no hvk flt 132 26 13	

Tafla 1: Virkni flettutengsla í orðapörum miðað við sameiginleg fylgdarorð. Samanburður á flettunum lygi og ósannindi sýnir að virkustu tengslin eru að mestu leyti við sömu orðin. Töludálkarnir sýna fjölda orðapara, sá fremsti á við flettuna sem er yfirskrift samanburðarins, sá í miðið við flettuna fremst í línunni og í þeim aftasta er fjöldi orðapara þar sem fylgdarorðið er sameiginlegt þeim báðum.

Í gagnagrunni orðanetsins er þegar kominn fram vitnisburður af þessu tagi um merkingarskyldleika fjölda orða, og stefnt er að því að hann verði sýnilegur á vefsíðunni *ordanet.is*. Að því tilskildu að orðapör séu virk gagnategund er hér um álitlegan mælikvarða að ræða sem tekur til mikils hluta flettufordans, sérstaklega nafnorða og lýsingarorða. Gagnvart samheitum og samheitalýsingu leggur hann til nýtt sjónarhorn með því að láta merkingarsamband orðanna styðjast við gildisröð og gera notkunarvirkni þeirra sýnilega. Um leið endurspegla þau merkingar- og notkunarþætti sem greina merkingarskyld orð að og draga fram sérkenni hvers um sig. Loks má bregða upp samanburði óháð fyrirfram hugmynd um merkingarskyldleika og leita stuðnings við ályktun um að tiltekna flettur séu merkingarlega óskyldar.

7 Birtingarform og vefsíðan *ordanet.is*

Hér að framan hefur innviðum orðanetsins verið lýst eins og þeim er fyrir komið í gagnagrunni verkefnisins. Grunnurinn er efnismeiri og margþættari en svo að honum sé ætlað að vera opinn til leitar. Í stað þess á orðanetið sér birtingarform á vefsíðunni *ordanet.is* þar sem greiningin skilar sér í áföngum eftir því sem flettur tengjast í einstökum

venslategundum og flokkuninni miðar áfram. Birtingarformið er enn í mótun og sumt af því sem hér verður nefnt er breytingum undirorpið. Því verður aðeins staldrað við meginrættina og þær hugmyndir um hlutverk og notagildi sem einkenna orðabókarlýsinguna.

Á vefsíðunni er athyglinni beint að innbyrðis venslum flettnanna, og fletturarnar koma fram sem grunneining, án þeirra orðasambanda og samsetninga sem þær kunna að geyma í gagnagrunninum. Aðild og hlutur einstakra flettna ræðst af því hvort og í hvaða mæli greiningin hefur skipað þeim í samband við aðrar flettur. Sums staðar hafa aðeins verið greind vensl milli tveggja flettna, annars staðar eru tengsl við hóp eða hópa flettna af ólíkum venslategundum.

Merkingarleg og hugtaksbundin vensl flettnanna eru í forgrunni og við greininguna byggist upp orðabókarlýsing sem sameinar hlutverk samheita- og hugtakaorðabókar. Með tilliti til þess að greiningin þarfnast frekari yfirferðar og samræmingar er fyrst í stað ekki greint til fulls á milli samheita og skyldheita. Sé um orðapör að ræða eru þau tilgreind sérstaklega, svo og andheiti. Merkingarflokkun undir hugtakaheitum er enn sem komið er einkum fyrir hendi meðal sagnaflettna. Þar er jafnframt vísað til skyldra merkingarflokka (hugtaka). Í stað þess að velja merkingarflokkunum heiti hefur hver flokkur tiltekna flettu í sínum hópi, eins konar forystuflettu, að yfirskrift. Með því er gert ráð fyrir að notendur nálgist lýsinguna og það sem leitað er að út frá formbundnum leitarstreng.

Flettulistinn veitir víða yfirsýn um formlega venslaðar flettur. Af flettustrengnum „bera“ spretta til dæmis fram nálega 500 stafrófsraðar flettur. Leit með algildisstaf getur svo tengt saman flettur með sameiginlegum lið. Þannig skilar leitarstrengurinn „*hest*“ yfir 800 flettum, strengurinn „* [hestur]“ rösklega 200 flettum með lýsingarorðum og strengurinn „* <hest*“ um 150 flettum í mynd sagnarsambanda.

Mörkunarstrengir fleiryrttra flettna eru ekki beinir leitarþættir á vefsíðunni en innan einstakra merkingarflokka má gera þá sýnilega og virkja þá til umröðunar á flettunum þar sem markið ræður röðinni. Á mynd 4 kemur fram breytileg fletturöð undir forystuflettunni **meiða sig**, eftir því hvort texti rofans hægra megin er **fela setningargerð** eða **sýna setningargerð**. Hér er miðað við að mörkunin nýtist notendum án þess að þeir þurfi að tileinka sér forsendur hennar og tilhögun og hafa fulla yfirsýn um skammstafanir og gildi þeirra.

HUGTAK
meiða sig (ss) fela setningargerð 

handleggsbrotna *so* hálsbrotna *so* hryggbrotna *so*
 höfuðkúpubrotna *so* lemstrast *so* limlestast *so* meiðast *so*
 nefbrotna *so* slasast *so* stórslasast *so* særast *so* ökklabrotna *so*
 verða fyrir <slysi> *so ao <no-d>* beinbrjóta sig *so fn-r-a*
 bráka sig *so fn-r-a* brenna sig *so fn-r-a* fleiðra sig *so fn-r-a*
 flumbra sig *so fn-r-a* fótbrjóta sig *so fn-r-a* handleggsbrjóta sig *so fn-r-a*
 hrómla sig *so fn-r-a* hrufla sig *so fn-r-a* hrumla sig *so fn-r-a*
 húðfletta sig *so fn-r-a* meiða sig *so fn-r-a* rispa sig *so fn-r-a*
 skaða sig *so fn-r-a* skaðbrenna sig *so fn-r-a* skera sig *so fn-r-a*
 skráma sig *so fn-r-a* slasa sig *so fn-r-a* snúa sig *so fn-r-a*
 stinga sig *so fn-r-a* hrufla sig til blóðs *so fn-r-a fs no-g*
 klóra sig til blóðs *so fn-r-a fs no-g* skera sig til blóðs *so fn-r-a fs no-g*
 stinga sig á <broddunum; nagla> *so fn-r-a fs <no-dg>*
 vera með glóðarauga *so fs no-a* vera með marblett *so fs no-a*

HUGTAK
meiða sig (ss) sýna setningargerð 

beinbrjóta sig *so* benja <hestinn> *so*
 bera blátt auga og brotið nef *so* bráka sig *so* bráka <beinið> *so* brenna sig *so*
 brjóta <beinið> *so* fá áverka *so* fá blátt auga og blóðugar nasir *so*
 fá glóðarauga *so* fá skeinu *so* fá skrámu *so* fá örkumsl *so* flaka í sárum *so*
 fleiðra sig *so* flumbra sig *so* fótbrjóta sig *so* handleggsbrjóta sig *so*
 handleggsbrotna *so* hálsbrotna *so* helmeiða <hestinn> *so* helta <hestinn> *so*
 hrómla sig *so* hrufla sig *so* hrufla sig til blóðs *so* hrufla <húðina> *so*
 hrumla sig *so* hryggbrotna *so* húðfletta sig *so* höfuðkúpubrotna *so*
 klóra sig *so* klóra sig til blóðs *so* laskast á <hendi, fæti> *so* lemjast *so*
 lemstrast *so* lesta <fótinn, handlegginn> *so* limlestast *so* meiða sig *so*
 meiðast *so*
meira ...

Mynd 4: Svipmynd af vefsíðunni ordanet.is þar sem raða má merkingarskyldum flettum á tvennan hátt, í stafrófsröð eða eftir setningargerð (mörkunarstreng).

8 Niðurlag

Í þessari grein hefur verið lýst orðabókarverkefni þar sem íslenskur orðaforði er látinn birtast í mynd rafrænnar orðabókar með áherslu á innbyrðis vensl, samstæðar heildir og samspil formbundinna og merkingarbundinna vensla. Í slíkri lýsingu er orðabókartextinn í sífelldri mótun og endurnýjun svo lengi sem haldið er áfram að greina gögnin og koma greiningunni á framfæri. Auk beinnar flokkunar getur gagnaefnið sjálft að nokkru leyti verið sýnilegt, notendum til yfirsýnar og glöggvunar.

Eins og ráða má af heiti greinarinnar er litið á gerð orðanetsins sem beint framlag til endurmótunar almennrar orðabókarlýsingar í samræmi við þær nýju forsendur sem rafræn gagna- og textavinnsla hefur skapað. Í stað þess að afmarka fjölda flettiorða og beina lýsingunni að einkennum hvers flettiorðs fyrir sig er tekist á við orðaforðann í heild sinni með áherslu á innbyrðis samhengi og vensl.

Lýsing á samheitum og öðrum merkingarskyldleika annars vegar og orðasamböndum hins vegar hefur ekki átt greiða samleið í hefðbundnum orðabókum. Hér kallast þetta tvennt á og orðasambandaefnið myndar uppistöðuna í merkingargreiningunni. Um leið taka merkingarbær orðasambönd sér stöðu sem sjálfstæðar flettur og flettulistinn umbreytist í safn merkingarlega einræðra flettna.

Orðabókarlýsing af þessu tagi á sér engin skýr takmörk og hún verður því fyllri því meira efni sem tiltækt er til greiningar. Meðal þess sem greiningin sýnir fram á er hversu virk orðin eru í setningarlegum og merkingarlegum venslum og þá virkni má bera að öðrum mælikvörðum á gildi orða innan orðaforðans, svo sem tölulegum upplýsingum um tíðni í rituðum textum.

Gagnvart hlutverki samheitaorðabókar er meginnýjungin fólgin í því að geta metið skyldleika tiltekinnar flettu við merkingarskyldar flettur með hliðsjón af setningarlegum venslum og styrkja með því þá samheita- og hugtakaflokkun sem byggð er á huglægu mati.

Með þessu sameinast ólík orðabókarhlutverk í samstæðri heild sem mikilvægt er að hafa yfirsýn um þegar staðnæmst er við lýsingu einstakra orða og gerð er grein fyrir breytileika þeirra í merkingarlegu og setningarlegu samhengi. Þá kemur m.a. í ljós hvaða merkingarbrigði eru virkari en önnur. Um leið skilar greiningin mikilvægum gögnum til margs konar athugana á íslenskum orðaforða, til annarra orðabókarverkefna og margs sem þar er við að fást. Það á meðal annars við um gerð merkingarskýringa, þar sem lengi hefur verið þörf á betri undirstöðugögnum.

Ritaskrá

- Anna Björk Nikulásdóttir og Matthew Whelpton. 2010. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17: 141–161.
- Dornseiff, Franz. 2004. *Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearbeitete und mit einem vollständigen alphabetischen Zugriffsregister versehene Auflage von Uwe Quasthoff. Berlin: Walter de Gruyter.

- Dönsk-íslensk orðabók*. 1992. Ritstjórar: Hrefna Arnalds og Ingibjörg Johanne-
sen. Reykjavík: Ísafoldarprentsmiðja.
- Ensk-íslenska orðabókin*. 2006. Jón Skaptason, ritstjóri. Reykjavík: JPV útgáfa.
Íslenskt orðanet: www.ordanet.is.
- Íslenskt textasafn*: http://arnastofnun.is/page/arnastofnun_gagnasafn_texta-safn.
- Jón Hilmar Jónsson. 2001. *Orðastaður. Orðabók um íslenska málnotkun*. Önnur
útgáfa, aukin og endurskoðuð. Reykjavík: JPV útgáfa.
- Jón Hilmar Jónsson. 2002. *Orðaheimur. Íslensk hugtakaorðabók með orða- og orða-
sambandaskrá*. Reykjavík: JPV útgáfa.
- Jón Hilmar Jónsson. 2005. *Stóra orðabókin um íslenska málnotkun*. Reykjavík:
JPV útgáfa.
- Jón Hilmar Jónsson. 2009a: Lexical description. An onomasiological approach
on the basis of phraseology. Í: Sandro Nielsen & Sven Tarp (ritstj.). *Lexi-
cography in the 21st Century. In honour of Henning Bergenholtz*, bls. 257–280.
Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Jón Hilmar Jónsson. 2009b. Lemmatisation of Multi-word Lexical Units: Mo-
tivation and Benefits. Í: Henning Bergenholtz, Sandro Nielsen & Sven
Tarp (ritstj.): *Lexicography at a Crossroads*, bls. 165–194. Bern: Peter Lang.
- Orðasambandaskrá Orðabókar Háskólans*: www.lexis.hi.is/osamb/osamb.pl
- Roget's Thesaurus of English Words and Phrases*. 1988. London: Penguin Books.
- Trap-Jensen, Lars. 2008. Tilgangs- og henvisningsstruktur i digitale ordbøger.
Overvejelser baseret på ordnet.dk. *Nordiske studier i leksikografi* 9. Rapport
fra konference om leksikografi i Norden, Akureyri 22.–26. maj 2007, bls.
447–464. Reykjavík.
- Whelpton, Matthew. 2012. From human-oriented dictionaries to computer-
oriented lexical resources – trying to pin down words. (Þetta hefti.)
- Þórdís Úlfarsdóttir. 2006. Málfræðileg mörkun orðasambanda. *Orð og tunga*
8: 117–144.
- Tímarit.is*: <http://timarit.is>

Abstract

The paper describes an ongoing lexicographic project, *Íslenskt orðanet* (Icelandic wordnet), which aims to analyse and describe the Icelandic vocabulary and its internal semantic relations on the basis of syntactic and morphological relations in word combinations and compounds. A comprehensive description of Icelandic phrasemes, which has been published in three phraseological dictionaries, together with a big collection of newspaper texts, provides the main material for the lexicographic analysis. Disambiguation and lemmatisation of multi-word lexical units plays a central role in the construction of the wordnet. The multi-word lemmas are grammatically tagged, which enables grammatical sorting and an interaction between semantic and grammatical features. The wordnet combines the functions of a synonym and a concept dictionary, paying special attention to measuring semantic relatedness between synonyms and near-synonyms on the basis of paratactic word-pairs. The product

of the analysis is open and accessible as an online dictionary on the website www.ordanet.is.

Lykilorð

orðanet, setningarleg vensl, fleiryrtar flettur, merkingarleg vensl, hugtakaorðabók, samheitaorðabók, orðapör, merkingarskyldleiki

Keywords

wordnet, syntagmatic relations, multi-word lemmas, semantic relations, thesaurus, synonym dictionary, word pairs, semantic relatedness

Jón Hilmar Jónsson

Stofnun Árna Magnússonar í íslenskum fræðum / Háskóli Íslands

Neshaga 16

107 Reykjavík

jhj@hi.is

