

Anna Björk Nikulásdóttir

Sjálfvirk greining merkingarvensla í *Íslenskri orðabók*¹

1 Inngangur

Með tilkomu orðabóka á tölvutæku formi opnast nýir möguleikar í orðabókahönnun. Þróunin í þá átt að hanna rafrænar orðabækur og önnur rafræn uppflettirit á sjálfstæðan hátt, óháð fyrirrennrum sínum á prenti, er þó enn á byrjunarstigi. Þar ræður mestu um að flestar tölvuorðabækur byggjast að einhverju eða öllu leyti á prentaðri útgáfu (Storrer 2001:54). Útgáfa á prenti lýtur öðrum lögmálum og því hefur ekki enn verið hægt að nýta möguleika rafræna formsins nema að hluta. Í bókaútgáfu er pláss veigamikill kostnaðarliður og er því kappkostað að hafa skýringar og aðrar upplýsingar í orðabókum sem knappastar. Skipulagi upplýsinganna eru einnig skorður settar, til að mynda skipa flestar orðabækur flettum í stafrófsröð. Fyrir orðabækur á tölvutæku formi hafa þessi atriði ekki eins mikið vægi. Ekki þarf að hugsa um að spara pláss, flettum og öðrum upplýsingum má raða á ýmsan hátt og gefa notanda kost á mismunandi upplýsingum allt eftir hans þörfum. Hér skiptir mestu að notendaviðmót og leitarmöguleikar séu sem best úr garði gerð.

¹Þessi rannsókn er hluti af lokaverkefni mínu til MA-prófs í tungutækni frá Ruprecht Karls-Universität Heidelberg, Þýskalandi.

Íslensk orðabók (ÍO) var gefin út af Eddu í tölvuútgáfu árið 2000 og í endurbætttri útgáfu árið 2003. Um nokkurt skeið hefur hún verið aðgengileg á vefnum (sjá *Vefbækur Eddu*) þar sem hún er í stöðugri þróun. Hvati rannsóknarinnar, sem hér verður lýst, var sá að það væri eftirsóknarvert að taka meira tillit til merkingarlegra þátta við uppsetningu orðabókarinnar en nú er gerlegt. Fyrsta skrefið í þessa átt hefur reyndar þegar verið tekið þar sem við leit að ákveðnu orði birtist listi þeirra flettna sem hafa leitarorðið í skýringu sinni og tengjast því leitarorðinu mjög líklega merkingarlega. Þessi breyting á framsetningu upplýsinga ætti að koma notendum til góða því ástæða þess að fólk flettir upp í orðabók er yfirleitt sú, að leitað er eftir merkingu ákveðins orðs (Herbst og Klotz 2003:33).

Markmið rannsóknarinnar var að kanna möguleikana á því að greina sjálfvirkt merkingarvensl milli nafnorðaflettna og orða úr skýringartextum þeirra. Í þeim tilgangi var forritið MERKOR þróað sem greinir 12 mismunandi merkingarvensl. Niðurstöðurnar sýna á ótvíræðan hátt að sjálfvirk greining merkingarvensla er raunhæfur möguleiki. Stór hluti nafnorðaflettna ÍO hefur verið tengdur við merkingarlega skyld orð, alls 96,45% merkingarliða.

Greining merkingarvenslanna byggist á orðflokkmynstrum skýringanna og því skilar rannsóknin einnig úttekt á því hvaða form skýringartexta eru best fallin til vélrænnar greiningar og hvaða form síður. Þessar upplýsingar gætu komið að notum við endurbætur á ÍO.

2 Aðferð

Lagt var upp með þá tilgátu að unnt sé að greina merkingarvensl milli flettna og orða í skýringum þeirra út frá orðflokkmynstri. Nokkrar rannsóknir af þessu tagi hafa verið gerðar fyrir önnur tungumál og má þar nefna greiningu ensku orðabókarinnar *Longman Dictionary of Contemporary English* (Alshawi 1987), greiningu basknesku orðabókarinnar *Euskal Hiztegia* (Agirre o. fl. 2000) og greiningu yfirheitastigveldis úr þýsku orðabókinni *Wörterbuch der deutschen Gegenwartssprache* (Geyken og Ludwig 2003). Skýringartextar orðabóka henta sérstaklega vel til greiningar eftir föstum mynstrum þar sem form þeirra er í nokkuð föstum skorðum. Algengast er að skýringar nafnorða tilheyri öðrum af eftirfarandi flokkum:

a) samheiti / jafnheiti²eða upptalning samheita / jafnheita.

(1) **fagnaður** 1 ánægja, gleði³

b) yfirheiti (*genus proximum*) ásamt nánari lýsingu, oft lýsingarorði eða -orðum (*differentia specifica*) (Svensén 1993:122).

(2) **breiðband** breitt tíðnisvið notað til fjarskipta, [...]

Einnig er samsetning beggja flokka algeng:

(3) **fagnaður** [...] 2 höfðinglegar móttökur, gleðskapur

Með því að greina orðflokkmynstur skýringanna og samhengi þeirra við merkingarvensl sést til dæmis að í (2) og (3) er höfuð (fyrsta) nafnliðarins yfirheiti flettunnar en hvert orð í upptalningu eins og í (1) er samheiti / jafnheiti hennar. Ýmis önnur mynstur koma fyrir í skýringatextum ÍO þó ekki séu þau eins algeng og þau í (1) – (3) og verður nánar fjallað um nokkur þeirra hér fyrir neðan.

Gögnin, sem rannsóknin byggist á, eru markaðir skýringartextar nafnorða úr gagnagrunni ÍO. Skýringartextarnir voru markaðir með TnT-málfræðimarkara Brants sem þjálfður hefur verið á íslensku.⁴ Stór kostur ÍO er að skýringum er skipt niður í svokallaðar flettutegundir í gagnagrunninum. Flettutegundir eru helsta viðfangsefni rannsóknarinnar og kýs ég að kalla þær skýringarhluta til einföldunar. Eftirfarandi skýring er til að mynda þrískipt:

(4) **dílaburkni** [1] íslensk burknategund [2] (*Dryopteris assimilis*) [3] af þrílaufungsætt, með fjaðurskiptum blöðum, vex í gjám og kjarri

Þetta skipulag auðveldar til muna vélrænan aðgang að einstökum hlutum skýringanna og skilar þar af leiðandi betri niðurstöðum en ef öll skýringin er vistuð sem ein heild. Í greiningu MERKOR er hver skýringarhluti meðhöndlaður sem sjálfstæð eining. Við hann er flettan tengd ásamt kenninúmeri sem og það orðflokkmynstur sem markarinn skilar. Skýringin í (4) hefur eftirfarandi greiningu:

³Sjá nánari umfjöllun um samheiti og jafnheiti í köflum 3.2 og 3.8.

³Öll orðabókardæmi eru úr vefútgáfu *Íslenskrar orðabókar* (sbr. *Vefbækur Eddu*). Málfræðiupplýsingum (kyni og beygingu) er sleppt í dæmunum.

⁴Gagnagrunnurinn var fenginn til rannsóknarinnar hjá Eddu útgáfu og kann ég þeim bestu þakkir fyrir og þá sérstaklega starfsmönnum orðabókadeildar, þeim Laufeyju Leifsdóttur og Marínó Njálssyni. Kærar þakkir einnig til Sigrúnar Helgadóttur hjá Orðabók Háskólans sem markaði skýringatextana.

- (5) íslensk burknategund l_nn⁵
 af þrilaufungsætt, með fjaðurskiptum blöðum, vex í
 gjám og kjarri⁶
 a_np_ _a_np_np_ _s_a_np_c_np

Markarinn skilar mun nákvæmari upplýsingum en hér sjást en einungis er notast við orðflokksameringuna, auk fallupplýsinga ef um nafnorð er að ræða. Upphafsorðflokkur hvers mynsturs segir mikið til um það hvaða merkingarvensl er von til þess að greina í viðkomandi skýringu. Mynstrunum er því skipt upp í flokka eftir því. MERKOR er skrifað á hlutbundna forritunarmálinu *Smalltalk* og greiningarklasinn hefur undirklasa fyrir hverja tegund af mynstrum. Greiningin byggist á reglustigveldum, einu fyrir hvern flokk af mynstrum. Í flestum tilvikum nægir að greina orðflokkamynstrið til þess að fá niðurstöðu, sú er raunin í fyrri hluta (5) með mynstrið l_nn. Stundum er þó nauðsynlegt að kanna skýringarstrenginn sjálfan og er seinni hluti (5) dæmi um það. Hér gefur orðið *af* sem upphaf texta með orðflokkamynstrið a_np_*, vísbendingu um að merkingarvenslin ÆTT sé að finna í skýringunni. Flettan *dílaburkni* hlýtur eftirfarandi greiningu í MERKOR:

- (6) YFIRHEITI (dílaburkni, burknategund)⁷
 EIGINLEIKI (dílaburkni, íslensk)
 ÆTT (dílaburkni, þrilaufungsætt)

Í næsta kafla verður fjallað nánar um merkingarvenslin og greiningu þeirra.

3 Merkingarvensl

Alls innihalda niðurstöðurnar 10 mismunandi merkingarvensl: yfirheiti, undirheiti, samheiti, eiginleiki, ætt, heildheiti, hlutheiti, hluti hóps, tengt lýsingarorð og tengt sagnorð. Auk þess voru jafnheiti og vísanir tekin með en þessi sambönd eru merkt sérstaklega innan orða-bókargagnagrunnsins. Ákveðið var að einskorða rannsóknina ekki

⁶Markarinn skilar niðurstöðum í samræmi við greiningarstrengi *Íslenskrar orðtíðni-bókar*: a = atviksorð, l = lýsingarorð, nn = nafnorð í nefnifalli, np = nafnorð í þágufalli, s = sögn, c = samtenging.

⁶*Dryopteris assimilis* er sleppt. Skýringarhluta á erlendu tungumáli er hægt að tengja við flettuna eftir á án þess að til komi sérstök greining.

⁷Lesist: *burknategund* er yfirheiti *dílaburkna*.

við gildandi merkingarvensl merkingarfræði orða heldur að leggja áherslu á að finna þau sambönd sem eru fyrir hendi í skýringatextum ÍO og gætu komið að notum við uppsetningu hennar.

3.1 Yfirheiti og undirheiti

Yfirheiti er algengustu merkingarvenslin sem er að finna í skýringatextum nafnorðaflettna ef jafnheitin eru ekki meðtalin, alls greindust 43.066 yfirheiti. Y er yfirheiti X ef $A \text{ er } X$ felur í sér $A \text{ er } Y$ en ekki öfugt:

- (7) Þetta er fólksbíll og þar af leiðandi er þetta farartæki.
*Þetta er farartæki og þar af leiðandi er þetta fólksbíll.

Fyrir yfirheitasambönd gildir, að ef Y er yfirheiti X þá er X undirheiti Y:

$$\text{YFIRHEITI}(X, Y) \rightarrow \text{UNDIRHEITI}(Y, X)$$

Þar af leiðir að ef yfirheiti flettu finnst í skýringartexta þá er flettan jafnframt undirheiti þess. Þannig geta orðið til yfirheitasigveldi innan flettna orðabókarinnar (undirstrikuðu orðin eru yfirheiti):

- (8) a. **glímukappi** frábær glímumaður
b. **glímumaður 1** keppandi í glímu **2** maður sem iðkar glímu
c. **keppandi** sá sem keppir, tekur þátt í keppni (t.d. í íþróttum)
d. **maður 1** tvífætt og tvíhent spendýr sem talar, [...] [...] 7 rún sem samsvarar m

Í (8c) kemur orðið *maður* ekki fyrir í skýringunni en er engu að síður greint sem yfirheiti. Þetta skýrist af því að skýringar, sem hefjast á orðunum *sá sem...*, vísa til þess að um manneskju sé að ræða og er reglan því sú að greina orðið *maður* sem yfirheiti flettna með þess konar skýringar (jafnvel þó að í einhverjum tilvikum gæti einnig verið vísað til dýra með viðkomandi flettu).

Þegar flettur eru tengdar saman í yfirheitasigveldi eins og í (8) verður að hafa í huga að yfirheitin, sem greind eru fyrir hverja flettu, geta haft margar mismunandi merkingar og í versta falli er um að ræða samhljóma flettur. Að svo stöddu er ekki mögulegt að segja til um hvaða merkingarliður flettu á við það orð sem greint er sem yfirheiti (eða önnur merkingarvensl) og því verður að taka stigveldunum

með fyrirvara. Í (8d) má til dæmis sjá að YFIRHEITI (maður, rún) er gilt en orðið *rún* er ekkert tengt orðunum *glímukappi* eða *keppandi* sem eru neðar í stigveldinu. Þrátt fyrir þetta gæti listi allra orða í slíku stigveldi með tengingu við viðkomandi flettu(r) komið að gagni við leit í orðabókinni. Greining yfirheita gefur einnig möguleika á því að finna öll undirheiti ákveðins orðs:

- (9) UNDIRHEITI (stelpa, (dugga, flekon, fruska, gafilhlað, gandála, glofra, hveðra, lausastelpa, skoffín, sleggja, stelpugopi, stelpuskjáta, stelputrippi, strákafledda, strækni, trilla, trýta, vammaskjóða)

Greining undirheita úr skýringum er ekki algeng. Eins og er skila einungis skýringar af forminu *X(-) eða (-)Y* undirheitum:

- (10) **fyrirtak 1** ágæti, úrvalsmaður eða -hlutur
 UNDIRHEITI (fyrirtak, úrvalsmaður)
 UNDIRHEITI (fyrirtak, úrvalshlutur)

3.2 Samheiti

Samheiti teljast þau orð sem í ákveðnu samhengi er hægt að skipta út hvort fyrir annað án þess að merking eða sannleiksgildi setningarinnar breytist:

- (11) a. *Merín* hans Jóns er inni í hesthúsi
 b. *Hryssan* hans Jóns er inni í hesthúsi

Algjör samheiti, það er að segja tvö orð sem geta í öllum hugsanlegum tilvikum komið hvort í stað annars, eru sjaldgæf eða ekki til (sjá t.d. Cruse 1986:265ff). Við greiningu samheita úr ÍO er heldur ekki leitað eftir algjörum samheitum í þessum skilningi heldur er gengið út frá því að flettur, sem eru skýrðar með einu orði eða upptalningu orða, hafi í ákveðnu samhengi sömu merkingu og orðið (orðin) í skýringunni, sbr. (1) hér að ofan. Samheiti eru reyndar ekki hátt hlutfall heildargreiningarinnar þar sem skýringar eins og í (1) eru yfirleitt merktar sem jafnheiti í gagnagrunni ÍO og hljóta greiningu samkvæmt því. Eftirfarandi eru dæmi um skýringar þar sem MERKOR greinir samheiti:

- (12) a. **rúmfatnaður 1** sængurfatnaður
SAMHEITI (sængurfatnaður, rúmfatnaður)
b. **hormotta** yfirvaraskegg
SAMHEITI (yfirvaraskegg, hormotta)
c. **garg** [...] fuglahljóð
SAMHEITI (fuglahljóð, garg)

Hér sést að niðurstöður samheitaagreiningarinnar eru ekki einsleitar. Dæmi (12a) sýnir samheiti sem geta yfirleitt, ef ekki alltaf, komið í stað hvort annars. Í (12b) er merking orðanna á vissan hátt sú sama en á þeim er stór stílmunur þannig að þau er ekki hægt að nota á víxl í hvaða samhengi sem er. Í (12c) er um að ræða yfirheiti og undirheiti: YFIRHEITI (garg, fuglahljóð). Upphaflega tilgátan um að skýringar sem eru eitt nafnorð (eða upptalning nafnorða) væru alltaf samheiti flettunnar stenst því ekki að öllu leyti.

3.3 Eiginleiki, tengt lýsingarorð

Merkingarvenslin eiginleiki og tengt lýsingarorð eru vensl milli nafnorðaflettu og lýsingarorðs úr skýringartextanum. Eiginleiki lýsir á einhvern hátt því sem flettan vísar til og oftast en ekki greinir eiginleiki flettuna frá systurum hennar í yfirheitastigveldi:

- (13) a. **busl** ólögulegt sund
b. **sprettsund** stutt sund

Yfirheiti beggja flettna í (13) er *sund* en þær hafa mismunandi eiginleika:

- c. EIGINLEIKI (busl, ólögulegt)
EIGINLEIKI (sprettsund, stutt)

Eins og er fylgja lýsingarorðin í niðurstöðunum eftirfarandi nafnorði í kyni og eru því ekki flettur í ÍÓ nema nafnorðið sé í karlkyni. Í útgáfu ÍÓ á netinu (sjá *Vefbækur Eddu*) er þó hægt að slá inn leitarorð í hvaða formi sem er þar sem flett er upp í *Beygingarlýsingu íslensks nútímamáls* (BÍN) ef leitarorðið finnst ekki sem fletta. Einnig stendur til að fara yfir niðurstöður MERKOR með hjálp BÍN, bæði til þess að setja orð merkingarvensla í kenniform og til þess að sigta út villur í greiningunni (sjá kafla 4). Hér má einnig benda á að vitanlega eru ekki öll orð, sem fram koma í greiningunni, flettur í ÍÓ (sjá einnig Kristínu Bjarnadóttur 1998:38–39).

Tengt lýsingarorð er lýsingarorð af sama stofni og flettan eða lýsingarorð með sömu eða svipaða merkingu og slíkt orð. Skýringar, sem innihalda tengt lýsingarorð, eru af forminu *það að vera X* eða *eitthvað X*:

- (14) a. **bæklun** það að vera bæklaður, lemstrun
TENGT_LO (bæklun, bæklaður)
- b. **dásemd** ágæti, e-ð aðdánlegt
TENGT_LO (dásemd, aðdánlegt)

3.4 Ætt

Skýringar orða úr plöntu- eða dýraríkinu hafa sérstöðu að því leyti að þær innihalda oft fræðilegar upplýsingar úr líffræði. Þetta eru upplýsingar sem notandi gæti verið að leita eftir þó að merking flettunnar sé honum ljós (Herbst og Klotz 2003:35). Sem dæmi má nefna að þó notandi viti að *grænlilja* sé plöntutegund áður en hann flettir upp í orðabók gæti vitneskjan um að hún sé af vetrarliljuætt nýst honum:

- (15) **grænlilja** íslensk plöntutegund (*Orthilia secunda*) af vetrarliljuætt, [...]
ÆTT (grænlilja, vetrarliljuætt)
YFIRHEITI (grænlilja, plöntutegund)
EIGINLEIKI (grænlilja, íslensk)

Þó segja megi að yfirheitið *plöntutegund* eigi einnig uppruna sinn í líffræði var ákveðið að halda þeirri greiningu í stað þess að setja upp fleiri sérhæfð sambönd fyrir þennan flokk flettna.

3.5 Tengt sagnorð

Nafnorð, sem leidd eru af sögnum (eða öfugt), eru oft skýrð með viðkomandi sögn í ÍO. Einnig nafnorð sem lýsa í raun verknaði án þess að til sé sögn af sama stofni. Þessar skýringar eru af forminu *það að X*:

- (16) a. **íhugun** það að íhuga
- b. **eftirför** það að elta

Í þessum tilvikum greinir MERKOR merkingarvenslin tengt sagnorð: TENGT_SO (eftirför, elta), TENGT_SO (íhugun, íhuga). Skýringar af þessu tagi innihalda oft orðasambönd:

- (17) **útungun** það að unga út eggjum, klak

MERKOR greinir einungis sambönd milli einstakra orða og greinir því ekki tengt sagnorð úr skýringum eins og þeirri í (17) enda væri TENGT_SO (útungun, unga) röng greining.

3.6 Heildheiti

„Heildheiti/holonym er heiti heildar e-s þegar hlutheiti/meronym X er heiti hluta þess. Y er heildheiti X ef X er hluti Y.“ (*Vefbækur Eddu*: Ensk orðabók). Greining heildheita úr ÍO fer eftir mynstrinu *hluti X / hluti af X*. Ef yfirheiti, sem finnst innan skýringartexta, er orðið *hluti* er næsta nafnorð á eftir að öllum líkindum heildheiti flettunnar. Í þessum tilvikum er hætt við að greina yfirheiti en heildheiti skilað í staðinn:

- (18) **drag** innsti hluti dals þar sem fjöllin eru lág og aflíð-andi
HEILDHEITI (drag, dals)

Skýringar af þessu tagi eru ekki hátt hlutfall skýringartextanna en þó greindust 382 heildheiti. Í einstaka tilvikum var sama heildheitið greint oftár en einu sinni og því hægt að fá lista yfir hlutheiti þess: HLOTHEITI (dals, (drag, framdalur, dalbotn, dalsmynni, uppdalur)). Heildheiti

er meðal þeirra merkingarvensla sem er nauðsynlegt að greina með hjálp BÍN til þess að hægt sé að hafa þau í nefnifalli í niðurstöðunum.

3.7 Hlutheiti

„Hlutheiti er heiti sem nær yfir samsetningarlið orðs, efni eða kjarna e-s, eða að eitthvað tilheyri ákv. hópi. X er hlutheiti Y ef X er hluti Y.“ (*Vefbækur Eddu*: Ensk orðabók). Hlutheiti eiga venjulega við áþreifanlega hluti og hluta þeirra, sígilt dæmi er *fingur* sem hlutheiti orðsins *hönd* (Cruse 1986:160–161):

- (19) Fingur er hluti handar
Hönd er með fingur

MERKOR greinir hlutheiti í skýringum sem hafa mynstur sem samsvara reglulegu segðinni $n_(_n_)*c_n_*$ þar sem samtengingin (c) er og:

- (20) **þvagfæri** nýru og þvaggangur
HLOTHEITI (þvagfæri, (nýru, þvaggangur))

Hlutheiti in í niðurstöðunum eru þó af mismunandi gerð og í mörgum tilfellum er ekki um að ræða eiginleg hlutheiti heldur tengd sambönd (sjá Cruse 1986:172ff). Prófunarsetningin *X er hluti Y* gildir til að mynda ekki fyrir öll samböndin. Það sem þau eiga sameiginlegt er að upptalning þeirra gefur nokkuð góða mynd af merkingu flettunnar. Eftirfarandi eru nokkur dæmi um hlutheiti og hvernig þau skiptast í flokka eftir því hvaða prófunarsetning gildir:

- (21) a. *X er hluti Y: HLUTHEITI* (Eyjaálfa, (Ástralía, Kyrrahafseyjar, Nýja-Sjáland))
- b. *X og Z mynda Y: HLUTHEITI* (brúðhjón, (brúðgumi, brúður))
- c. Það að vera *X* og *Z* felst í því að vera *Y: HLUTHEITI* (trúbador, (skáld, tónlistarmaður))

Sérstakur flokkur hlutheita er sá sem lýsir meðlimum hóps. Fyrir þessi vensl gildir að myndi hópur af *X Y*, þá er *X* hluti hóps *Y*. Þessi gerð greindist einungis 87 sinnum í ÍO en mynstrið, sem gildir fyrir þessi vensl, er *hópur (af) X*:

- (22) a. **holl** hópur manna, hluti heildar
HLUTI_HÓPS (holl, manna)
- b. **grind** [...] 5 hópur af hvölum
HLUTI_HÓPS (grind, hvölum)
- c. **gæslusveit** hópur eftirlitsmanna við tiltekið verkefni, einkum friðargæslu
HLUTI_HÓPS (gæslusveit, eftirlitsmanna)

Síðan kemur til kasta BÍN til þess að setja orðin *manna*, *hvölum*, *eftirlitsmanna* í nefnifall eintölu. Þá lítur (22a) svona út: HLUTI_HÓPS (holl, maður).

3.8 Jafnheiti og vísanir

Jafnheiti og vísanir eru sérstaklega merkt innan gagnagrunns ÍO og MERKOR framkvæmir enga frekari greiningu á þeim skýringum nema skipta þurfi upptalningum upp í einstök orð.

Hugtakið *jafnheiti* er reyndar ekki viðeigandi í þessu samhengi. Þetta hugtak er notað um „orð með samsvarandi merkingu“ (Jón Hilmar Jónsson 2005:26) í tvímála orðabókum. Hafa ber því í huga að jafnheiti í niðurstöðum MERKOR eru merkt sem slík í gagnagrunni

ÍO en lýsa ekki eiginlegum merkingarvenslum. Jafnheitaflokkun ÍO er notandanum ósýnileg og ekki er alltaf hægt að útskýra mun á flokkun skýringa:

- (23) **leppur 1** tuska [merkt í gagnagrunni sem skýring]
 [...]

 4 handbendi [merkt í gagnagrunni sem jafnheiti]

Við yfirferð niðurstaðna MERKOR komu í ljós nokkrir flokkar jafnheita. Þetta er ekki tæmandi listi þar sem dæmi voru valin af handahófi:

- (24) a. Jafnheitið er nánast sama orð og flettan:
afrak afhrak; **snjár** snjór; **ábreiðsla** ábreiða; **sauð-arlús** sauðalús
 b. Jafnheitið er samheiti flettunnar:
hettuselur blöðruselur; **aflgjafi** orkugjafi; **vefdag-bók** blogg síða
 c. Jafnheiti hefur samsvarandi merkingu og flettan en háð samhengi. Jafnheitin í skýringunni eru ekki endilega samheiti innbyrðis:
krakki barn, barnungi, krógi
blástur kaldur, stinningsgola, vindur, þurrkur
 d. Jafnheiti er undirheiti flettunnar:
sjómaður farmaður, fiskimaður
feiti viðbit

Einnig koma fyrir skýringar þar sem jafnheitin eru af mismunandi gerð. Í eftirfarandi dæmi eru samheiti, yfirheiti og undirheiti í sömu upptalningunni:

- (25) **dalur** dollari, peningur, ríkisdalur

Í öllum tilfellum ætti þó tenging við viðkomandi jafnheiti sem flettur (ef þær eru fyrir hendi í ÍO) að koma notandanum að gagni en hann verður að meta að hvaða leyti merking þeirra er samsvarandi því sem hann er að leita að.

Töluvert algengt er að lengri skýringartextar séu merktir sem jafnheiti. Þessi tilfelli eru ekki greind í þessari fyrstu útgáfu MERKOR en þau þyrfti að taka til athugunar, sem og hvort ekki væri ástæða til að merkja þessa skýringarhluta öðruvísi í gagnagrunninum. Eftirfarandi er dæmi um lengri skýringu sem er merkt sem jafnheiti:

- (26) **keppni** [...] 3 skipuleg samkoma eða mót þar sem menn reyna með sér í íþróttum eða öðru kappi

Vísanir eru merktar fyrir notanda og benda honum á að gagnlegt gæti verið að fletta upp á viðkomandi orði. Hér eru eftirfarandi merkingarvensl algeng:

- (27) a. undirheiti
gafi fótagafl, höfðagafl
 b. yfirheiti
bandprjónn prjónn
 c. andheiti
vörn sókn
 d. hlutheiti
brot nefnari, teljari

Einnig sambönd eins og lýst er í (24a):

- (28) a. **pepílrás** pöpulrás
 b. **fögnuður** fagnaður

Vísanir eru því tengdar flettunni á ólíkan hátt og ekki er alltaf ljóst af hverju eitt orð er skráð sem vísun en annað ekki. Af hverju er til dæmis *pöpulrás* í (27a) vísun en *snjár* í (23a) jafnheiti? Það að ákveðið ósamræmi sé til staðar skýrist væntanlega að einhverju leyti af því að efni ÍO er frá mismunandi tímum. Hluti skýringanna er óbreyttur síðan í upphaflegri útgáfu ÍO frá árinu 1963, aðrar hafa verið endurskoðaðar fyrir síðari útgáfur verksins og enn öðrum bætt við. Spurning er hvort vísanir í því formi sem þær birtast í prentaðri útgáfu eigi alltaf erindi í tölvuútgáfu orðabókar. Ef núverandi uppsetning ÍO á vefnum er tekin sem dæmi, má sjá að til vinstri við skýringartextana birtist listi af tenglum sem allir vísa í flettu innan orðabókarinnar og þetta er því eins konar listi með vísunum.

4 Niðurstöður

Reynslan af þessari fyrstu útgáfu af MERKOR sýnir að hægt er að ná góðum árangri í sjálfvirkri greiningu merkingarvensla. Með frekari fínstillingu reglnanna má reikna með að árangurinn geti batnað enn frekar. Þær niðurstöður sem þegar hafa fengist mætti byrja að nýta til þess að prófa áfram notendaviðmót og leitarmöguleika ÍO á netinu þó

ljóst sé að nokkur fjöldi orða sé tengdur merkingarvenslum við röng orð.

	alls	heimt
merkingarliðir	77.348	96,45%
skýringarhlutar	106.972	92,61%

Tafla 1: Heimt MERKOR úr skýringartextum nafnorðafletta Íslenskrar orðabókar

Þar sem hver merkingarliður flettu er greindur sjálfstætt er fjöldi merkingarliða notaður sem viðmiðun en ekki fjöldi flettna. Hver merkingarliður getur svo samanstðið af tveimur eða fleiri skýringarhlutum en hver skýringarhluti myndar eigið orðflokkamynstur sem er grundvöllurinn fyrir greiningu merkingarvensla. Eins og sjá má í töflu 1 hafa fundist merkingarvensl við 96,45% allra merkingarliða en það þýðir að við stærsta hluta merkingarliða nafnorðaflettna er hægt að setja vísun í orð sem stendur í merkingarvenslum við flettuna. Þessi tala er mikilvæg fyrir hönnuði og höfunda orðabókarinnar til þess að sjá hve algengt það er að fletta hafi vísun í merkingarlega tengt orð.

Hin talan, hlutfall greindra mynstra, sýnir hve hátt hlutfall orðflokkamynstra hlaut greiningu. Með því að skoða þau mynstur sem MERKOR hafnar mætti annars vegar bæta hönnun tólsins og hins vegar benda á aðra möguleika til þess að orða skýringar. Skýringarnar verða þó alltaf skrifaðar með notendur efst í huga og því er ekki við því að búast að hægt verði að greina allt vélrænt.

	fjöldi merkingarliða	hlutfall
rétt greining	786	82,13%
ófullnægjandi greining	121	12,64%
röng greining	50	5,22%
heimt	957	92,55%

Tafla 2: Niðurstöður prófunar (prófunarsett: 1034 merkingarliðir sem greindir voru handvirkir)

Prófunarsettið er tilviljunarúrtak heildargagnanna, um 1,34%. Þessir 1034 merkingarliðir voru greindir handvirkir og síðan keyrðir saman við niðurstöður úr MERKOR. Ef merkingarliður telst rétt greindur þýðir það að MERKOR fann öll merkingarvensl sem greind voru handvirkir

og ekki fleiri. Þegar MERKOR finnur eingungis hluta þeirra merkingarvensla sem greind voru handvirkt en greinir þó ekkert rangt er um ófullnægjandi greiningu að ræða. Samanlagt voru því 94,77% allra greindra merkingarliða villulausir. Rangt greindur er hver sá merkingarliður þar sem MERKOR greinir að minnsta kosti eitt rangt orð eða röng merkingarvensl, jafnvel þó hluti greiningarinnar sé réttur. Þessi aðferð prófunar var valin þar sem mikilvægt er fyrir höfunda ÍO að vita hve hátt hlutfall merkingarliða gætu haft rangt greind merkingarvensl til þess að meta hvort bæta eigi niðurstöðunum hráum inn í ÍO. Mögulegt væri einnig að gera villutölfræðina eftir flettum eða mynstrum en þær tölur gæfu ekki nógu nákvæma mynd af því hve margir merkingarliðir væru rangt greindir.

4.1 Helstu vandamál

Algengasta ástæða rangt greindra merkingarvensla eru villur í greiningu markarans. Nákvæmni markarans var ekki prófuð enda leiða ekki allar villur markarans til rangrar greiningar merkingarvensla og þar af leiðandi er ekki beint samband þar á milli. Í sumum tilvikum voru reglur MERKOR einnig lagaðar að villum markarans til þess að ná betri árangri. Eins og Þórdís Úlfarsdóttir bendir á (2006:141) er í „þessum stuttu textum [orðabókarskýringanna] lítið setningarlegt samhengi fyrir markarann að styðjast við“⁸. Árangur markarans er engu að síður ágætur eins og sjá má af því að villur, sem leiða til rangt greindra merkingarvensla, eru aðeins um 2% af prófunarsettinu.

Lýsingarorð eru oft greind sem nafnorð og MERKOR greinir þau þar af leiðandi sem yfirheiti:

(29) **bylgusa** skammvinn hríð, snjógangur [...]

Hér er greining markarans sú að *skammvinn* sé nafnorð og niðurstaða MERKOR var YFIRHEITI (bylgusa, skammvinn) í staðinn fyrir YFIRHEITI (bylgusa, hríð), EIGINLEIKI (bylgusa, skammvinn).

Önnur tegund af villum er þegar vensl eru út af fyrir sig rétt greind en gefa samt sem áður engar nothæfar upplýsingar. MERKOR útilokar nokkur slík orð við greininguna en lista yfir þessi orð þarf að lengja. Dæmi um slík orð úr niðurstöðum prófunarinnar:

⁸Þórdís Úlfarsdóttir fjallar í grein sinni um mörkun orðasambanda og tekur einnig dæmi um mörkun orðabókartexta.

- (30) a. **bull**a svipuð stöng í vél eða dælu
EIGINLEIKI (bull, svipuð)
b. **spaði** þannig lagað kvenhöfuðfat
EIGINLEIKI (spaði, lagað)

5 Form skýringartexta

Í sambandi við form skýringatexta eru tveir meginþættir sem þarf að athuga: Hvaða form eru vel eða illa til þess fallin að greina þau vélrænt og hvaða form eru algengust í nafnorðaskýringum ÍO. Nákvæmri greiningu er ekki lokið en hér verður bent á nokkur dæmi. Hafa ber í huga að tölurnar eru byggðar á niðurstöðum markarans og verður því að lesa þær með þeim fyrirvara að greining hans er ekki alltaf rétt. Eins og sést í töflu 1 hér að ofan er heildartala skýringarhluta 106.977. Flestir þeirra hefjast á nafnorði, alls 73.290 og þar af eru 35.158 einungis eitt nafnorð. Upptalning á nafnorðum eru 6.913 skýringarhlutar og 26.085 skýringarhlutar passa við mynstrið nafnorð_komma_frekarí skýring. Önnur mynstur eru til dæmis nafnorð_atviksorð_nafnorð og nafnorð_eignarfallseinkunn. Í skýringarhlutum, sem hafa mynstur sem hefist á nafnorði, er yfirleitt að finna samheiti eða yfirheiti og eru þær því vel fallnar til greiningar merkingarvensla. Sem stendur er litið fram hjá því sem stendur í svigum og því eru skýringar eins og eftirfarandi ekki greindar:

- (31) **handaburður** handa- (og handleggja)hreyfing(ar)

Önnur form skýringarhluta, sem skila áreiðanlegum niðurstöðum, eru form sem hefjast á lýsingarorði. Alls eru slík mynstur í nafnorðaskýringum ÍO 14.684 talsins, þar af kemur mynstrið lýsingarorð_nafnorð 5.672 sinnum fyrir. Þessi mynstur skila yfirleitt merkingarvenslunum eiginleika og yfirheiti. Að síðustu má nefna mynstur sem hefjast á fornafni en þau eru 7.362 talsins. Úr þessum mynstrum eru það fyrst og fremst merkingarvenslin tengd sögn (*það að X*) og orðið *maður* sem yfirheiti (*sá / sú sem...*) sem koma til greina.

Auk skýringarhluta eins og í (17) *það að unga út eggjum*, eru ýmiss konar form skýringa sem ekki er hægt að greina með tilliti til merkingarvensla:

- (32) a. **horgemlingur** einnig um horað fólk
 b. **umritun** um að venjulegum orðum er skipt út fyrir
 ný orð [...]

Þetta þýðir þó ekki endilega að engin merkingarvensl finnist fyrir viðkomandi merkingarlið flettu. Til dæmis skila aðrir skýringarhlutar viðkomandi merkingarliðs fyrir flettuna *umritun* (*það að umrita* og *um-orðun*) merkingarvenslum TENGTSO (*umritun*, *umrita*) og JAFNHEITI (*umritun*, *umorðun*).

Notkun sviga í skýringatextum skapar oft vandamál í vélrænni greiningu. Svigar eru oft notaðir til þess að draga fram atriði sem skýra merkingu orðsins betur eða til þess að benda á atriði sem yfirleitt (en ekki endilega alltaf) tilheyra merkingu orðsins. Herbst og Klotz benda á að innihald sviga skýri betur viðkomandi „Prototyp“ (2003:36). Eins og er er það sem stendur í svigum ekki tekið með í greiningu MERKOR og þær skýringar sem hefjast á sviga eru í heild sinni látnar eiga sig. Þetta verður þó skoðað betur í áframhaldandi þróun tólsins.

Í skýringum ÍO er að finna „margvíslegt misræmi“ (Mörður Árnason 1998:4) sem skýrist, eins og áður sagði, af sögu bókarinnar og af því hvernig hún hefur verið unnin. Notkun sviga er eitt af því sem þyrfti að vera skýrt skilgreint við endurskoðun skýringatexta ef auðvelda á vélræna greiningu. Í inngangi þriðju útgáfu *Íslenskrar orðabókar* eru skýringar um notkun tákna og greinimerkja. Þar segir: „Svigar eru notaðir í dæmum um uppflettiorð þannig að það sem er innan þeirra getur komið í stað þess sem stendur á undan (einnig notaðir á hefðbundinn hátt í skýringartextum, þ.e. um upplýsingar sem eru til nánari útskýringar á einhvern hátt).“ (*Íslensk orðabók* 2002: xiii). Dæmi um sviganotkun, sem hefur áhrif á núverandi greiningu MERKOR, eru lýsingarorð í sviga í upphafi skýringarhluta. Í eftirfarandi dæmum eru skýringarhlutar með lýsingarorðinu *lítill* skoðaðir. Það liggur í eðli lýsingarorða að þau eru til nánari skýringar og því fellur það hlutverk svigans niður. Annað hlutverk svigans, að innihald hans geti komið í stað orða utan svigans, virðist ekki vera algengt.

- (33) a. **smástrákur** (lítill) ungur drengur
 b. **stráklíngur** lítill, ungur drengur
 c. **prammi** (lítill eða) stór flatbotna bátur [...]
 d. **sproti** (lítill eða allvænn) silungur
 e. **sprotasilungur** (lítill) silungur
 f. **branda** (lítill) fiskur
 g. **skítseiði** lítill fiskur

Í (33a) gæti *smástrákur* þýtt til dæmis a. *lítill* og *ungur drengur*, b. *lítill eða ungur drengur* (orð í sviga kemur í stað orðs utan hans) eða c. *ungur drengur*, *yfirleitt lítill*. Til samanburðar má líta á skýringuna við *stráklíngur* og leiða að því getur að sviginn í (33a) sé í raun óþarfur. Í (33c) virðist lýsingarorðunum vera ofaukið því væntanlega getur *prammi* líka lýst bát sem hvorki er stór né lítill, *flatbotna* er hér lýsingarorðið sem skiptir máli. Sama á við í (33d) en hér er sviginn settur utan um bæði lýsingarorðin. Í greiningu merkingarvensla getur komið til mótsagna þegar skýringar eru orðaðar á þennan hátt: EIGINLEIKI (*prammi*, lítill), EIGINLEIKI (*prammi*, stór). Eins og áður segir er þessum skýringum sleppt í greiningunni eins og er. Annað dæmi um sviganotkun er þegar yfirheiti og undirheiti er í raun skeytt saman: (*ára*)*bátur*, (*vín*)*sopi* eða orðum eins og *breyting – tilbreyting*: (*til*)*breyting*, *skipti – umskipti* (*um*)*skipti*. Þetta skapar ákveðin vandamál við vélræna greiningu þar sem textanum er skipt upp í tóka fyrir mörkun. Þá stendur eitt orð eða eitt greinamerki í hverri línu. Fyrir orðið (*til*)*breyting* lítur það svona út:

(
 til
)
 breyting

Ekki er lengur hægt að sjá hvort orð í sviga tilheyra eftirfarandi orði (eins og í dæmunum hér á undan), undanfarandi orði eða eru sjálfstæð orð. Mögulega er hægt að skrifa reglur sem leysa úr þessu en það hefur ekki verið kannað til hlítar.

Þó hér hafi verið bent á dæmi um form skýringartexta, sem skapa vandamál við vélræna greiningu, er rétt að benda aftur á að 92,61% skýringarhlutanna voru greindir af MERKOR (sjá töflu 1).

6 Lokaorð

Í þessari grein var lýst tilraun til sjálfvirkrar greiningar merkingarvensla í *Íslenskri orðabók*. Greiningartólið MERKOR, sem þróað var í þessum tilgangi, greindi merkingarvensl milli nafnorðaflettna og orða úr skýringartextum þeirra. Alls voru greind merkingarvensl í 74.633 merkingarliðum sem jafngildir 96,45% heimt. Í prófunarsetti var heimtin lægri eða 92,55%. Nákvæmni var 82,13% ef einungis eru taldir merkingarliðir sem hlutu fullnægjandi greiningu en 94,77% ef allir merkingarliðir eru meðtaldir sem voru greindir villulaust. Þessar tölur benda til þess að með áframhaldandi þróun aðferðarinnar væri hægt að ná mjög góðum árangri.

Tenging á milli flettna og orða, sem standa í merkingarvenslum við þær, býður upp á margvíslega möguleika við uppsetningu orðabóka á tölvutæku formi. Einnig verður notagildi orðabókarinnar margþætta og má nefna sem dæmi að með hjálp niðurstaðna MERKOR er hægt að fá lista yfir allar löggildar iðngreinar sem er lýst í ÍO.

Stefnan í áframhaldandi þróun MERKOR er þríþætt: a. að kanna forsetningarliði í skýringartextum með það í huga að greina fleiri merkingarvensl, b. að kanna möguleikana á greiningu merkingarvensla fyrir sagnorðaflettur og lýsingarorðaflettur og c. að kanna hvort ekki sé hægt með lítilli fyrirhöfn að greina aðrar orðabækur og uppflettirit með aðstoð MERKOR.

Heimildir

Agirre, Eneko o. fl. 2000. *Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar*.

<http://arxiv.org/abs/cs.CL/0010025>. sótt: 21.01.2005

Alshawi, Hiyan. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *Computational Linguistics* Vol. 13, Nr. 3–4: 195–202.

Beygingarlýsing íslensks nútímamáls. Á vefsíðu Orðabókar Háskólans: <http://www.lexis.hi.is/beygingarlysing/>

BÍN = *Beygingarlýsing íslensks nútímamáls*.

Cruse, Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

- Geyken, Alexander og Rainer Ludwig. 2003. *Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache*. www.dwds.de/help/pages/ExtrHyp.pdf. sótt: 28.05.2005
- Herbst, Thomas og Michael Klotz. 2003. *Lexikografie*. Paderborn: Ferdinand Schöningh.
- ÍO = Íslensk orðabók.
- Íslensk orðabók. 2002 (3. útgáfa). Ritstjóri: Möður Árnason. Reykjavík: Edda.
- Jón Hilmar Jónsson. 2005. Aðgangur og efnisskipan í íslensk-erlendum orðabókum — vandi og valkostir. *Orð og tunga* 7: 21–40.
- Kristín Bjarnadóttir. 1998. Um skýringarorðaförðann. *Orð og tunga* 4: 33–43.
- Möður Árnason. 1998. Endurútgáfa „Íslenskrar orðabókar“. Stefna — staða — horfur. *Orð og tunga* 4: 1–8.
- Storrer, Angelika. 2001. Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. Í: Ingrid Lemberg, Bernhard Schröder og Angelika Storrer (ritstj.). *Chancen und Perspektiven computergestützter Lexikographie*, bls. 53–69. Tübingen: Max Niemeyer.
- Svensén, Bo. 1993. *Practical Lexicography. Principles and Methods of Dictionary-Making*. Oxford/New York: Oxford University Press.
- Vefbækur Eddu: Íslensk orðabók; Ensk orðabók: edda.is/vefbaekur. sótt: 31.08.2006
- Þórdís Úlfarsdóttir. 2006. Málfræðileg mörkun orðasambanda. *Orð og tunga* 8: 117–144.

Lykilorð

merkingarfræði orða, merkingarvensl, orðabókaskýringar, orðflokka-mynstur

Keywords

lexical semantics, semantic relations, dictionary definitions, POS-patterns

Abstract

In the design of electronic dictionaries it is possible to organize the information due to the meaning of the lexems. In this article a method for automatic extraction of semantic relations from dictionary definitions is demonstrated. Definitions of all noun lexems of a monolingual Icelandic dictionary, *Íslensk orðabók*, were analyzed. First, the definitions were tagged with Brants TnT-tagger which had been trained on an Icelandic corpus. From the tagged data the POS-patterns of the definitions were extracted and rules for extracting the semantic relations were developed. The rule algorithm was implemented in Smalltalk, resulting in the tool MERKOR. The results of the analysis were promising. The test was made with a random set of lexemes, about 1,34% of the data. For each lexeme the result could be completely right, that is all semantic relations from the reference data was found in the analysis of MERKOR, or it could be partly right, that is MERKOR did not find every relation compared to the reference data, but did nevertheless not extract any wrong word or relation. The recall turned out to be 92,55% and the precision was 82,13% (completely right analyzed) up to 94,77% (completely right plus partly right).

Anna Björk Nikulásdóttir
Seminar für Computerlinguistik
Ruprecht Karls-Universität Heidelberg
anna.b.nik@gmx.de