Erla Hallsteinsdóttir

Íslenskur orðasjóður

1 Inngangur

Í þessari grein mun ég lýsa tilurð og eiginleikum íslensks textagrunns – Íslensks orðasjóðs – sem unnið er að við Háskólann í Leipzig. Íslenskur orðasjóður er íslenskur textagrunnur með innbyggðu orðasafni sem samanstendur af u.þ.b. 400 milljónum orða og orðmynda úr íslensku nútímamáli. Íslenskur orðasjóður var unninn við Háskólann í Leipzig sem hluti af rannsóknarvinnu við verkefnið Leipzig Corpora Collection (http://corpora.informatik.uni-leipzig.de/). Textarnir í textagrunninum eru úr vefsíðusöfnun Landsbókasafns Íslands – Háskólabókasafns haustið 2005 og notendaumhverfið var þróað við tölvunarfræðideild Háskólans í Leipzig, sbr. mynd 1 (hér sem þróunarútgáfa með einum árgangi af *Morgunblaðinu*: http://wortschatz.uni-leipzig.de/ws_ice/index.php):

Textagrunnurinn verður birtur á Veraldarvefnum til frjálsrar notkunar. Í þessari grein mun ég lýsa notkunarmöguleikum textagrunnsins sem orðasafns fyrir almenna notendur og sem rannsóknartækis fyrir rannsóknir bæði í hagnýtum og almennum málvísindum.

Orð og tunga 9 (2007), 83–99.

Mynd 1: Vefsíða Íslensks orðasjóðs.

2 Textagrunnar: „orðasjóðir“

2.1 Orðasjóðir í Leipzig

Deutscher Wortschatz („Þýskur orðasjóður“) er textagrunnur með innbyggðri orðabók (málheild þýsks ritmáls með leitarmöguleikum) á Veraldarvefnum sem gerð var og viðhaldið er af Uwe Quasthoff og samstarfsfólki hans við Háskólann í Leipzig. Þýski orðasjóðurinn inniheldur nú texta m.a. frá Institut für Deutsche Sprache í Mannheim, stórum þýskum bókaforlögum (t.d. dtv, Reclam, Walter de Gruyter), frá ýmsum stærri blöðum og tímaritum (t.d. *Der Spiegel*, *TAZ*, *Süddeutsche Zeitung*, *Die Zeit*), opinberum aðilum (t.d. þýsku ríkisstjórninni, Frauenhofer-stofnununum). Vefsíður þýska orðasjóðsins (www.wortschatz.uni-leipzig.de) eru á þýsku en textagrunna 17 annarra tungumála er að finna á ofangreindu vefsvæði með enskum skýringum. Að auki er hægt að nálgast afmarkaða textagrunna á 15 tungumálum til notkunar í rannsóknum á vefsvæðinu http://corpora.informatik.uni-leipzig.de/download.html. Aðalmarkmiðið með þessum textagrunnum er að veita frjálsan aðgang að stöðluðum, sambærilegum málgögnum og

tölfræðilegum upplýsingum um þessi gögn (sbr. Quasthoff, Richter og Biemann 2006).

Hugmyndin að íslenskum textagrunni kom fram í framhaldi af samstarfi mínu við Uwe Quasthoff um rannsókn á tíðni þýskra orðtaka[1] í textagrunninum Deutscher Wortschatz. Tæknilega hliðin á Íslenskum orðasjóði var prófuð með einum árgangi af *Morgunblaðinu* (1 milljón orða) og næsta skref er að skipta um textasafn og ná tökum á að tengja beygingarform íslensku orðanna við uppflettiorðið. Stefnt er að því að nota til þess beygingarlýsingu fyrir íslensku sem unnin var sem tungutækniverkefni við Orðabók Háskólans.

2.2 Textagrunnur: Vefsíðusafn Landsbókasafns Íslands

Gerður hefur verið samningur við Landsbókasafn Íslands – Háskólabókasafn um að nota safn íslenskra vefsíðna (vefsíður sem enda á.is) sem grunn í Íslenskan orðasjóð. Textagrunnurinn verður byggður upp á sama hátt og Deutscher Wortschatz í Leipzig.

Töluvert magn erlendra texta var í vefsíðusafninu. Það olli þó engum vandræðum við úrvinnslu þess þar sem þróuð hefur verið mjög virk aðferð til að finna og fjarlægja önnur tungumál og „rusl" eins og leifar af forritunarmálum úr textagrunnum (sbr. Quasthoff og Biemann 2006). Eftir að hafa notað þessa aðferð stendur eftir íslenskur textagrunnur með u.þ.b. 400 milljónum orðmynda. Ótvíræður kostur við vefsíðusafnið er að það inniheldur íslenskt nútímamál í víðum skilningi, bæði texta frá opinberum aðilum og einkaaðilum, texta sem fylgja reglum ritmáls og texta sem telja má vera talmál á rituðu formi.

2.3 Notendahópur og notkunargildi „Íslensks orðasjóðs"

Íslenskur orðasjóður hefur tvenns konar notkunarmöguleika:

[1]Rannsóknin var unnin sem hluti af rannsóknastöðuverkefni RANNÍS við Hugvísindadeild HÍ 2001–2004, einnig styrkt af Launasjóði fræðiritahöfunda 2005, vinnan við íslenska textagrunninn er styrkt af DAAD. Mig langar að koma á framfæri þakklæti til þessara aðila fyrir fjárhagslegan stuðning.

(1) Orðasafn Íslensks orðasjóðs er „öðru vísi" orðabók sem inniheldur upplýsingar um heildartíðni og hlutfallslega tíðni orðmynda, notkunardæmi, merkingar- og textaumhverfi, þ.e. hægri og vinstri nágranna í textum og orð með marktæka tíðni í sömu setningum og leitarorðið. Orðasafnið er í íslensku notendaumhverfi sem ætlað er almennum íslenskum notendum.

(2) Textagrunnar eru mikilvægt hjálpartæki í tungumálarannsóknum og tungutækniverkefnum. Íslenskur orðasjóður er einn umfangsmesti textagrunnur á íslensku sem er ætlaður til notkunar í rannsóknum á íslensku nútímamáli, en öflugar rannsóknir eru mikilvæg undirstaða varðveislu og eflingar íslenskrar tungu. Sem dæmi um rannsóknir sem eru í undirbúningi má nefna tíðnirannsóknir á íslenskum orðtökum og rannsóknir á nýyrðum, orðmyndunarmöguleikum og úreltum orðum í íslensku.

2.3.1 Almennir notendur

Notendahópur þýska orðasjóðsins er mjög fjölbreyttur. Allir sem eitthvað hafa með málnotkun (skrifa og þýða texta), þýskukennslu eða tungumálarannsóknir að gera virðast nota hann og hafa gagn af honum, bæði notendur með þýsku sem móðurmál og einnig þeir sem eru að læra þýsku. Miðað við notendahóp þýska orðasjóðsins má ætla að Íslenskur orðasjóður muni einnig nýtast mjög fjölbreyttum íslenskum notendahópi og jafnvel einnig málnotendum sem eru að læra íslensku sem erlent tungumál.

2.3.2 Sérhæfðir notendur

Það er viðurkennd staðreynd að textagrunnar eru nauðsynleg hjálpartæki í tungumálarannsóknum (sbr. Quasthoff, Richter og Biemann 2006). Textasafn Orðabókar Háskólans inniheldur samtals um 52 milljónir lesmálsorða úr fjölbreyttum textum (sbr. upplýsingar á heimasíðu Orðabókar Háskólans, http://www.lexis.hi.is/ts_umsafnid.htm) en vegna höfundarréttar er vefaðgangur[2] að safninu einskorðaður við texta án

[2]Starfsmenn Orðabókarinnar hafa reynst mjög hjálpsamir við aðstoð við og aðstöðu fyrir rannsóknarverkefni, þ.e. aðgangur að öllu textasafninu hefur verið mögu-

höfundarréttar. Við Orðabók Háskólans er einnig verið að vinna að markaðri íslenskri málheild (sbr. Sigrún Helgadóttir 2004).

Markmiðið með Íslenskum orðasjóði er að veita aðgang að málnotkun í íslensku eins og hún er í dag í textum (hugsanlegur möguleiki er að leyfa val á milli textategunda eftir uppruna textanna, t.d. úr *Morgunblaðinu*, til að kynna sér málnotkun í þeim). Notkunargildi Íslensks orðasjóðs felst einkum í orðfræðilegum upplýsingum um notkun orða í textum; þessar upplýsingar eru skýrðar nánar í kafla 3. Gagnagrunnurinn sem geymir textana er þannig byggður upp að ekki er hægt að endurgera texta úr honum; þetta er nauðsynleg ráðstöfun til að tryggja að farið sé eftir lögum um höfundarrétt.

Íslenski textagrunnurinn verður hluti af fjölmála textagrunni á vefsvæði þýska orðasjóðsins sem ætlaður er til notkunar í tungumálarannsóknum. Hugsanlegt er að nota þessa textagrunna meðal annars við (sbr. Quasthoff, Richter og Biemann 2006):

- vinnu að einmála orðabókum,
- leit að svörum við málfræðilegum spurningum,
- tölfræðilega unninn samanburð á mismunandi tungumálum,
- gerð mállíkana, t.d. fyrir talgreiningu,
- rannsóknir á orðum sem haga sér tölfræðilega á líkan hátt,
- val á orðum í tilraunir, t.d. í sálfræðilegum málvísindum.

Þetta er ekki tæmandi listi, möguleikarnir eru margvíslegir, m.a. við rannsóknir á tíðni, orðmyndun, merkingu og merkingarlegu umhverfi orða. Dæmi um önnur áhugaverð rannsókna- og tungutækniverkefni sem byggja á gögnum úr textagrunnum má finna í greinum Richter, Quasthoff, Erla Hallsteinsdóttir og Biemann (2006) og Quasthoff, Richter og Biemann (2006) um notkun textagrunna í tungumálarannsóknum.

Eins og áður var nefnt hefur þýski orðasjóðurinn verið notaður sem grunnur í rannsókn á tíðni þýskra orðtaka. Niðurstöðurnar úr þeirri rannsókn hafa þegar verið nýttar á margvíslegan hátt, m.a. við að velja orðtök í þýsk-íslenskan orðtakagagnagrunn (sbr. Erla Hallsteinsdóttir 2005, 2006b), við að velja þýsk orðtök í grunnorðaforða þýsku sem erlends tungumáls (sbr. Erla Hallsteinsdóttir, Sajankova

legur ef unnt er að vinna rannsóknarvinnuna í húsakynnum Orðabókarinnar.

og Quasthoff 2006) sem og við endurskoðun fræðilegra hugmynda um margræðni orðtaka og þróun aðferðafræði við rannsóknir á orðtökum (sbr. Erla Hallsteinsdóttir í prentun). Íslenski textagrunnurinn mun verða notaður við tíðnirannsóknir á íslenskum orðtökum og niðurstöður þeirra rannsókna munu verða nýttar á sama hátt og niðurstöður þýsku rannsóknarinnar. Þrátt fyrir miklar og góðar orðtakarannsóknir í íslensku eru rannsóknir á orðtökum í textagrunni, t.d. á tíðni orðtaka, á frumstigi og það vantar enn skilgreiningu á þeim orðtökum sem tilheyra grunnorðaforða, þ.e. hvaða orðtök eru æskilegur hluti orðaforða við nám í íslensku sem erlendu máli.

3 Íslenskur orðasjóður

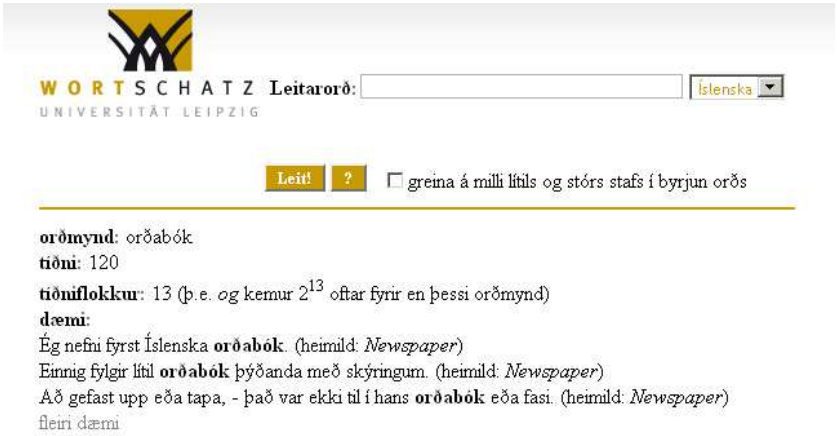
Íslenski *Morgunblaðs*-orðasjóðurinn eins og hann er vistaður á vefsíðu Projekt Deutscher Wortschatz í Leipzig í dag býður m.a. upp á eftirfarandi notkunarmöguleika:

(1) leit að orðum/orðmyndum,
(2) leit að notkunarumhverfi,
(3) leit að notkunardæmum,
(4) leit að samsettum og afleiddum orðum með algildistáknum (* og ?),
(5) listar með tíðni orðmynda.

Hér á eftir fylgir stutt lýsing Íslenska orðasjóðnum með dæmum. Að auki mun ég lýsa þeim möguleikum sem eingöngu eru í þýska orðasjóðnum og hægt væri að yfirfæra á orðasafnið í Íslenskum orðasjóði.

3.1 Leit að orðum

Leitarniðurstöður í Íslenskum orðasjóði sýna tíðni, tíðniflokk (miðað við „og" sem er algengasta orð) og textadæmi með tengli til fleiri dæma, sbr. niðurstöður leitar að orðinu *orðabók* í mynd 2:



Mynd 2: Upplýsingar um tíðni og tengill við fleiri dæmi.

Þar sem ekki er búið að tengja saman beygingarmyndir orða eru eingöngu sýndar upplýsingar um orðið í eintölu í nefnifalli, þolfalli og þágufalli, þ.e. upplýsingar um beygingarmyndina *orðabók*. Leita verður sérstaklega að öðrum myndum orðsins. Einnig er sýnt merkingarlegt umhverfi leitarorðsins, þ.e. gefin eru upp orð sem hafa marktæka tíðni sem nágrannar leitarorðsins í textum. Þessar upplýsingar verða útskýrðar nánar í eftirfarandi köflum.

3.2 Notkunarumhverfi

Með notkunarumhverfi er ekki eingöngu átt við fastar orðastæður (e. collocations) í hefðbundnum skilningi heldur einnig þau orð sem hafa háa tíðni sem nágrannar leitarorðsins í textum. Við þau orð sem hafa marktæka tíðni sem nágrannar er sýnd heildartíðni í gagnagrunninum. Einnig er greint á milli orða sem koma fyrir sem hægri og vinstri nágrannar leitarorðsins og við þau orð er gefin upp sú tíðni sem þau hafa með leitarorðinu, sbr. mynd 3:

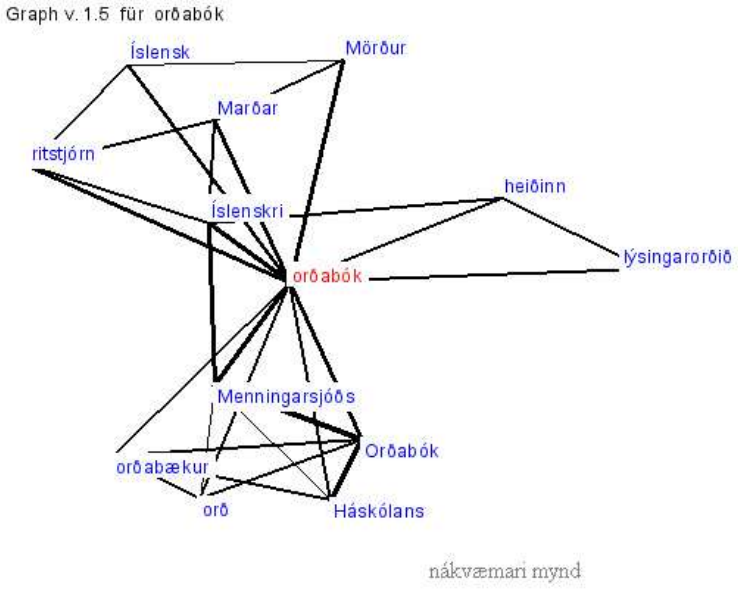
orð sem koma oft fyrir sem nágrannar orðabók:
Menningarsjóðs (85), Íslenskri (64), Íslensk (42), Orðabók (23), Mörður (21), ritstjórn (20), Marðar (20), Í (17), Orðastað (17), lýsingarorðið (15), heiðinn (15), orðabækur (14), orð (14), Háskólans (14), þreyja (13), Árnasonar (13), stórfiskaleikur (13), prentútgáfna (13), lýðveldistímans (13), klyfberi (13), Bókaútgáfu (13), ÍSLENSK (12), uppgjöf (12), delicious (12), útgáfudegi (11), merking (11), lsquo (11), gefast (11), eða (11), Freysteins (11), orðsins (10), orðið (10), orðinu (10), merkir (10), hugum (10), færeyskt (10), fletta (10), ekki (10), dægrastytting (10), Grunnavík (10), Örlygs (9), Íslenska (9), syndrome (9), glöggva (9), forsölu (9), bók (9), Árna (8), viðhorfa (8), slangur (8), samkvæmt (8), samanlögðu (8), ríkjandi (8), nýrri (8), metsölubók (8), heimspekideild (8), fornit (8), endurbætt (8), Blöndals (8), nefni (7), merkingar (7), keyptum (7), er (7), alist (7), Starfaði (7), Orðið (7), Böðvarssonar (7), íslenskri (6), Ö (6), skýringum (6), selst (6), samantekt (6), lektor (6), hinni (6), hin (6), gefin (6), eintök (6), dósent (6), Færeyingar (6), íslensku (5), íslenska (5)

orð með háa tíðni sem vinstri nágrannar orðabók:
Íslenskri (64), Íslensk (34), ÍSLENSK (12), íslenskri (4), Úr (4), Íslenska (4), íslenska (3), samkvæmt (3)

orð með háa tíðni sem hægri nágrannar orðabók:
Menningarsjóðs (50), Freysteins (11), Blöndals (8), ríkjandi (5), Háskólans (5)

Mynd 3: Dæmi um notkunarumhverfi orða: nágrannar í textum.

Upplýsingar um merkingarumhverfi orða eru sýndar á grafísku formi með n.k. merkingarneti, sbr. mynd 4:



Mynd 4: Merkingarnet orðmyndarinnar *orðabók*.

Upplýsingar um orðastæður og textanágranna eru mikilvægar í málnotkun. Þær sýna hvaða orð er hægt að nota saman (sbr. muninn á að

bursta tennurnar í íslensku og „hreinsa tennurnar" (*die Zähne putzen*) í þýsku). Þessar upplýsingar mynda grunn fyrir tungumálarannsóknir, t.d. merkingarfræði og setningafræði og einnig við orðabókagerð.

3.3 Notkunardæmi

Við hvert orð eru sýndar tvær setningar úr gagnagrunninum sem sýna notkun orðsins. Að auki er tengill við síðu með fleiri notkunardæmum, sbr. eftirfarandi notkunardæmi fyrir leitarorðið *orðabók* (öll dæmi úr *Morgunblaðinu*):

Dæmi:

- Í samantekt Minjasafnsins á Akureyri sem byggir m.a. á bók Hallgerðar Gísladóttur, Íslensk matarhefð, kemur fram að elstu rituðu heimildir um hátíðarbrauð Íslendinga, laufabrauðið, séu í **orðabók** Jóns Ólafssonar frá Grunnavík frá árinu 1736.
- Það er greinilegt að blaðamaður hefur ekki ómakað sig við að fletta upp í nýútkominni **orðabók** til að glöggva sig á málinu, því það er ljóst að hefði hann gert það, hefðu hinar gnarrísku fullyrðingar aldrei komist á prent.
- Í nýrri og prýðilegri **orðabók** Eddu - miðlunar er merking orðsins „skipuleg samtök til að berjast fyrir ákveðinni stefnu og markmiðum í stjórnmálum".
- Vol og væl var ekki til í hennar **orðabók** og ekki minnist ég þess að hafa nokkurn tíma hitt hana í slæmu skapi.
- Sjálfsvorkunn var hreinlega ekki til í hans **orðabók**.
- Í Íslenskri **orðabók** stendur um lýsingarorðið heiðinn: 1) sem er heiðingi, ókristinn; guðlaus; heiðinn siður Ásatrú; heiðinna manna heilsa fornmannaheilsa, góð heilsa. 2) ófermdur, illa upplýstur um trúmál. 3) sem vantar á: heiðinn klyfberi gjarðalaus klyfberi; verlaus (um sæng): sofa í heiðnu rúmi; bryddingalaus; auður, óskrifaður: heiðin blaðsíða; sviplaus, eyðilegur: þetta er svo heiðið.
- Ný íslensk **orðabók** var kærkomin sending inn í það ógnargímald, en betur má ef duga skal.
- Því er vitnað í þessa bók nú, að Íslensk **orðabók** hefur nú verið gefin út, mikið aukin og endurbætt.

- En eigum við að fljóta sofandi að feigðarósi og bíða þess að lýsingarorðið „delicious“ rati inn í íslenska **orðabók** og sé þar sjálfsögð fletta og eðlilegur hluti talaðs máls á Íslandi?
- Enska lýsingarorðið „delicious“ er hinsvegar ekki að finna í hinni nýju íslensku **orðabók**, þótt það hafi ratað inn í auglýsingu kartöflubænda.
- Margar slettur hafa skamma viðdvöl á vörum fólks, – sem betur fer og raunar þurfa kannski ýmsir á því að halda að slettunum sé hægt að fletta upp í **orðabók**.
- HIN nýja Íslenzka **orðabók** hefur komið af stað miklum umræðum, ekki sízt um það hvort ýmis orð (t.d. sjitt eins og frægt er orðið) megi vera í bókinni eða hvort úthýsa beri „röngu“ máli úr jafnvirðulegri heimild um íslenzkt mál.
- Þau tímamót sem mörkuð eru með nýrri **orðabók** eru hvatning til þess að almenningur taki afstöðu til þess málfars sem nú tíðkast og leiði jafnframt hugann að því það hvernig hann telur ákjósanlegt að málið þróist
- Þegar orðabókinni var fylgt úr hlaði í síðustu viku kom það fram í máli Marðar Árnasonar að vonir stæðu til „að hér eftir liðu ekki nema 5–10 ár á milli prentútgáfna af Íslenskri **orðabók**“.

Greining á þessum dæmum leiðir m.a. í ljós að orðið *orðabók* er oft hluti af sérheitinu *Íslensk orðabók*, það kemur fyrir sem hluti af orðasambandinu „eitthvað er ekki til í hans/hennar orðabók“ og í merkingunni 'uppsláttarrit um orðaforða'.

3.4 Leit að samsettum og afleiddum orðum

Með því að nota algildistáknin stjörnu eða spurningarmerki (* eða ?) er hægt að leita að orðum sem innihalda ákveðna bókstafi eða orðhluta. Leitarorðið *orðab** gefur t.d. eftirfarandi niðurstöður (tíðni í hornklofa):

orðabanka [4]	orðabókarforrit [1]	orðabókarsmíðinni [1]
orðabankann [1]	orðabókargerð [2]	orðabókarstjóri [3]
orðabankanum [5]	orðabókargerðina [2]	orðabókarstörf [1]
orðabankinn [1]	orðabókargerðinni [1]	orðabókarverk [1]
orðabilum [8]	orðabókarhefð [1]	orðabókarverkefnið [1]
orðablaðra [1]	orðabókarhöfunda [3]	orðabókarvinnslunni [1]
orðablaðran [1]	orðabókarhöfundar [1]	orðabókasmíð [1]
orðabrunnurinn [1]	orðabókarhöfundarnir [1]	orðabókaverkefni [1]
orðabækur [29]	orðabókarhöfundi [1]	orðabókaútgáfa [1]
orðabækurnar [2]	orðabókarinnar [16]	orðabókaútgáfu [2]
orðabók [120]	orðabókarlýsingu [2]	orðabókin [34]
orðabóka [9]	orðabókarmaður [2]	orðabókina [11]
orðabókagerð [2]	orðabókarnotenda [1]	orðabókinni [27]
orðabókagerðina [1]	orðabókarritstjóri [2]	orðabókum [20]
orðabókanotenda [1]	orðabókarskráin [1]	orðabókunum [1]
orðabókar [21]	orðabókarsmiðsins [1]	orðabólgu [1]
orðabókarbreytingar [1]	orðabókarsmíð [1]	

Tafla 1: Orð sem innihalda *orðab.*

Eins og sést í töflu 1 koma öll beygingarform fram sem sérstök orð, án tengsla sín á milli, sbr.:

orðabækur [29], orðabækurnar [2], orðabók [120], orðabóka [9], orðabókar [21], orðabókarinnar [16], orðabókin [34], orðabókina [11], orðabókinni [27], orðabókum [20]. Til þess að öll beygingarform birtist í leitarniðurstöðum verður að samtengja beygingarform og grunnform orða eins og gert hefur verið í þýska orðasjóðnum.

3.5 Orðalistar

3.5.1 Listar með tíðni orðmynda

Upplýsingar um tíðni orða og orðasambanda eru mikilvægar, bæði í hagnýtum tungumálarannsóknum og í málvísindum. Á einfaldan hátt er hægt að gera lista með t.d. algengustu eða sjaldgæfustu orðum í gagnagrunninum. Í töflu 2 eru sýnd dæmi fyrir þýsku og ensku:

þýsk orð		**ensk orð**	
1:	*der*	*of*	:1
2:	*die*	*to*	:2
3:	*und*	*and*	:3
4:	*in*	*a*	:4
5:	*den*	*in*	:5
6:	*von*	*for*	:6
7:	*zu*	*is*	:7
8:	*das*	*the*	:8
9:	*mit*	*that*	:9
10:	*sich*	*on*	:10
	Niðurhal algengustu orða	*Niðurhal algengustu orða*	
	++ 100 ++ 1000 ++ 10000 ++	*++ 100 ++ 1000 ++ 10000 ++*	

Tafla 2: Algengustu orð í þýsku og ensku.

3.5.2 Nýyrði og úrelt orð

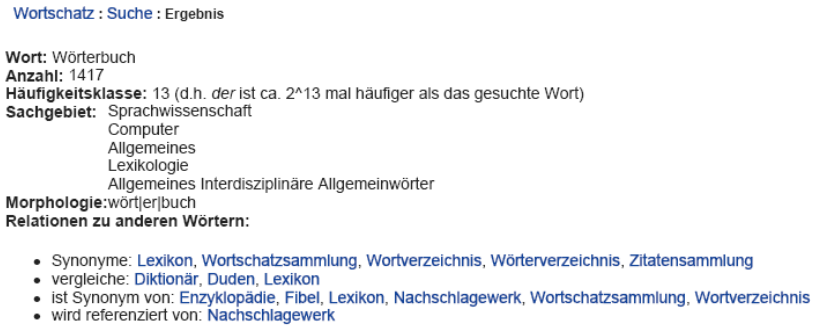
Með því að bera saman orð og beygingarmyndir úr *Beygingarlýsingu íslensks nútímamáls* og orð og beygingarmyndir sem koma fyrir í textagrunninum fást áhugaverðar upplýsingar um nýyrði, orðmyndunarmöguleika og úrelt orð í íslensku. Gera má ráð fyrir að þau orð sem ekki eru í beygingarlýsingunni séu nýyrði eða slangur og að þau orð sem ekki koma fyrir í textagrunninum séu orðin úrelt eða séu beygingarform sem af einhverjum ástæðum eru ekki notuð. Gögn úr þannig samanburði mynda grunn fyrir rannsóknir á þróun og stöðu orðaforðans.

3.6 Deutscher Wortschatz

Eins og eftirfarandi dæmi sýna hefur verið bætt við töluverðu magni af upplýsingum í þýska orðasjóðnum, m.a. um beygingu, merkingu og merkingartengsl (þó ekki merkingarlýsingu), orðmyndun, orðatengsl o.fl. Þessar upplýsingar, sem byggja að hluta til á lokaverkefnum nemenda við Háskólann í Leipzig, eru unnar bæði sjálfvirkt, hálfsjálfvirkt og handvirkt. Notuð er sjálfvirk tenglasetning til að samtengja upplýsingarnar í orðasjóðnum.

Í mynd 5 hafa auk tíðni (Anzahl) og tíðniflokks (Häufigkeitsklasse) verið tilgreindir fagflokkar sem orðið *Wörterbuch* tilheyrir (Sachge-

biet); sýnd er orðhlutagreining orðsins (Morphologie) og gefin eru upp
merkingarleg tengsl (samheiti, svipuð merking o.s.frv.) við önnur orð (Relationen zu anderen Wörtern):



Wortschatz : Suche : Ergebnis

Wort: Wörterbuch
Anzahl: 1417
Häufigkeitsklasse: 13 (d.h. *der* ist ca. 2^13 mal häufiger als das gesuchte Wort)
Sachgebiet: Sprachwissenschaft
Computer
Allgemeines
Lexikologie
Allgemeines Interdisziplinäre Allgemeinwörter
Morphologie:wört|er|buch
Relationen zu anderen Wörtern:

- Synonyme: Lexikon, Wortschatzsammlung, Wortverzeichnis, Wörterverzeichnis, Zitatensammlung
- vergleiche: Diktionär, Duden, Lexikon
- ist Synonym von: Enzyklopädie, Fibel, Lexikon, Nachschlagewerk, Wortschatzsammlung, Wortverzeichnis
- wird referenziert von: Nachschlagewerk

Mynd 5: Upplýsingar við orðið *Wörterbuch* í þýska orðasjóðnum.

Í mynd 6 (Links zu anderen Wörtern) eru sýnd dæmi um fleiri tegundir merkingartengsla og einnig eru gefin upp samsett orð, orðastæður og orðasambönd sem orðið er hluti af og sem hægt er að slá upp sem sjálfstæðum flettum í orðasjóðnum. Að auki koma fram beygingarform og skammstafanir orðsins.

Links zu anderen Wörtern:

- falls positiv bewertet Originalwörterbuch
- Grundform: Wörterbuch
- ist ein(e) Buch, Nachschlagewerk, Wortsammlung
- Teilwort von: im Wörterbuch nachschlagen, Wörterbuch zusammenstellen, kurzgefaßtes Wörterbuch, fremdsprachliches Wörterbuch, ein einsprachiges Wörterbuch, rückläufig sortiertes Wörterbuch, automatisches Wörterbuch, ein gutes Wörterbuch, rückläufiges Wörterbuch, computerisiertes Wörterbuch, in einem Wörterbuch nachschlagen, ein wandelndes Wörterbuch, übersetzende Wörterbuch, ein Wörterbuch kürzen
- Form(en): Wörterbuch, Wörterbücher, Wörterbüchern, Wörterbuchs, Wörterbuches, Wörterbuche
- Abkürzung: WB, Wtb., Wb.

Dornseiff-Bedeutungsgruppen:

- 11.30 Kenntnis: Bibliografie, Buch, Enzyklopädie, Handbuch, Lexikon, Pflichtlektüre, Sekundärliteratur, Standardwerk, Vademekum, Vokabular, Wörterbuch
- 12.16 Bezeichnung, Wort: Duden, Fremdwörterbuch, Glossar, Grundwortschatz, Lexikon, Phrasenkatalog, Sprachschatz, Verzeichnis, Wortschatz, Wörterbuch, Zitatenschatz
- 12.43 Erklärung: Enzyklopädie, Lexikon, Wörterbuch
- 12.55 Schriftliche Überlieferung: Auflistung, Bibliografie, Enzyklopädie, Index, Katalog, Kompendium, Konkordanz, Lexikon, Liste, Register, Tabelle, Verzeichnis, Werkverzeichnis, Wörterbuch

Mynd 6: Upplýsingar við orðið *Wörterbuch* í þýska orðasjóðnum.

Sýnd eru hugtök sem orðið heyrir undir í *Dornseiff*-hugtakaorðabókinni (Dornseiff-Bedeutungsgruppen, sbr. Dornseiff 2004) sem er ein stærsta hugtakaorðabók í þýsku, en Uwe Quasthoff ritstýrði nýjustu útgáfu hennar.

4 Íslensk-þýsk orðabók

Saga þýsk-íslenskra orðabóka hefur verið hálfgerð sorgarsaga (sbr. Erla Hallsteinsdóttir 2004). Fyrir utan skólaorðabók Steinars Matthíassonar (Steinar Matthíasson 2004) hafa ekki verið gefnar út neinar orðabækur með þýsku fyrir íslenska notendur á síðustu árum. Ein af hugmyndum um notkun á Íslenskum orðasjóði í Leipzig er að byrja á grunni fyrir þýsk-íslenska netorðabók á svipuðu formi og þýsk-enska orðabókin sem þegar hefur verið gerð í Leipzig, sjá eftirfarandi dæmi í mynd 7:

Abfrageergebnis für Ihre Anfrage nach »Wörterbuch«

18 Treffer aus dem **deutsch->englisch** Lexikon:		**#Belege**
Wörterbuch (652)	dictionary (4641)	7
Wörterbuch (652)	vocabulary (1007)	2
Wörterbuch (652)	wordbook	1
Wörterbuch (652)	glossary (689)	1
ein gutes (8783) Wörterbuch (652)	a good dictionary (4641)	1
ein Wörterbuch (652) kürzen (8322)	abridge (18) a dictionary (4641)	1
übersetzende (9) Wörterbuch (652)	translating (1086) dictionary (4641)	1
kurzgefaßtes (2) Wörterbuch (652)	concise (839) dictionary (4641)	1
rückläufiges (38) Wörterbuch (652)	reverse (6828) dictionary (4641)	1
automatisches (152) Wörterbuch (652)	automatic (16704) dictionary (4641)	3
ein wandelndes (59) Wörterbuch (652)	a walking (4486) dictionary (4641)	1
Wörterbuch (652) zusammenstellen (592)	compile (2688) a dictionary (4641)	1
im Wörterbuch (652) nachschlagen (118)	consult (1853) the dictionary (4641)	1
computerisiertes (5) Wörterbuch (652)	computerized (4588) dictionary (4641)	1
ein einsprachiges Wörterbuch (652)	a monolingual (7) dictionary (4641)	1
fremdsprachliches (2) Wörterbuch (652)	foreign-language (218) dictionary (4641)	1
rückläufig (1711) sortiertes (24) Wörterbuch (652)	backward-sorted dictionary (4641)	1
in einem Wörterbuch (652) nachschlagen (118)	consult (1853) a dictionary (4641)	1

Mynd 7: *Wörterbuch* í þýsk-ensku orðabókinni.

Í því sambandi munu möguleikar á sjálfvirkri málgreiningu verða skoðaðir (sbr. t.d. Cysouw, Biemann og Ongyerth 2006 og Rapp 1994) með það í huga að nýta þá möguleika sem til eru á notkun textagrunnsins í sjálfvirkum þýðingum eða orðabókagerð, t.d. með tölfræðilegri greiningu orða, orðastæðna og setninga í textum eða með samanburði á frumtextum og þýðingum þeirra.

5 Samantekt

Íslenskur orðasjóður er verkefni sem unnið er við Háskólann í Leipzig. Orðasjóðurinn, sem byggir á vefsíðusafni Landsbókasafns Íslands – Háskólabókasafns og nýtir tækniþekkingu úr verkefninu Deutscher

Wortschatz, veitir almennum notendum aðgang að upplýsingum um málnotkun í íslensku nútímamáli og fræðimönnum textagrunn sem hægt er að nýta á margvíslegan hátt í rannsóknum á tungumálum. Orðasjóðurinn mun nýtast bæði í hagnýtum rannsóknum eins og orðabókagerð, gerð kennsluefnis og tungumálakennslu og í fræðilegum rannsóknum, t.d. við þróun kenninga og aðferðafræði í tungumálarannsóknum.

Ritaskrá

Beygingarlýsing íslensks nútímamáls. Vefslóð: *http://www.lexis.hi.is/beygingarlysing*.

Cysouw, Michael, Biemann, Christian og Ongyerth, Matthias. 2006. Using strong's numbers in the bible to test an automatic alignment of parallel texts. Í: Michael Cysouw og Bernhard Wälchli (útg.): Parallel Texts: Using translational equivalents in linguistic typology. *Special issue of Sprachtypologie und Universalienforschung (STUF)*, bls. 66–79.

Dornseiff, Franz. 2004. *Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearb. und mit einem alphabetischen Zugriffsreg. vers. Aufl. von Uwe Quasthoff. Berlin, New York: de Gruyter.

Erla Hallsteinsdóttir. 2004. En kort oversigt over islandsk ↔ tysk leksikografi. Í: *LexicoNordica* (11), bls. 51–65.

Erla Hallsteinsdóttir. 2005. Vom Wörterbuch zum Text zum Lexikon. Í: Ulla Fix, Gotthard Lerchner, Marianne Schröder og Hans Wellmann (útg.): *Zwischen Lexikon und Text – lexikalische, stilistische und textlinguistische Aspekte*, bls. 325–337. Leipzig: Verlag der Sächsischen Akademie der Wissenschaften zu Leipzig.

Erla Hallsteinsdóttir. 2006a. Phraseographie. Í: *HERMES Journal of Language and Communication Studies* (36), bls. 91–128.

Erla Hallsteinsdóttir. 2006b. Konzeption und Erstellung einer computergestützten zweisprachigen Phraseologiesammlung Isländisch – Deutsch. Í: Annelies Häcki Buhofer og Harald Burger (útg.): *Phraseology in Motion*. Proceedings zu Europhras Basel 2004, bls. 211–222. Baltmannsweiler: Schneider Verlag.

Erla Hallsteinsdóttir. Í prentun. Wörtliche, freie und phraseologische Bedeutung. Eine korpusbasierte Untersuchung des Vorkommens von freien und phraseologischen Lesarten bei deutschen Idiomen.

Í: Erika KrûiŽnik (útg.): *Phraseologie in der Sprachwissenschaft und anderen Disziplinen*. Akten der Europhras-Tagung in Strunjan/Slowenien, 19.–22. September 2005.

Erla Hallsteinsdóttir, Uwe Quasthoff og Monika Sajankova. 2006. Vorschlag eines phraseologischen Optimums für Deutsch als Fremdsprache auf der Basis von Frequenzuntersuchungen und Geläufigkeitsbestimmungen. Í: *Linguistik online* 27, 2/06: Neue theoretische und methodische Ansätze in der Phraseologieforschung, bls. 117–136.
(Vefslóð: *www.linguistik-online.de/27_06/hallsteinsdottir_et_al.pdf*.)

Quasthoff, Uwe og Christian Biemann. 2006. Measuring Monolinguality. Í: *Proceedings of the LREC-06 workshop on Quality assurance and quality measurement for language and speech resources*, Genoa, Italy.

Quasthoff, Uwe og Matthias Richter. 2005. Projekt Deutscher Wortschatz. Í: *Babylonia* 3/2005.
Vefslóð: *http://www.babylonia-ti.ch/BABY305/quaride.htm*.

Quasthoff, Uwe, Matthias Richter og Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. Í: *Proceedings of LREC-06*, Genoa, Italy.

Rapp, Reinhard. 1994. Die maschinelle Generierung von Wörterbüchern aus zweisprachigen Texten. Í: Susanne Beckmann og Sabine Frilling (útg.): *Satz – Text – Diskurs*. Akten des 27. Linguistischen Kolloquiums, Münster, 1993, Bd. I, bls. 203–209. Tübingen: Niemeyer.

Richter, Matthias, Uwe Quasthoff, Erla Hallsteinsdóttir og Christian Biemann. 2006. Exploiting the Leipzig Corpora Collection. Í: *Proceedings of IS-LTC'06*, Ljubljana, Slovenia.

Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild. Í: *Tunga og tækni*, bls. 67–71. (Vefslóð: www.tungutaekni.is/news/sigrun2.pdf.)

Steinar Matthíasson. 2004. *Þýsk-íslensk*, *íslensk-þýsk orðabók*. Aukin og endurbætt útg. Reykjavík: Iðnú.

Abstract

Corpora are important linguistic resources. In this paper I describe the details of an Icelandic corpus, that is a part of a collection of corpora in 17 different languages which can be accessed online from *http://corpora.informatik.uni-leipzig.de*, and, with Icelandic instructions, at

http://wortschatz.uni-leipzig.de/ws_ice/index.php. I discuss the possible usage of the corpus as a corpus based dictionary for non-linguistic users, and as a research tool for linguistic purposes in both applied and theoretical linguistics.

Keywords:

corpora, dictionary, linguistics, Icelandic

Lykilorð:

textagrunnur, orðabók, málvísindi, íslenska

Erla Hallsteinsdóttir
Vægtens Kvarter 336
DK-5220 Odense
erlahall@yahoo.dk