Anna B. Nikulásdóttir

# Tölvutækur merkingarbrunnur fyrir íslenska máltækni

## *Grunnur lagður að því að tölvur skilji merkingu í íslenskum textum*

## 1 Inngangur

Í þessari grein verður fjallað um þróun gagnagrunns sem inniheldur merkingarupplýsingar um íslensk orð. Gagnagrunninum er ætlað að koma að notum í íslenskri máltækni en einnig hefur máltækniaðferðum verið beitt við gerð hans. Þessar máltækniaðferðir sem og einstök merkingarvensl og formgerð merkingarbrunnsins eru viðfangsefni greinarinnar.

Orðanet er ungt hugtak yfir orðabókargögn sem mynda einskonar net merkingarlega tengdra orða. Orðunum geta fylgt hefðbundnar orðabókarskýringar en einkennandi fyrir orðanet er að frá hverju orði eru beinar vísanir í þau orð sem eru því merkingarlega tengd. Tölvutæk orðanet gegna orðið mikilvægu hlutverki í máltækni. Aðgangur að merkingarupplýsingum orða nýtist í fjölmörgum máltæknilausnum, svo sem merkingareinræðingu, upplýsingaheimt og stafsetningarleiðréttingu. Merkingareinræðing er mikilvæg fyrir allan hugbúnað sem greinir merkingu í textum. Lesendur velja yfirleitt ómeðvitað hvaða merking margræðs orðs á við í ákveðnu samhengi og leiða sjaldnast hugann að öðrum mögulegum merkingum. Ef hugbúnaður hinsvegar greinir t.d. orðin *heimsmeistarakeppni* og *box* í *heimsmeistarakeppnin í boxi* þarf hann að hafa aðgang að merkingar-

Orð og tunga 14 (2012), 19–38. 

upplýsingum sem tengja þessi orð þannig að merkingin ‚hnefaleikar' sé valin en ekki merkingin ‚kassi' fyrir orðið *box*. Slík greining er nauðsynleg til að mynda fyrir upplýsingaheimt sem miðar að því að greina efni fyrirspurna og texta og skila notanda efni sem er líklegt til þess að fela í sér svör við fyrirspurn hans.

Með hugtakinu **tölvutækur** er hér átt við að unnt sé að nýta gögnin við hugbúnaðargerð, að þau séu á formi sem hugbúnaður geti lesið og túlkað. Engin slík gögn með merkingarupplýsingum orða eru til fyrir íslensku. Þau íslensku orðabókargögn sem er að finna á vefnum, vefbókasafnið Snara[1] og *Íslenskt orðanet* (sjá grein Jóns Hilmars Jónssonar (2012) í þessu hefti), eru á tölvutæku formi í þeim hefðbundna skilningi að tölva getur lesið og sýnt gögnin en þau eru ekki hönnuð með það fyrir augum að hugbúnaður geti túlkað innihald þeirra. Með túlkun er hér átt við að hugbúnaður geti fengið svör við spurningum eins og t.d. *Hvaða orð tengjast orðinu* box*? Hvað hefur orðið* fugl *margar merkingar? Hvaða merkingarsviði tengist orðið* gjaldmiðill*? Hvaða orð eiga það sameiginlegt að hafa eiginleikann ‚soðinn'?* og unnið svo með svörin til þess að leysa það verkefni sem honum er ætlað.

Mikilvægt skref í áframhaldandi þróun gagna og tóla fyrir íslenska máltækni er því að til verði gagnagrunnur með merkingarupplýsingum um íslensk orð. Þegar þetta er skrifað er þróun slíks gagnagrunns vel á veg komin. Hann hefur hlotið nafnið *MerkOr — íslenskur merkingarbrunnur*, til aðgreiningar frá *Íslensku orðaneti*, og er væntanlega orðinn aðgengilegur nú (2012) í frumútgáfu[2].

Fjölmörg orðanet hafa verið byggð um allan heim að fyrirmynd Princeton WordNet, orðanets fyrir ensku (Fellbaum 1998; sjá einnig grein Matthew Whelpton (2012) í þessu hefti). Það hefur verið þýtt (hálf-) sjálfvirkt eða handvirkt á ýmis mál (sjá t.d. Fernández-Montraveta, Vázquez & Fellbaum 2008 og Lindén & Carlson 2010) og einnig hefur uppbygging þess verið lögð til grundvallar orðanetum sem byggjast á einmála nálgun (Pedersen et al. 2009). Við upphaf verkefnisins sem hér er kynnt var einmála nálgun valin þannig að þar eru íslensk gögn grundvöllur merkingarbrunnsins en ekki enska orðanetið.

Það er gífurlega tímafrekt og krefst mannafla að vinna orðanet

---

[1] http://snara.is

[2] Verkefnið er doktorsverkefni greinarhöfundar. Aðrir þátttakendur í því eru Dr. Matthew Whelpton sem aðalleiðbeinandi og verkefnisstjóri og Kristín Bjarnadóttir sem sérfræðingur og ráðgjafi. Það er hluti af verkefni sem hlaut Öndvegisstyrk RANNÍS árið 2009, *Hagkvæm máltækni utan ensku – íslenska tilraunin.*

handvirkt eins og gert hefur verið með WordNet. Einungis fámennur hópur íslensks fræði- og vísindafólks vinnur að því að koma upp sambærilegum gögnum og tólum fyrir íslenska máltækni og þegar eru til fyrir stærri málsamfélög. Til þess að gera það mögulegt að nothæf útgáfa af merkingarbrunninum yrði tilbúin sem fyrst, þrátt fyrir stærð verkefnisins og takmarkaðan mannafla, var ákveðið að beita sem mest sjálfvirkum aðferðum við vinnslu hans. Þær aðferðir ættu jafnframt að geta nýst við gerð samskonar merkingargagnagrunna fyrir önnur tungumál.

Áætlað er að merkingarbrunnurinn muni innihalda um 134 þúsund orð, 110.300 nafnorð, 6.300 sagnorð og 17.600 lýsingarorð. Þessar tölur gætu þó hafa breyst fyrir fyrstu útgáfu. Nafnorð eru meginuppistaðan í merkingarbrunninum og miða flestar greiningaraðferðirnar við að tengja nafnorð við önnur nafnorð en einungis að hluta til við sagnorð eða lýsingarorð.

Rannsóknin skiptist í þrjá meginhluta: a) undirbúning gagna; b) greiningu merkingarupplýsinga með mismunandi aðferðum; c) samþættingu niðurstaðna úr öðrum hluta. Fyrstu tveimur liðunum er lokið, búið er að greina mikinn fjölda merkingarvensla með mismunandi greiningaraðferðum en síðasti hluti verkefnisins mun felast í því að samþætta niðurstöðurnar og þannig að flokka vensl eftir áreiðanleika, einræða orð og vensl og jafnvel bæta við venslum.

Í þessari grein verður aðferðum sem notaðar hafa verið við greininguna lýst, sem og fjallað um einstök merkingarvensl og formgerð merkingarbrunnsins eins og hún er nú, fyrir samþættingu niðurstaðna.

Fyrsti kaflinn lýsir stuttlega þeim gögnum sem unnið var með og tilreiðslu þeirra. Meginhluti greinarinnar fjallar um greiningu merkingarvensla með mynsturgreiningu og greiningu merkingarupplýsinga með hjálp tölfræðiaðferða. Tölfræðiaðferðunum verður einungis lýst á almennan hátt. Áhugasömum lesendum er bent á Manning & Schütze (1999) þar sem er að finna nánari lýsingar og formúlur tengdar tölfræðiaðferðunum. Síðasta aðferðin sem notuð er í frumgerðinni byggist hvorttveggja í senn á mynsturgreiningu og tölfræði. Í sjötta kafla er sýnt dæmi um formgerð merkingarbrunnsins og borið saman við dæmi úr öðru merkingarneti og að lokum verður fjallað stuttlega um mat á niðurstöðum.

## 2 Gögn

Þær sjálfvirku aðferðir til greiningar merkingarupplýsinga sem notaðar voru byggjast á því að beita þeim á mikið magn texta. Í fyrstu voru aðferðir þróaðar og prófaðar á hluta *Markaðrar íslenskrar málheildar (MÍM)* (Sigrún Helgadóttir 2004) en lokagreining var gerð á *Íslenskum orðasjóði* (Erla Hallsteinsdóttir et al. 2008), textasafni sem safnað var af .is lénum frá árinu 2005, alls um 250 milljón orð. Textarnir voru markaðir og hlutaþáttaðir með *IceNLP* tólinu[3] (Hrafn Loftsson 2008, Hrafn Loftsson og Eiríkur Rögnvaldsson 2007).

Textar af netinu eru misjafnir að gæðum og Orðasjóðurinn inniheldur því töluvert af villum: stafsetningarvillum, innsláttarvillum o.fl. ásamt ýmsum upphrópunum, „áherslustafsetningu“ (t.d. *roooosalega*) og *ad hoc* orðmyndunum. Til þess að forðast það að slíkir strengir yrðu vistaðir í merkingarbrunninum voru öll orð með mörkum borin saman við gagnagrunn *Beygingarlýsingar íslensks nútímamáls (BÍN)*.[4] Orð sem höfðu sömu beygingarlýsingu í BÍN og samkvæmt *IceTagger* úr IceNLP voru lemmuð með viðeigandi uppflettiorði í BÍN. Þannig er tryggt að öll orð í merkingarbrunninum séu gild íslensk orð, þó að vissulega komi villur fyrir í lemmuninni.

## 3 Merkingarvensl og mynsturgreining

Þekkt aðferð til þess að greina merkingarvensl úr textum er að líta til ákveðinna setningafræðilegra mynstra (sjá t.d. Hearst 1992 og Girju & Badulescu 2006). Þessi aðferð hefur mér vitanlega þó ekki verið notuð á íslenska texta, ef frá er talin greining merkingarvensla úr *Íslenskri orðabók* (Anna B. Nikulásdóttir 2007).

Aðferðin eins og hún var kynnt hjá Hearst byggist á því að með hjálp orðapara sem standa í ákveðnum merkingarvenslum er leitað að setningafræðilegum mynstrum í textum sem eru líkleg til þess að vera lýsandi fyrir merkingarvenslin. Þannig voru til dæmis orðin *England* og *country* notuð til þess að finna mynstur sem gefa til kynna yfirheitavensl:

(1) *Countries* such as *England,* France and Spain

[3] http://icenlp.sourceforge.net
[4] http://bin.arnastofnun.is

(2) $NP_0$ such as {$NP_1$, $NP_2$ ... , (and | or)} $NP_n$ [5]

Mynstrið í (2) er dæmi um orða- og setningahlutamynstur (e. *lexico-syntactic pattern*) sem hægt er að nýta til þess að greina yfirheitavensl í textum. Fyrir hvert mynstur er skrifuð regla sem segir til um hvaða orð í birtingarmyndum mynstranna á að skrá og hvaða vensl gilda milli þeirra. Reglan tengd mynstrinu í (2) hljóðar þannig: $NP_0$ er yfirheiti $NP_1$ til og með $NP_n$. Þetta mynstur og fleiri mynstur sem Hearst kynnti í sinni grein hafa þá eiginleika að vera áreiðanleg en að vera jafnframt sjaldgæf í textum. Það er því einungis hægt að búast við að greina takmarkaðan fjölda af merkingarvenslum með þessari aðferð, jafnvel úr stórum textasöfnum.

Við þróun merkingarbrunnsins var mynstraaðferðinni beitt á nokkuð annan hátt. Markmiðið var að finna sem flest mynstur sem mögulega gæfu einhvers konar merkingarvensl til kynna, án þess að skilgreina fyrirfram hvaða vensl ætti að greina. Í stað þess að nota orð sem vitað er að standa í ákveðnum venslum til þess að finna mynstur í textunum (eins og *England* og *country* í dæminu hér að ofan, e. *seed-words*) var hlutaþáttað textasafn greint með tilliti til nafnliða og forsetningarliða. Hvert mynstur er samsett úr nafnliðum eða nafnlið(um) og forsetningarlið(um). Allar birtingarmyndir mynstranna voru vistaðar í gagnagrunni og þau mynstur sem komu minnst tíu sinnum fyrir í textasafninu voru rannsökuð sérstaklega. Mynstrin voru merkt eftir því hvort þau sýndust almennt innihalda merkingarlega tengd orð eða ekki og þá af hvaða tagi venslin voru. Dæmi:

(3) Gilt mynstur: [NP nheng][PP í aþ [NP nkeþg]][6]
Birtingarmynd: [NP *lánið* nheng][PP *í* aþ [NP *bankanum* nkeþg]]
Vensl: *lán* – í – *banki*

(4) Ógilt mynstur: [NP feveo [AP lveoof] nveo]]
Birtingarmynd: [NP *mína* feveo [AP *eigin* lveoof] *lopapeysu* nveo]]
Engin vensl

[5] NP: nafnliður

[6] Markastrengir IceTagger samsvara að mestu mörkunum sem notuð eru í *Íslenskri orðtíðnibók* (Jörgen Pind o.fl. 1991). Þannig merkir ‚nheng' nafnorð í hvorugkyni, eintölu, nefnifalli með greini og ‚aþ' atviksorð eða forsetningu sem stýrir þágufalli. Nákvæman lista er að finna í skjölun IceNLP. Við mynsturgreininguna var ekki tekið tillit til kyns orða.

Yfir 2.600 mynstur reyndust gefa einhverskonar merkingarvensl til kynna. Með því að nýta algrím[7] sem fellir saman mjög lík mynstur (e. *minimum edit distance*) (Ruiz-Casado, Alfonseca & Castells 2005) og reglulegar segðir var unnt að þjappa þessum mynstrum saman í 30 reglur fyrir greiningu merkingarvensla. Með þessum reglum voru 39 mismunandi vensl greind: yfirheiti, hliðstæð nafnorð, hliðstæð lýsingarorð, eiginleiki (no. – no.), eiginleiki (lo. – no.) auk 34 forsetningavensla. Tíðni venslanna er mjög mismunandi. Hliðstæð nafnorð og eiginleikavensl eru langalgengust en vensl byggð á forsetningunum *meðfram, eftir* (+ þf.) og *andspænis* eru mjög fá. Sem dæmi um merkingarvenslagreiningu fyrir eitt orð sýnir (5) hluta orða sem standa í merkingarvenslum við *málverk*:

(5) *málverk* – yfirheiti – *listmunur, listaverk*
*málverk* – hliðstæð no. – *teikning, ljósmynd, höggmynd, listaverk, ...*
*málverk* – eiginleiki (no.-no.) *listamaður, meistari, listasaga, málari*
*málverk* – eiginleiki (lo.-no.) *stór, nýr, frægur, fallegur, ...*
*málverk* – af – *stóll, landslag, atburður, haf*
*málverk* – úr – *myndröð*

Þessi vensl hafa verið greind úr textabútum eins og til dæmis *málverk og önnur listaverk; málverk, teikningar og ljósmyndir; málverk meistaranna; málverk af hafinu* o.s.frv. Venslin eru ýmist algild eins og *málverk* – yfirheiti – *listaverk*, eða eru einungis gild í ákveðnum tilfellum (ekki eru öll málverk fræg eða af landslagi). Orðið *listaverk* er að finna á tveimur stöðum í dæminu: sem yfirheiti (*málverk og önnur listaverk*) og sem hliðstætt orð (*málverk og listaverk*). Það er ekki óalgengt að mynsturgreiningin finni fleiri en ein vensl á milli tveggja orða og það verður hluti af vinnunni við samþættingu niðurstaðna að velja ein ákveðin vensl sem eiga að gilda fyrir hvert orðapar.

Forsetningavensl lýsa oft og tíðum sterkum venslum en samt sem áður er ekki unnt að skilgreina hver forsetningavensl á ótvíræðan hátt. Venslin *ull* – af – *kind* eru til dæmis annars eðlis en *málverk* – af – *landslag*. Í fyrri venslunum er um hlutheitavensl að ræða, *ull* – hluti_af – *kind*, en það er útilokað að skilgreina *málverk* – hluti_af – *landslag*. Hér stendur fyrra orðið en ekki það seinna fyrir heildina og

[7] **algrím** (e. *algorithm*): ákveðin röð af reglum og aðgerðum sem segir til um hvernig leysa eigi ákveðið verkefni.

efni málverksins, hér *landslag*, er ekki tengt við hlutinn *málverk* heldur einungis mynd af því. Einn þáttur í því að samþætta niðurstöður, sem er næsti áfangi verkefnisins, mun felast í því að kanna hvernig orð sem hafa sömu vensl við ákveðið eða ákveðin orð tengjast. Til að mynda finnast venslin *ull* – af – *X* fyrir orðin *fé*, *kind*, *sauðfé* og *rolla* sem sýnir að í einhverjum tilfellum gæti þessi aðferð verið árangursrík til þess að tengja skyld orð en þetta á þó eftir að kanna nánar.

# 4 Merkingartengsl

## 4.1 Útreikningur tengsla samkvæmt samhengi orða

Greining merkingarvensla með mynstraaðferðinni beinist að venslum tveggja orða sem koma fyrir í ákveðnu setningaliðamynstri (sjá (3)). Þá er litið til raðvensla orðanna. Við útreikning merkingartengsla (e. *semantic relatedness*) er hinsvegar litið til umhverfis orða. Merkingartengsl tengjast því frekar staðvenslum, þó ekki sé nauðsynlega hægt að skipta út merkingarlega tengdum orðum hverju fyrir annað.

Fyrir útreikning á merkingartengslum þarf að velja markorð og samhengisorð. Markorðin eru þau orð sem á að reikna út tengsl fyrir en samhengisorð eru þau orð sem tekið er tillit til við athugun á umhverfi markorðanna. Þessi orð er hægt að velja á ýmsan hátt, allt frá því að öll orð texta teljist hvort tveggja í senn, markorð og samhengisorð (Bullinaria 2008), til þess að velja einungis takmarkaðan fjölda og/eða flokka orða. Sem dæmi notuðu Cederberg & Widdows (2003) í sinni rannsókn 1000 algengustu orðin í málheildinni sem þeir unnu með sem samhengisorð og skilgreindu öll önnur orð sem markorð og Schütze (1998) valdi 2000 samhengisorð á móti 20 þúsund markorðum. Það hefur ekki verið sýnt fram á að ákveðið val markorða og samhengisorða gefi bestu niðurstöður. Við val á þessum orðalistum þarf m.a. að hafa í huga stærð málheildarinnar sem unnið er með og markmið útreikninganna. Í útreikningunum fyrir merkingarbrunninn voru 50 þúsund algengustu nafnorðin skilgreind sem markorð. Markmiðið var að vinna tengslaupplýsingar fyrir sem flest íslensk nafnorð. Stór hluti orðanna hefur þó mjög lága tíðnitölu (sbr. lögmál Zipf, sjá t.d. Manning & Schütze 1999:23) og þar sem ákveðin tíðni er nauðsynleg til þess að mögulegt sé að draga ályktanir út frá tölfræði er ekki hægt að nota öll nafnorðin í málheildinni. Fyrir

valið á samhengisorðunum var sett saman tíðnitafla nafnorða, sagnorða og lýsingarorða, eitt hundrað algengustu orðunum var sleppt og næstu 5000 skilgreind sem samhengisorð. Algengustu orðin voru ekki notuð þar sem þau eru ekki nægilega aðgreinandi, það eru til dæmis ekki sérkennandi upplýsingar fyrir orð að það komi fyrir í námunda við sögnina *vera.* Fjöldi samhengisorða var valinn með það í huga að geta lýst dæmigerðu umhverfi markorðanna sem best en að samhengisorðin hefðu samt sem áður ákveðna tíðni í málheildinni.

Þegar markorð og samhengisorð hafa verið valin þarf að skilgreina umhverfið eða samhengið sem á að kanna. Samhengið getur til að mynda verið afmarkað af ákveðnum fjölda orða í kringum markorð, svokölluðum orðaglugga, og einnig er hægt að tiltaka hvort kanna á samhengi vinstra megin, hægra megin eða báðum megin við markorðin. Margar rannsóknir hafa verið gerðar með mismunandi gerðum orðaglugga, en ekki hefur verið hægt að sýna fram á að ein ákveðin skilgreining sé árangursríkust (Sahlgren 2006). Í þessari rannsókn voru nokkrar tilraunir gerðar með mismunandi stærðir orðaglugga. Stærri orðagluggar, t.d. af stærðinni 25 (12 orð vinstra megin og 12 orð hægra megin við markorðin), reyndust nýtast vel til þess að skipta orðum upp í merkingarsvið. Til þess að marka sérkenni orðanna enn frekar skiluðu smærri orðagluggar betri niðurstöðum. Að endingu var orðagluggi af stærðinni sjö notaður, þ.e. þrjú orð vinstra megin og þrjú orð hægra megin við markorðin voru könnuð. Fyrir greininguna var búin til tafla þar sem hver lína stendur fyrir eitt markorð og hver dálkur fyrir eitt samhengisorð. Hver reitur í fylkinu[8] stendur fyrir það hve oft viðkomandi markorð (=lína) kemur fyrir með ákveðnu samhengisorði (=dálkur). Í upphafi stóð því talan 0 í öllum reitum og þegar samhengisorð fannst innan orðaglugga ákveðins markorðs var talan í viðkomandi reit hækkuð um einn. Að greiningu lokinni var því hvert markorð tengt við vektor[9] sem sýnir dreifingu orðsins í námunda við ákveðin samhengisorð og vektorinn er þannig lýsandi fyrir það samhengi sem orðið kemur fyrir í í textasafninu (skv. fyrirfram skilgreinda samhengishugtakinu). Kenningin sem liggur til grundvallar útreikningum á merkingartengslum er sú, að orð sem koma fyrir í svipuðu samhengi séu merkingarlega tengd (sjá t.d. Schütze 1993). Til þess að reikna út merkingartengsl markorða þarf

[8] **fylki** (e. *matrix*): tafla með línum og dálkum.

[9] **vektor** (e. *vector*): hverja línu eða hvern dálk í fylki má skilgreina sem vektor. Línuvektor samanstendur af reitum úr dálkunum í fylkinu, hver reitur stendur fyrir einn dálk. Línuvektorar fylkis með tíu dálka telja því tíu reiti.

því einungis að bera saman vektorana úr samhengisgreiningunni – því líkari sem vektorarnir eru því skyldari eru markorðin merkingarlega.

Tafla 1 sýnir tilbúið dæmi um fylki með tölum fyrir nokkur markorð með samhengisorðum. Fyrir hvert markorð er hægt að mynda vektor, sem dæmi *borðstofa* [7, 0, 5, 10, 0, 0]. Samanburður tveggja vektora felst í því að bera saman tölurnar í hverjum reit: fyrstu tölu í vektor a með fyrstu tölu í vektor b o.s.frv. Í töflu 1 eru vektorarnir fyrir *borðstofa*, *baðherbergi* og *þvottahús* svipaðir en vektorarnir fyrir *hljómplata* annarsvegar og *þorskur* hinsvegar skera sig úr og teljast því ekki tengjast öðrum markorðum í fylkinu.

| | *innrétting* | *hljómsveit* | *forstofa* | *borðkrókur* | *ýsa* | *afli* |
|---|---|---|---|---|---|---|
| *borðstofa* | 7 | 0 | 5 | 10 | 0 | 0 |
| *baðherbergi* | 11 | 0 | 9 | 9 | 0 | 0 |
| *þvottahús* | 8 | 0 | 9 | 11 | 0 | 0 |
| *hljómplata* | 0 | 8 | 0 | 0 | 0 | 0 |
| *þorskur* | 0 | 0 | 0 | 0 | 14 | 23 |

Tafla 1. *Tilbúið dæmi um fylki sem sýnir hve oft ákveðin markorð (línur) koma fyrir með samhengisorðum (dálkar).*

Frekari ákvarðanir sem þarf að taka við útreikning merkingartengsla lúta að vali á reikniaðferðum. Yfirleitt er samanburður vektoranna ekki framkvæmdur með því að bera beint saman niðurstöður greiningarinnar sem lýst var hér að ofan. Þær tölur segja ekki endilega til um hve sterk tengsl eru á milli markorðs og samhengisorðs. Til að mynda er dreifingin meiri og tölurnar hærri hjá algengum markorðum en þau gætu engu að síður verið merkingarlega skyld sjaldgæfari orðum. Á tölunum eru því framkvæmdir útreikningar sem auka upplýsingagildið, til dæmis með því að reikna út hve líklegt er að ákveðið markorð og ákveðið samhengisorð komi fyrir saman í textanum. Vektorarnir eru síðan bornir saman. Hér var notuð kósínus formúla sem mælir hve líkir vektorarnir eru (e. *cosine similarity*, sjá t.d. Manning & Schütze 1999:299, einnig almennt um þetta efni í kafla 8.5 í sömu bók). Með niðurstöðum samanburðarins er hægt að flokka markorðin eftir merkingartengslum: því nær tölunni 1,0 sem niðurstaða samanburðar tveggja vektora er, því skyldari eru orðin (sjá t.d. einnig rannsókn Bullinaria 2008).

Að ofangreindum útreikningum loknum var hvert markorð vistað með 14 skyldustu orðunum. Markorðin sem vistuð voru koma fyrir með minnst 10 samhengisorðum en þó ekki með fleiri en 3000

samhengisorðum en eins og áður sagði skila tölfræðiútreikningar fyrir mjög sjaldgæf og mjög algeng orð ekki góðum niðurstöðum. Dæmi um lista merkingarlega skyldustu orða er sýndur í (6):

(6) þorskur, tonn, ýsa, afli, fiskur, síld, steinbítur, veiðar, ufsi, kvóti, loðna, fisktegund, rækja, kolmunni, heildarafli

Í stað þess að telja einfaldlega orð innan orðaglugga má setja frekari skorður á samhengið og líta til setningahlutverka. Orð sem standa sem andlög með ákveðinni sögn hafa til að mynda oft einhverja sameiginlega eiginleika. Andlög með sögninni *að drekka* til dæmis vísa til einhvers konar vökva. Til þess að finna orð með svipaða eiginleika voru um 1.000 sagnir valdar sem samhengisorð og talið var hve oft markorð koma fyrir sem bein andlög þessara sagna. Sömu útreikningar voru svo framkvæmdir og fyrir talningu orða innan orðaglugga og sömuleiðis settir saman listar með tengdustu orðum. Dæmi um þetta er sýnt í (7).

(7) **þorskur,** fiskur, síld, ýsa, rjúpa, hvalur, rækja, tonn, fugl, ufsi, silungur, lax, sjóbirtingur, bleikja, loðna

Hér má greina nokkur merkingarsvið (e. *domain*) sem orðið *þorskur* tengist. Í (6) eru það ‚fiskur' og ‚fiskveiðar' og í (7) bætast við dýr sem tengja má við annars konar veiðar eins og ‚hvalveiðar' (*hvalur)* og ‚sportveiði' (*rjúpa, silungur*). Ef orð tengd orðunum í listunum eru skoðuð kemur í ljós að orð sem tengjast fiskveiðum (*kvóti, afli* o.s.frv.) koma oft fyrir með orðum í (6), og merkingarsviðið ‚matur' bætist við þar sem orð eins og *sósa* og *grænmeti* finnast í nokkrum listum. Með því að bera saman tengd orð á þennan hátt og jafnframt að kanna merkingarvenslin úr mynsturgreiningunni er stefnt að því að tengja orð við mismunandi merkingarsvið og greina hvaða sviði/sviðum orðin tengjast sterkast. Einnig verður litið til niðurstaðna úr þyrpingagreiningu í því samhengi (sjá kafla 4.2.).

## 4.2 Merkingarþyrpingar

Niðurstöður úr útreikningum á merkingartengslum er hægt að nýta til þess að mynda þyrpingar (e. *clusters*) merkingarlega tengdra orða. Þá er vektor orðs eða meðaltal vektora orða skilgreint sem miðja þyrpingar og orð sem hafa vektora sem reiknast nálægt þessari

miðju „þyrpast“ um hana (sjá t.d. Manning & Schütze 1999). Fyrir merkingarbrunninn voru tvær mismunandi þyrpingaraðferðir notaðar: *Clustering by Committee (CBC)* (Pantel & Lin 2002) og *Pole-Based Overlapping Clustering (PoBOC)* (Cleuziou, Martin & Vrain 2004). Fyrri aðferðin skilar frekar löngum listum orða sem tilheyra ákveðnum merkingarsviðum en niðurstöður PoBOC sýna heldur minni þyrpingar, allt niður í tvö náskyld orð (*almanaksár – reikningsár; tað – mykja*). Báðar aðferðirnar leyfa það að sama orðið tilheyri fleiri en einni þyrpingu og þannig geta mismunandi merkingar eða merkingaráherslur orða komið fram. Til að mynda má sjá í tveimur mismunandi þyrpingum úr PoBOC greiningunni að *þorskur* tengist merkingarsviði sjávarútvegs (sbr. (8)) en tilheyrir einnig merkingarþyrpingu sem inniheldur afurðir almennt (sjá (9)):

(8) **þorskur,** koli, kvóti, ufsi, krókabátur, línubátur, smábátur, steinbítur, þorskkvóti, útgerð, kvótasetning, ívilnun, grálúða, aflaheimild, línuveiði

(9) **þorskur,** fuglakjöt, kindakjöt, nautakjöt, innanlandsmarkaður, þorskafli, söluaukning, búvara, afurðaverð, mjólkurafurð

Orðið *þorskur* tilheyrir aftur á móti bara einni þyrpingu í CBC-greiningunni eins og sýnt er í (10):

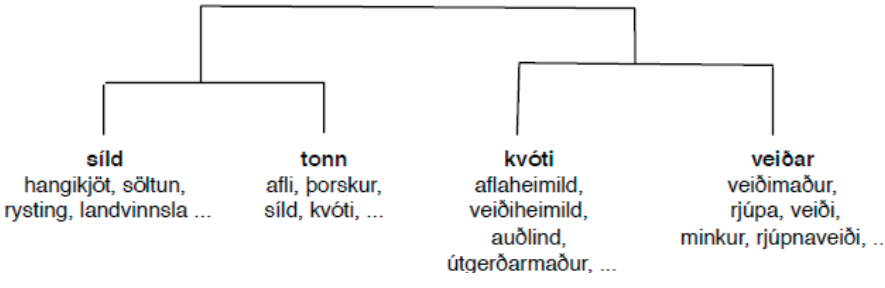
(10) **tonn,** afli, þorskur, síld, kvóti, veiðar, skip, togari, ýsa, loðna, kolmunni, króna, heildarafli, milljón, aflaverðmæti, útgerð, ufsi, steinbítur, vertíð, verðmæti, fisktegund, löndun, [...]

Þyrpingaraðferðirnar sem lýst er hér að ofan skila svokölluðum flötum þyrpingum. Þær mynda þyrpingar sem eru óháðar hver annarri og hver þyrping tilheyrir ákveðnu merkingarsviði. Þó myndast í sumum tilvikum fleiri en ein þyrping sem tilheyra sama merkingarsviði. Til þess að leitast við að tengja þessar þyrpingar innbyrðis og jafnframt að tengja þyrpingar með skyld merkingarsvið var annarri þyrpingaraðferð beitt, svokallaðri stigveldaaðferð (e. *hierarchical clustering*)[10]. Þá er fyrst leitað að þeim tveimur þyrpingum sem eru næstar hvor annarri og þær tengdar saman til þess að mynda nýja þyrpingu. Þannig vinnur algrímið sig upp þar til búið er að tengja allar þyrpingar

[10] Notast var við stigveldisþyrpingaralgrím úr LingPipe máltæknitólinu (http://alias-i.com/lingpipe, 30.06.2011).

saman. Ein allsherjarþyrping er vitanlega ekki það sem verið er að stefna að og því er leitað að þeim stað í sameiningarferlinu sem sýnir merkingarfyllstu skiptinguna. Allar merkingarlega skyldar þyrpingar ættu því að tengjast en óskyldar þyrpingar ekki.

Á mynd 1 er dæmi um stigveldisþyrpingu. Þyrpingarnar með orðin *síld* og *tonn* næst miðju eru skyldastar og mynda fyrst nýja þyrpingu. Þá eru þyrpingarnar með orðin *kvóti* og *veiðar* næst miðju tengdar saman og ný þyrping mynduð, og að síðustu eru þessar tvær nýju þyrpingar tengdar saman. Þyrpingarnar eru misjafnar að gæðum eins og búast má við af sjálfvirkri greiningu og einhver kann t.d. að undrast það að *hangikjöt* kemur fyrir í þyrpingu með *síld* og *söltun.* Það þýðir að þessi orð standa að einhverju leyti í svipuðu samhengi í málheildinni og í raun ekki svo fráleitt að *hangikjöt* tengist a.m.k. *söltun* að einhverju marki. Þess má geta að *hangikjöt* er einnig að finna með orðinu *jóladagur* í annarri þyrpingu tengdri merkingarsviðinu ‚veisluhöld'.




Mynd 1: *Stigveldisþyrping tengir saman skyldar þyrpingar*

## 5 Blönduð aðferð – mynsturgreining og tölfræði

*Structured Dimension Extraction and Labeling* (STRUDEL) (Baroni et al. 2010) er aðferð til þess að greina merkingarvensl milli orða samkvæmt mynstrum og reikna út líkindin á því að venslin eigi við. Þannig er mynstraaðferðinni og tölfræði blandað saman til þess að freista þess að bæta niðurstöður. STRUDEL vinnur ekki með fyrirfram skilgreind mynstur heldur notar einungis leiðandi reglur (e. *heuristics*) og takmarkanir (e. *constraints*) til þess að greina mynstur sem líkleg eru til þess að vísa á merkingarvensl. Markorð eru merkt sérstaklega í mörkuðum texta fyrir greiningu og forritið kannar umhverfi orðanna og greinir mynstur samkvæmt takmörkunum sem gefnar eru. Orða-

pörin sem tengd eru með þessum hætti lýsa oft óhefðbundnum venslum en samt sem áður lýsa tengdu orðin markorðinu oft á tíðum vel. Slík vensl er t.d. að finna í dæmi sem Baroni og félagar nefna í grein sinni um markorðið *book* sem stendur í venslum við orð eins og *reader* (*book – for – reader, reader – of – book*), *author* (*author – of – book, book – by – author*) og *library* (*library – of – book, book – in – library*). Eins og sjá má er hér notast við forsetningavensl eins og í mynsturgreiningaraðferðinni fyrir íslenska merkingarbrunninn.

Reglurnar og takmarkanirnar í STRUDEL miðast við ensku. Með lágmarksaðlögun forritsins var *Íslenskur orðasjóður* greindur með forritinu en eflaust væri hægt að bæta niðurstöður með því að bæta inn reglum og takmörkunum sem sérstaklega ættu við íslensku þótt ekki sé ljóst hvernig slíkar reglur myndu líta út. Fara þyrfti yfir kóðann í STRUDEL forritinu til þess að kanna að hvaða marki væri hægt að laga reglurnar að íslensku og hvort að einhverju leyti þyrfti að skrifa nýjar reglur. Um það bil 340.000 vensl úr greiningu á orðasjóðnum voru yfir þeim líkindamörkum sem höfundar STRUDEL miðuðu við í rannsókn sinni. Dæmi um vensl orðisins *mjólk* sem hafa há líkindagildi eru: *ábót – við – mjólk, drekka – mjólk, flóaður – mjólk, hella – mjólk, framleiða – mjólk, lítri – af – mjólk.*

Niðurstöðum STRUDEL greiningarinnar svipar að mörgu leyti til niðurstaðna mynsturgreiningarinnar: vensl eru ekki skilgreind fyrirfram og hér er einnig að finna forsetningavensl. Engin sagnorð koma þó fyrir í greiningu mynstraaðferðarinnar en hún skilar mun fleiri venslum. Fyrstu tilraunir með að tengja tölfræði við niðurstöður mynsturgreiningarinnar líkt og gert er í STRUDEL gáfu yfir 1 milljón vensla (af um 3,4 milljónum) sem eru nógu há líkindamörk til þess að teljast líkleg vensl. Við endanlegt mat á niðurstöðum verða niðurstöður þessara tveggja aðferða bornar saman sérstaklega til þess að greina nánar sameiginlega og mismunandi eiginleika.

## 6 Formgerð merkingarbrunnsins

Mikilvægustu venslin í orðanetum að WordNet fyrirmyndinni eru samheiti og yfirheiti (sjá einnig grein Matthew Whelpton (2012) í þessu hefti). Þau eru byggð upp sem heildstæð yfirheitastigveldi út frá grunnhugtaki eða -hugtökum. Frá öllum orðum í orðanetinu liggur leið upp eftir stigveldinu að einhverju grunnhugtaki sem getur

verið t.d. TILFINNING eða HLUTUR. Þannig mætti hugsa sér að í íslensku hefði orðið *ofsagleði* yfirheitið *gleði* sem aftur tengdist grunnhugtakinu TILFINNING.

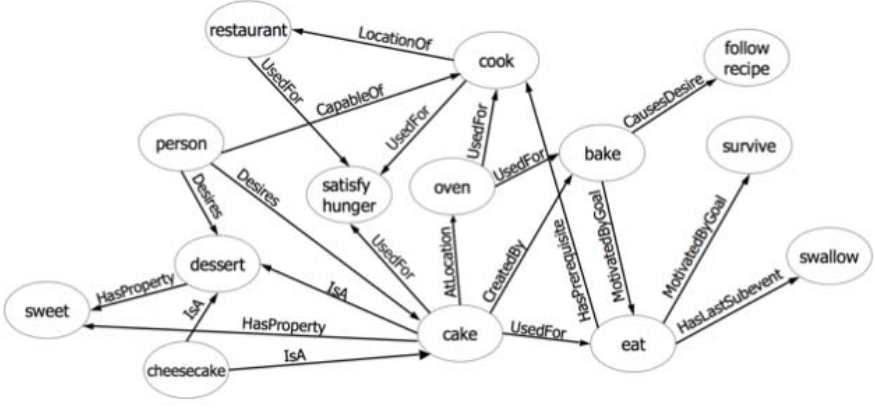
Formgerð íslenska merkingarbrunnsins hefur ekki verið greind að fullu en eins og sjá má hér að ofan eru merkingarupplýsingarnar um einstök orð margvíslegar og ekki alltaf nákvæmlega skilgreinanlegar. Ólíklegt er einnig að merkingarbrunnurinn myndi heildstætt net orða. Frekar má búast við að orð innan einstakra merkingarsviða tengist innbyrðis og myndi þannig þyrpingar sem eru einangraðar að mestu leyti.

Einstök hefðbundin merkingarsvið geta svo myndað einskonar undirsvið. Í tengslum við dæmin í kafla 3, orð sem tengjast *þorskur*, má til dæmis nefna að merkingarsviðinu ‚fiskur' má mögulega skipta í þrjú svið eftir niðurstöðunum: a) svið sem tengist umræðu um fiskveiðar og útgerð (*þorskur, loðna, kolmunni*), b) svið sem tengist sportveiði (*lax, sjóbirtingur, silungur*) og c) svið sem tengist mat (*ýsa, skötuselur, rauðspretta*). Þannig fást viðbótarupplýsingar sem tengjast daglegu máli og almennri þekkingu, sem sjaldan er að finna í hefðbundnum orðabókum. *Íslensk orðabók* til að mynda skilgreinir orðin *ýsa* og *kolmunni* á sama hátt: „fiskur [latneskt heiti] af þorskaætt" (Snara, 30.06.2011). Í merkingarbrunninum hins vegar er að finna upplýsingar um að *ýsa* sé borðuð, ýmist steikt, soðin eða djúpsteikt, geti verið í kvöldmatinn og verið með kartöflum. Orðið *kolmunni* tengist hins vegar eingöngu öðrum fisktegundum og orðum tengdum útgerð og fiskveiðum.

Merkingarnetið *ConceptNet* (Havasi, Speer og Alonso 2007) inniheldur merkingarvensl milli hugtaka. Takmark höfunda þess er að til verði gagnagrunnur sem nýta má í ýmsum hugbúnaði sem þarfnast merkingarupplýsinga sem tengjast almennri reynslu og þekkingu. Stór hluti af hæfileikum okkar til þess að skilja skilaboð byggist á því sem við vitum og höfum reynt í umhverfinu, þekkingu sem oft er sameiginleg hverju samfélagi. Ef einhver segir til að mynda *ég bakaði köku í gær* er ólíklegt að hann taki sérstaklega fram að kakan hafi verið bökuð í ofni, því það er sjálfgefið að bakstur fer fram í ofni. Þekking af þessu tagi þarf hins vegar að vera fyrir hendi í tölvutækum merkingarnetum því tölvan býr ekki yfir neinni fyrirfram gefinni þekkingu.

Á mynd 2 er lítið dæmi úr ConceptNet. Grunneiningin er hugtak en ekki orð eins og í merkingarbrunninum og því er að finna fleiryrtar framsetningar eins og *satisfy hunger* og *follow recipe*. ConceptNet

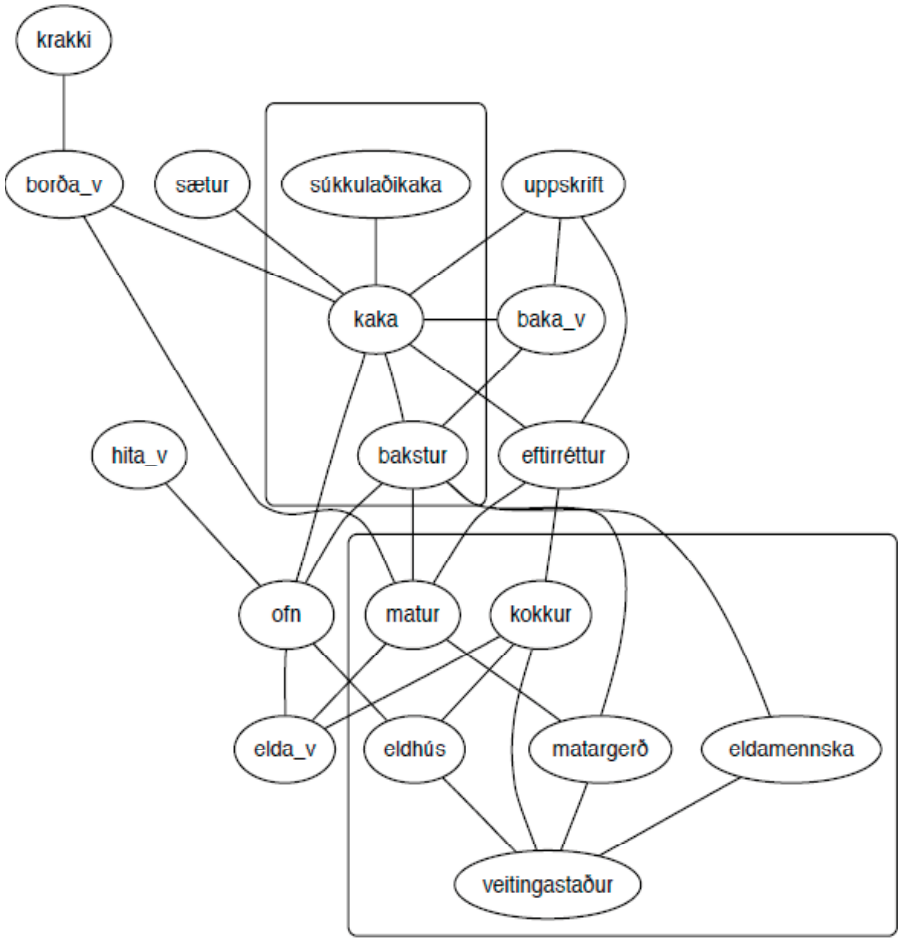
inniheldur 21 merkingarvensl og að auki ein vensl sem kallast *ConceptuallyRelatedTo* sem eru ekki nánar skilgreind. Á mynd 2 má sjá vensl eins og *UsedFor, HasProperty* og *IsA*. Venslin *HasProperty* eru sambærileg við venslin eiginleiki í merkingarbrunninum og *IsA* eru yfirheitavensl. Merkingarbrunnurinn inniheldur einnig vensl sem kalla mætti *ConceptuallyRelatedTo*, til að mynda merkingarlega skyld orð sem tilheyra sama merkingarsviði án þess að hægt sé að skilgreina venslin nákvæmlega.




Mynd 2: *Dæmi um vensl í ConceptNet*[11]

Mynd 3 sýnir sambærilegt dæmi úr merkingarbrunninum. Eins og sjá má er margt sameiginlegt með dæmunum. Munurinn orsakast fyrst og fremst af grunneiningum gagnagrunnanna. Merkingarbrunnurinn er byggður út frá nafnorðum þannig að eins og er eru vensl á milli sagna ekki fyrir hendi (sbr. *eat – swallow, bake – eat* á mynd 2). Eins er þar einungis að finna einstök orð en ekki fleiryrt hugtök eins og til að mynda *seðja hungur* eða *fylgja uppskrift*. Venslin eru ekki merkt inn á íslenska dæmið þar sem merkingarbrunnurinn er ennþá í vinnslu, þ.e. ekki hafa öll vensl fengið nafn og einnig geta fleiri en ein vensl verið á milli orða. Kassarnir tveir tákna að orðin innan þeirra tilheyra sömu merkingarþyrpingu. Þannig tengjast *kaka, súkkulaðikaka* og *bakstur* sérstaklega sem og *matur, matargerð, eldamennska* o.s.frv. Einnig er vert að taka fram að hér eru einungis sýnd einstök dæmi, mun fleiri orð tengjast þeim sem hér eru sýnd.

[11] http://csc.media.mit.edu/conceptnet




Mynd 3: *Dæmi um vensl í Íslenskum merkingarbrunni*

## 7 Samþætting og mat á niðurstöðum

Dæmin sem hér hafa verið sýnd eru úr merkingarbrunninum eins og staða hans er eftir að einstökum greiningaraðferðum hefur verið beitt. Gera má ráð fyrir að töluvert sé um villur í sjálfvirku greiningunni og því er næsta skref að bera saman niðurstöður mismunandi aðferða og nýta samanburðinn til þess að reikna út áreiðanleika venslanna. Til að mynda má gera ráð fyrir að ef tvö orð tengjast samkvæmt mörgum greiningaraðferðum auki það líkurnar á því að orðin séu í

rauninni tengd. Í þessu ferli verða einnig möguleikar til einræðingar kannaðir, meðal annars með því að athuga hvort þau orð sem tengjast ákveðnu orði tilheyri mismunandi merkingarsviðum. Sem dæmi má nefna að orð sem tengjast orðinu *olía* tengjast líka ýmist orðum af merkingarsviðinu ‚orka' (*bensín, kol*), ‚matargerð' (*panna, smjör*), ‚myndlist' (*strigi, pensill*) eða ‚snyrting og vellíðan' (*krem, nudd*). Út frá þessu væri hægt að skilgreina fjórar merkingar orðsins *olía* og aðskilja þær í gagnagrunninum.

Til þess að meta gæði sjálfvirku greiningarinnar og hvort samþætting niðurstaðna skilar árangri verður tilviljunarúrtak metið. Matið verður í höndum meistaranema sem mun fara yfir úrtak úr niðurstöðunum fyrir og eftir samþættingu.

Óhjákvæmilegt er að í niðurstöðum sjálfvirkrar greiningar, eins og hér hefur verið lýst, leynist villur. Til þess að auðvelda vinnu við að fara yfir gagnagrunninn handvirkt verður þróað notendaviðmót með verkferlum til þess að bæta við, eyða út og leiðrétta vensl. Þess konar leiðrétting mun vitanlega taka töluverðan tíma en vonast er til að merkingarbrunnurinn nýtist frá upphafi þrátt fyrir að eitthvað verði um villur. Tilraunir með tengingar við máltæknihugbúnað munu leiða það í ljós.

## 8 Lokaorð

*Íslenskur merkingarbrunnur* er tölvutækt merkingarnet sem unnið er með sjálfvirkum aðferðum. Aðferðirnar byggjast á mynsturgreiningu og tölfræði og miða að því að greina merkingarupplýsingar orða úr stóru textasafni.

Niðurstöðurnar sýna fjölbreytt merkingarvensl og flokkun orða eftir merkingarsviðum. Alls eru um 134 þúsund nafnorð, sagnorð og lýsingarorð í merkingarbrunninum og vel á aðra milljón vensla. Þessar tölur eru þó ekki endanlegar þar sem enn er unnið að síðasta hluta verkefnisins, sem felst í því að samþætta niðurstöður einstakra greiningaraðferða. Markmiðið er að kanna hvernig niðurstöður mismunandi aðferða styðja eða hrekja einstök vensl og meta þannig hvaða vensl eru líkleg til þess að vera rétt og hver síður. Einnig verða möguleikar einræðingar kannaðir.

Þótt formgerð merkingarbrunnsins sé nokkuð önnur en formgerð *Princeton WordNet*, er stefnt að því að gera tilraun með að tengja hluta

merkingarbrunnsins við svokallaðan kjarnaorðaforða WordNet (e. *core WordNet*) í tengslum við verkefnið MetaNord[12].

Merkingarbrunnurinn hefur verið öllum opinn frá því í byrjun árs 2012. Fyrst og fremst er stefnt að því að hann komi að gagni í hugbúnaðarþróun, en einnig má hugsa sér annars konar nýtingu, til dæmis við rannsóknir og sem viðbót við hefðbundna orðabókanotkun.

## Heimildir

Anna B. Nikulásdóttir. 2007. Sjálfvirk greining merkingarvensla í *Íslenskri orðabók*. *Orð og tunga* 9: 5–24.

Baroni, Marco, Brian Murphy, Eduard Barbu & Massimo Poesio. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science* 34: 222–254.

*BÍN = Beygingarlýsing íslensks nútímamáls.* http://bin.arnastofnun.is. (30. júní 2011)

Bullinaria, John A. 2008. Semantic Categorization Using Simple Word Co-occurrence Statistics. Í: M. Baroni, S. Evert & A. Lenci (útg.). *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics,* bls. 1–8. Hamburg, Þýskalandi.

Cederberg, Scott & Dominic Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. Í: *Proceedings of the International Conference on Natural Language Learning (CoNLL),* bls. 111–118. Edmonton, Kanada.

Cleuziou, Guillaume, Lionel Martin & Christel Vrain. 2004. PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classication and Textual Data. Í: *Proceedings of the 16th European Conference on Artificial Intelligence,* bls. 440–444. Valencia, Spáni.

*ConceptNet.* http://csc.media.mit.edu/conceptnet. (30.06.2011)

Erla Hallsteinsdóttir, Thomas Eckart, Chris Biemann, Uwe Quasthoff & Matthias Richter. 2007. Íslenskur orðasjóður – Building a Large Icelandic Corpus. Í: Joakim Nivre, Heiki-Jaan Kaalep & Kadri Muischnek (útg.). *Proceedings of NODALIDA-07*, bls. 288–291. Tartu, Eistlandi.

Fellbaum, Christiane (útg.). 1998. *WordNet. An Electronic Lexical Database*. Cambridge Mass., London: MIT Press.

Fernández-Montraveta, Ana, Gloria Vázquez & Christiane Fellbaum. 2008. The Spanish Version of WordNet 3.0. Í: A. Storrer, A. Geyken, A. Siebert & K.-M. Würzner (útg.). *Text Resources and Lexical Knowledge*, bls. 175–182. Berlin, New York: Mouton de Gruyter.

[12] http://www.meta-n et.eu/projects/meta-nord/

Girju, Roxana & Adriana Badulescu. 2006. Automatic Discovery of Part-Whole Relations. *Computational Linguistics* 32(1): 83–134.

Havasi, Catherine, Robert Speer & Jason B. Alonso. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. Í: *Proceedings of Recent Advances in Natural Language Processing.* Borovets, Búlgaríu.

Hearst, Marti A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. Í: *Proceedings of COLING-92*, bls. 539–545. Nantes, Frakklandi.

Hrafn Loftsson. 2008. Tagging Icelandic Text: A Linguistic Rule-Based Approach. *Nordic Journal of Linguistics* 31(1): 47–72.

Hrafn Loftsson & Eiríkur Rögnvaldsson. 2007. Ice-Parser: An Incremental Finite-State Parser for Icelandic. Í: Joakim Nivre, Heiki-Jaan Kaalep & Kadri Muischnek (útg.). *Proceedings of NODALIDA-07*, bls. 128–135. Tartu, Eistlandi.

*IceNLP.* http:// icenlp.sourceforge.net. (30.06.2011)

*Íslenskt orðanet.* http://www.ordanet.is. (30.06.2011)

Jón Hilmar Jónsson. 2012. Að fanga orðaforðann: orðanet í þágu orðabókar. (Þetta hefti).

Jörgen Pind (ritstj.), Friðrik Magnússon og Stefán Briem. 1991. *Íslensk orðtíðnibók.* Reykjavík: Orðabók Háskólans.

Lindén, Krister & Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17: 119–140.

Manning, Christopher & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge Mass., London: MIT Press.

Pantel, Patrick & Dekang Lin. 2002. Discovering Word Senses From Text. Í: *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, bls. 613–619. Edmonton, Kanada.

Pedersen, Bolette Sandford, Sanni Nimb, Jörg Asmussen, Nicolai Hartvig Sörensen, Lars Trap-Jensen & Henrik Lorentzen. 2009. DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. *Language Resources and Evaluation*, 43: 269–299.

Ruiz-Casado, Maria, Enrique Alfonseca & Pablo Castells. 2005. Automatic Extraction of Semantic Relationships for WordNet by means of Pattern Learning from Wikipedia. Í: A. M. R. Munos & E. Métais (útg.). *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, bls. 67–79. Alicante, Spáni. Volume 3513 of Lecture Notes in Computer Science, Heidelberg: Springer.

Sahlgren, Magnus. 2006. *The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces.* Doktorsritgerð. Háskólinn í Stokkhólmi. Sjá heimasíðu M. Sahlgren: http://www.sics.se/~mange/publications.html. (20.10.2011).

Schütze, Hinrich. 1993. Word Space. Í: S. J. Hanson, J. D. Cowan & C. L. Giles (útg.). *Advances in Neural Information Processing Systems*, 5, bls. 895–902. San Mateo, Kaliforníu: Morgan Kaufmann.
Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild. Í: *Samspil tungu og tækni*, bls. 65–71. Reykjavík: Menntamálaráðuneytið.
*Snara.* http://snara.is. (30.06.2011)
Whelpton, Matthew. 2012. From human-oriented dictionaries to computer-oriented lexical resources – trying to pin down words. (Þetta hefti).
*WordNet.* http://www.princeton.edu/wordnet/. (20.10.2011)

## Abstract

This article describes the work on a semantic database for Icelandic language technology. The database is being developed using a monolingual approach with automatic methods for the extraction of semantic information from texts. Both pattern based and statistical methods are used, as well as a hybrid methodology. The database already contains about 134,000 words, primarily nouns, and more than one million relations. The number of relations might change during the last stage of the development which consists of automatically validating the results. This will be done e.g. by using results of one extraction method to support or reject the results of another.

The structure of the database is not based on hierarchies, like for example the Princeton WordNet, but rather on clusters of strongly related words and semantic relations often describing common sense knowledge and associations.

After release, in the beginning of 2012, the database will be freely available.

## Lykilorð

merkingarbrunnur, orðanet, máltækni, merkingarvensl, merkingarupplýsingar

## Keywords

semantic database, wordnet, language technology, semantic relations, semantic information

*Anna B. Nikulásdóttir*
*Háskóli Íslands*
*anna.b.nik@gmk.de*