# Orð og tunga

14

*Orð og tunga* birtir fræðigreinar um mál og málfræði með áherslu á greinar sem snerta orðaforðann, þ.á m. nöfn og íðorð, og greinar um orðabókafræði og orðabókagerð. Í hverju hefti eru annars vegar greinar um fyrirfram ákveðið þema og hins vegar aðrar greinar á efnissviði tímaritsins almennt. Auk þess eru birtir ritdómar, bókafregnir og frásagnir af rannsóknarverkefnum og ráðstefnum. Greinar skulu að jafnaði ritaðar á íslensku með útdrætti á ensku en einnig eru birtar greinar á norðurlandamálum og ensku með íslenskum útdrætti. Allar greinar í tímaritinu eru ritrýndar þannig að auk ritstjóra lesa a.m.k. tveir ónafngreindir sérfræðingar hverja grein. Frekari upplýsingar um ritstjórnarstefnu, skil og frágang greina má finna á vefsíðu tímaritsins: http://www.arnastofnun.is/page/arnastofnun_timarit_ot.

# Orð og tunga
# 14

Ritstjóri
Ásta Svavarsdóttir

Stofnun Árna Magnússonar í íslenskum fræðum
Reykjavík 2012

ISSN 1022-4610

Umbrot: Bessi Aðalsteinsson.
Hönnun kápu: Björg Vilhjálmsdóttir.
Prentun og bókband: Prentmet ehf.

# Efnisyfirlit

# Formáli ritstjóra

Með þessu hefti af Orði og tungu verða ritstjóraskipti á tímaritinu. Guðrún Kvaran, sem hefur ritstýrt því allt frá árinu 1997, lætur nú af því starfi og Ásta Svavarsdóttir tekur við sem ritstjóri. Hún sat áður í ritnefnd, sem starfaði um nokkurra ára skeið við hlið ritstjóra (2005–2010), en frá og með síðasta hefti var ákveðið að leggja ritnefndina af og að ritstjórinn tæki einn ábyrgð á útgáfu tímaritsins. Með þessu hefti gengur aftur á móti öflug ráðgefandi ritnefnd til liðs við tímaritið en henni er fyrst og fremst ætlað það hlutverk að vera bakhjarl þess og tengiliður við fræðasamfélagið. Einkum var leitað til fræðimanna utan stofnunarinnar, bæði innan lands og utan, og ritnefndina skipa nú sautján einstaklingar við fjórtán rannsóknarstofnanir og háskóla í átta löndum.

Allt frá árinu 2005, þegar gerðar voru gagngerar breytingar á útgáfu og búningi tímaritsins, hefur það verið fastur liður í útgáfu þess að efna til málþings um tiltekið efni sem síðan hefur orðið þema í næsta hefti á eftir. Á síðasta ári var 100 ára afmæli Háskóla Íslands og af því tilefni var boðað til fjölda málþinga á ýmsum sviðum auk annarra viðburða. Hugvísindastofnun efndi til tvöfalds Hugvísindaþings í marsmánuði, um svipað leyti og efnt hefur verið til málþings á vegum Orðs og tungu, og var brugðið á það ráð að bjóða upp á málstofu innan þingsins að þessu sinni. Hún var haldin í samvinnu við Máltæknisetur og var sameiginleg yfirskrift hennar *Stefnumót: Á mörkum málfræði og tölvutækni*. Síðari hluti málstofunnar var helgaður fyrirhuguðu þema þessa heftis og nefndist *Net til að fanga orð. Greining og lýsing á merkingu og merkingarvenslum*. Þar voru svonefnd **orðanet** í brennidepli, en um þessar mundir er unnið að tveimur íslenskum verkefnum á því sviði, *Íslensku orðaneti* og *Íslenskum merkingarbrunni*. Þrír fyrirlestrar voru haldnir um þetta efni. Matthew Whelpton, dósent í ensku við Háskóla Íslands, hélt inngangsfyrirlestur þar sem hann fjallaði um orðanet sem gerð hafa verið fyrir ýmis tungumál síðastliðna áratugi og gerði grein fyrir einkennum þeirra, hvað þau ættu sameiginlegt og hvað skildi þau að. Anna Björk Nikulásdóttir,

doktorsnemi við Háskóla Íslands, sagði frá gerð *Íslensks merkingarbrunns*, þeim vélrænu aðferðum sem þar er beitt og markmiðum verkefnisins en merkingarbrunnurinn er einkum ætlaður til nota við ýmis máltæknileg úrlausnarefni, þ.e.a.s. hann er ætlaður vélum fremur en fólki. Loks fjallaði Jón Hilmar Jónsson, rannsóknarprófessor við Stofnun Árna Magnússonar í íslenskum fræðum, um *Íslenskt orðanet* sem hann og Þórdís Úlfarsdóttir hafa unnið að um árabil. Sjónarhornið í því verkefni er fyrst og fremst orðabókarfræðilegt, aðferðir um margt ólíkar þeim sem viðhafðar eru í verkefni Önnu og megintilgangurinn annar. Greiningin fer í miklu ríkari mæli fram í höndunum, ef svo má segja, og markmiðið er ekki síst að birta lifandi notendum ýmiss konar merkingarleg og formleg vensl orða og orðasambanda. Málstofunni lauk með pallborðsumræðum þar sem fyrirlesarar brugðust við erindum hver annars og áheyrendum gafst kostur á að bera fram fyrirspurnir og athugasemdir. Á grundvelli erinda sinna og þeirra umræðna sem á eftir fylgdu hafa fyrirlesararnir nú samið greinar sem eru uppistaðan í þessu hefti. Í sameiningu gefa þær áhugaverða innsýn í það rannsóknar- og þróunarstarf sem nú fer fram við greiningu á merkingu og merkingarvenslum íslenskra orða og orðasambanda og með grein Matthew Whelpton eru íslensku verkefnin sett í samhengi við það sem verið er að gera og gert hefur verið annars staðar.

Auk þemagreinanna þriggja eru í heftinu þrír ritdómar um nýleg verk á sviði tímaritsins. Anna Helga Hannesdóttir, dósent við Gautaborgarháskóla, fjallar um nýja útgáfu á orðabók frá 17. öld, *Specimen Lexici Runici* eftir séra Magnús Ólafsson í Laufási sem kom út í ritröðinni *Orðfræðirit fyrri alda*. Þá fjallar Ásgrímur Angantýsson, lektor við Háskólann á Akureyri, um *Handbók í íslensku* sem kom út á síðasta ári og loks fjallar Veturliði G. Óskarsson, dósent í Uppsölum, um ritröðina *Moderna importord i språka i Norden* þar sem gerð er grein fyrir niðurstöðum samnefndrar rannsóknar á nýlegum aðkomuorðum í norðurlandamálum en í dómnum horfir hann einkum til þeirra bóka og bókarkafla sem fjalla um íslensku.

Að vanda eru svo í heftinu bókafregnir þar sem sagt er í stuttu máli frá nokkrum nýjum og nýlegum verkum sem vakið gætu áhuga lesenda, þ. á m. einni orðabók, ÍSLEX *orðabókinni*, sem er sérstaklega samin til vefbirtingar og því ekki bók í hefðbundnum skilningi. Loks eru sagðar fréttir af nýliðnum og væntanlegum ráðstefnum sem tengjast fræðasviði tímaritsins.

Matthew Whelpton

# From human-oriented dictionaries to computer-oriented lexical resources – trying to pin down words

## 1 Introduction

Dictionaries are one of the most familiar linguistic resources to which an ordinary native speaker of a (standardised) language is likely to have access; indeed the process of dictionary creation has served a crucial role historically in the standardisation of European languages and represents an important activity in the creation and maintainance of standards today. In the age of information technology, the successors of the paper dictionary have continued to exert great influence, playing a central role in the field of natural language processing (NLP), supplying text and speech processors with essential information on word form, category, and meaning in activities as diverse as grammatical parsing, information extraction, and machine translation.

Although some of these NLP resources are essentially electronic versions of paper dictionaries, the demands of computer applications are in many ways much greater than those of human users: computers require that information be presented in a manageable format for algorithmic manipulation (e.g. in a relational database where each piece of information is classified and linked explicitly to others) and that the information itself be systematic and absolutely explicit. The

Orð og tunga 14 (2012), 1–18. 

human user of a dictionary brings to the dictionary a host of implicit knowledge and cognitive skills that aid in dictionary use, in particular a range of assumed world and cultural knowledge and the common sense ability to deploy that knowledge appropriately in interpreting the dictionary entry. A computer on the other hand comes to the electronic "dictionary" knowing nothing at all in advance and has only the "sense" that is represented by the processing algorithms available to it. To be effective, a computational lexical resource must therefore represent the relevant information in a fully explicit and systematic way, encoding information that to human users would seem obvious and unnecessary.

This paper focuses on the meaning of words (lexical semantics) and some important computational resources that have been developed to make lexical semantic information available to computers for a variety of NLP tasks. It is intended as a survey article, describing the properties of three major lexical semantic resources (WordNet, DanNet and SALDO), which provide a frame of reference for current work on two Icelandic projects, reported in this volume. Anna Björk Nikulásdóttir reports on a project developing semi-automatic means for extracting information on lexical semantic relations from text corpora (*Íslenskur merkingarbrunnur*, cf Nikulásdóttir & Whelpton 2009, 2010a, 2010b); Jón Hilmar Jónsson reports on a project which is manually developing a network of lexical sense relations (*Íslenskt orðanet*, cf Jónsson 2008, 2009a, 2009b, 2009c; Úlfarsdóttir 2006).

Section 2 introduces one of the oldest and most influential lexical semantic resources, the Princeton WordNet, and reviews some of the central lexical semantic relations around which the resource is structured: synonymy, hyponymy, meronymy, antonymy, and troponymy. Section 3 introduces DanNet, a lexical semantic resource for Danish, conforming to the international standards of wordnet development; a number of challenges faced by DanNet are reviewed, in particular the challenge of converting traditional dictionary information into a computer-tractable form and the challenge of addressing deficiencies in the relation set of the original Princeton WordNet. Section 4 introduces SALDO, a morphological and lexical semantic database for Swedish, organised on radically different lines to the wordnets, as it attempts to model the degree of centrality of lexicalised concepts in Swedish rather than encoding specific lexical semantic relations between them. Section 5 concludes this survey and points on to the papers introducing the two Icelandic resources.

# 2 Wordnet

## 2.1 Background

The Princeton WordNet[1] (Miller 1995, Fellbaum 1998) is a lexical database of English constructed to represent word sense relations. It was developed under the direction of the psychologist, George Miller, and its original aims were explicitly psycholinguistic in nature. As Miller (1998a: xv) explains, the original WordNet project included two psycholinguistic hypotheses: (i) the separability hypothesis "that the lexical component of language can be isolated and studied in its own right", i.e. that the mental lexicon has a distinct organisation and identity from the combinatorial systems of grammar and the expressive system of phonology; (ii) the patterning hypothesis "that people could not master and have readily available all the lexical knowledge needed to use a natural language unless they could take advantage of systematic patterns and relations among the meanings that words can be used to express". The WordNet project was always, however, a project in computational psycholinguistics and another important hypothesis is related to the issue of computational tractability and scalability: the comprehensiveness hypothesis "that computational linguistics, if it were ever to process natural languages as people do, would need to have available a store of lexical knowledge as extensive as people have".

The challenge was to decide how a comprehensive lexical semantic database for computation might be structured. One of the earliest and most influential forms of lexical semantic analysis was componential analysis, i.e. the analysis of the meaning of a word like *man* as HUMAN + MALE + ADULT. However, by 1985 it was becoming clear that there was no easily identifiable list of "conceptual atoms" and following contemporary developments in the field, Miller adopted the idea that word meaning could be characterised in terms of systematic relationships to other words (Miller 1998a: xvi): for instance, *table* could be related to *furniture* by an IS-A-KIND-OF relation: this would not make the claim that *furniture* was a component of the meaning of *table*, merely that there was a systematic relationship of a particular kind between the meaning of *table* (whatever that was) and the meaning of *furniture* (whatever that was).

1 http://wordnet.princeton.edu/

The WordNet database is therefore structured in terms of a number of sense relations which appear to be psychologically relevant in the characterisation of word meaning. Further the database is organised around part of speech, on the basis of evidence that word storage in the mental lexicon is sensitive to part of speech. The current discussion relates to WordNet 3.0, which contains around 155,000 word forms (unique strings), of which just over 115,000 are nouns; the rest are verbs, adjectives and adverbs. In the following sections, we will review some of the main lexical sense relations that determine the organisation of WordNet.

## 2.2 Synonyms and synsets

The basic building block of WordNet is the synset or "set of synonyms" (Icelandic: *samheiti*; Greek: *syn* 'same' + *onyma* 'name'). In WordNet, synonymy is defined as having the same sense in a particular context.

(1) the nurse gave him a flu shot/injection/*pellet

- synset: = {shot, injection}

(2) the shot/pellet/*injection buzzed past his ear

- synset: = {shot, pellet}

Sentence (1) identifies a particular "sense", glossed in WordNet 3.0 as "the act of putting a liquid into the body by means of a syringe". This sense can be expressed by *shot* and by *injection* but not by *pellet*; *shot* and *injection* are therefore synonyms and form a synset. Sentence (2) identifies another "sense", glossed in WordNet 3.0 as "a solid missile discharged from a firearm". This second sense can be expressed by *shot* and by *pellet* but not by *injection*. This illustrates two important points about the organisation of WordNet.

First, the basic building block of the network is in fact a particular sense or concept; that sense can be expressed by one or more different word forms. This is very different from a traditional dictionary, whose basic building block is the word itself: the forms *shot* and *injection* would be listed separately in a traditional dictionary and each would be listed with the relevant sense as part of its entry. In WordNet, the sense itself represents a unique entry and the forms associated with

it are grouped in a synset. WordNet is sense-oriented; a traditional dictionary is word-oriented.

Second, WordNet does not distinguish between polysemy and homonymy. Polysemy is when a single word (lexeme) is associated with more than one sense. The word *shot* would be a good example, as it can express the sense associated with either Sentence (1) or Sentence (2). Homonymy is when two different words (lexemes) happen to have the same form: the classic example of this in English is the word-form *bank*, which can refer either to the side of a river or to a particular kind of financial institution. The intuition here is that the two senses are completely unrelated and that it is no more than a historical coincidence that they are expressed by the same word-form. WordNet remains completely agnostic on this distinction between polysemy and homonymy because its basic building block is the sense, each sense having one entry and being associated with a set of one or more word-forms which can express that sense in a certain context, i.e. the synset. It is the synset in WordNet which stands in sense-relations to other synsets and we will now review some of the main relations around which the database is structured.

## 2.3 Hyponymy~Hypernymy

Hyponymy is also known as the IS_A relation, typically the subkind relation. For instance, *mare* is a hyponym (Icelandic *undirheiti*; Greek: *hypo* 'under' + *onyma* 'name') of *horse*; and conversely, *horse* is a hypernym (Icelandic *yfirheiti*; Greek: *hyper* 'over' + *onyma* 'name') of *mare*, because *a mare is a (kind of) horse*. Hyponymy naturally creates hierarchies:

(3) a mare IS_A horse IS_A mammal IS_A animal

This is especially true of natural kinds, for which the hyponymy hierarchy can become quite articulated.

According to the hierarchy in Figure 1, both *mare* and *stallion* are hyponyms of *horse*, i.e. they are co-hyponyms; *animal* is the root of this hierarchy. In fact, WordNet has a considerably more articulated hierarchy than is shown here, with much greater depth. For instance, *stallion* is in fact a co-hyponym with *gelding*: both are *male horses* but the latter is castrated and the former not: this means that there is a lexical gap in the hierarchy because there is no specialised lexeme in

English for a male horse which covers both castrated and uncastrated varieties. WordNet sometimes fills this gap with multiword expressions: in this case, the hypernym for *stallion* and *gelding* is given as *male horse*, and it is *male horse* which is the co-hyponym of *mare*. At the top of the hierarchy, are a number of abstract terms which root the tree: so the top of the hierarchy for *mare* is not in fact *animal* but *entity* (*entity* is in fact at the root of all noun hyponymy hierarchies).

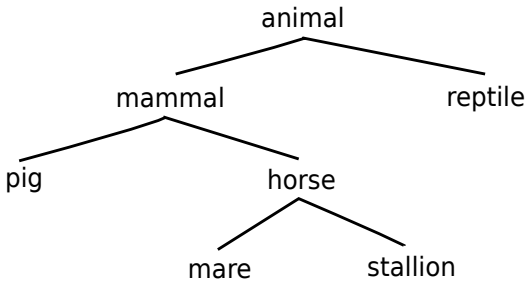



Figure 1. *A (partial) hyponymy hierarchy*

Such hierarchies are therefore lexical ontologies, i.e. classifications of the kinds of things that can be referred to in the language. Lexical ontologies must therefore confront the tension between scientific ontologies and folk ontologies, i.e. between the classification established as objective by the natural sciences and the classification established by popular usage and belief. A wordnet is often a compromise between these two and not always a consistent one. So, for instance, WordNet conforms to the scientific ontology for *whale*: it is given as a hyponym of *mammal* and is glossed as "any of the larger cetacean mammals having a streamlined body and breathing through a blowhole on the head". However, *tomato* is given as a hyponym of *vegetable* despite biologically being a *fruit*; nevertheless the gloss acknowledges the scientific classification and hints at the reason for the *vegetable*-classification: "mildly acid red or yellow pulpy fruit eaten as a vegetable", i.e. the hyponym relation is assigned on the basis of the use that is made of the entity, rather than its biological status – this is a functional hyponym not a nature-kind hyponym. It is important to stress here the difference between WordNet and a traditional dictionary: the main semantic information is the lexical semantic relation (hyponymy) and not the gloss; a computer using WordNet to build a semantic representation will treat *tomato* as a *vegetable*.

## 2.4 Meronymy~holonymy

Meronymy is the part-relation. For instance, *nose* is a meronym (Icelandic: *hlutheiti*; Greek *meros* 'part' + *onyma* 'name') of *face*; conversely, *face* is the holonym (Icelandic: *heildheiti*; Greek *holos* 'whole' + *onyma* 'name') of *nose*. The meronymy relation raises the important issue of modality: whether the relation actually must hold or merely can hold. With natural-kind hyponymy, the relation is necessary: every mare is a horse and no mare is not a horse. With meronymy, the relation is often one of possibility rather than necessity. So, for instance, meronyms of *face* include *beard*, which is only possible on some faces and never necessary. This shows that meronymy in WordNet is not even associated with typicality, as *beard* is not a typical part of *face* in general: women´s faces don´t typically have beards and even for men´s faces beards would only be typical in some cultures.

## 2.5 Antonymy

Antonymy is the relation of oppositeness and is important for the classification of adjectives.

(4) If the water is hot, then the water is not cold, and vice versa.

*Hot* is the antonym (Icelandic *andheiti*; Greek *anta* 'opposite' + *onyma* 'name') of *cold* and vice versa. It turns out, however, that not every adjective has an antonym, even when it is a synonym for an adjective that does. For instance, *torrid* is a synonym of *hot* (*hot/torrid weather*); *hot* is an antonym of *cold*; yet *torrid* is not an antonym of *cold*. Adjective networks in WordNet therefore often have a "bicycle" structure (cf. Figure 2).

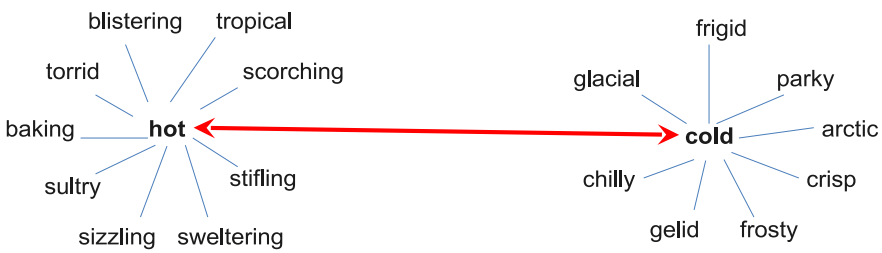



Figure 2. *The "bicycle" structure of antonymy in WordNet*

## 2.6 Troponymy

Verbs in WordNet 3.0 are largely organised in terms of the hyponymy relation, just like nouns. For instance, *to amble* is listed as a hyponym of *to walk*. Fellbaum (1998: 79) points out that verb hyponymy is not straightforwardly equivalent to noun hyponymy, however. She cites Lyons (1977: 294) for the observation that verbs do not fit naturally into the hyponymy paraphrase for nouns, without nominalisation:

(5) x is a (kind of) y

(6) ?To amble is (*a kind of) to walk

(7) ?Ambling is (a kind of) walking

Even Sentence (7) would not necessarily be accepted by all native speakers. However, if the manner aspect of the relation is emphasised, then the paraphrase works much more effectively:

(8) To amble is to walk in an ambling manner.

(9) To $V_1$ is to $V_2$ in a particular manner/way.

Fellbaum & Miller (1990) dub such manner hyponyms, **troponyms**. As Fellbaum goes on to observe, however, the complexity of the relation between a verb and its hypernym is much richer and more complex than this paraphrase suggests and I will leave this issue here.

We will now turn to another major wordnet resource which explicitly addresses some of the problems with the original Princeton WordNet – DanNet, a wordnet for Danish.

# 3 DanNet

DanNet[2] is a lexical semantic database for Danish, conforming to international wordnet standards; it was developed from a monolingual Danish dictionary (*Den Danske Ordbog*, DDO) as a collaborative project between SIMPLE-DK (Centre for Language Technology at the University of Copenhagen) and the publisher of DDO (the Society for Danish Language and Literature, Danish Ministry of Culture). As of 2009, it con-

2 See www.wordnet.dk for download and general information; see http://andreord.dk/ord/ to browse DanNet

tained 50,000 synsets. Pedersen et al. (2009) report on the problems of converting a human-use dictionary to a lexical semantic database and the limitations of using the classical lexical semantic relations of the Princeton WordNet. A good example of both problems comes in the discussion of the word *butik* ‚shop':

> For example all hyponyms of *butik* (shop) inherit the involved agent *handlende* (shopkeeper). Thus, the DanNet editor is prompted to identify the involved agent of the more restricted hyponym: that the shopkeeper of a pharmacy is a pharmacist, the shopkeeper of a bakery is a baker and so on. Such information is only rarely specified in DDO definitions (although sometimes provided implicitly as examples of word formation), but this information is seen as highly relevant in a wordnet.
>
> (Pedersen et al. 2009: 273)

The DanNet developers address two issues in this passage.

The first issue is the gap between a traditional dictionary and an NLP resource: the traditional dictionary entries of DDO leave implicit the relation between a pharmacy and a pharmacist, a bakery and a baker, and so on, because the inference of such a relation can be left to the world knowledge and common sense of a human user; a computer however will not automatically make such inferences. The DanNet developers therefore state the shopkeeper for every subkind of shop, where such an entity is lexicalised. The manual process of adding such information is streamlined by exploiting the inheritance relation: an algorithm is set to prompt the developer for "missing information" that can be inferred on the basis of established relations; so, if a relation is explicitly stated for a hypernym, then it is likely that all hyponyms will have a specific equivalent of this relation: if all shops have a shopkeeper (involved agent) and a bakery is shop (hyponymy), then a bakery will have a shopkeeper, for whom the language may well have a specialised lexical item.

The second issue implied in the quoted passage concerns the limits of the original WordNet relation set. Notice that the information being added to DanNet here (involved agent) is not a relation in the original WordNet: speakers of a language implicitly understand a systematic relation between an entity and its typical owner or user; this information is therefore added systematically to DanNet. In fact, the developers of DanNet extend the classical relations of WordNet, following the work of the computational semanticist, Pustejovsky,

whose *Generative Lexicon* (Pustejovsky 1995) proposes four "qualia roles" that form part of the representation of noun meaning:

(10) Formal: contrastive relation to other objects (affecting inheritance relations)

(11) Agentive: origin

(12) Constitutive: how an object is made up (its parts and organisation)

(13) Telic: purpose or function

The range of relations used in DanNet are shown in Figure 3 (based on Pedersen et al. 2009, Figure 12).

| **FORMAL** | **AGENTIVE** | **CONSTITUTIVE** | **TELIC** |
|---|---|---|---|
| has_hypernym | made_by | has_holo_made_of | used_for |
| has_hyponym | | has_mero_made_of | used_for_object |
| is_a_way_of (troponym) | | has_holo_part | role_agent |
| | | has_mero_part | role_patient |
| | | has_holo_member | |
| | | has_mero_member | |
| | | has_holo_location | |
| | | has_mero_location | |
| | | concerns | |
| | | involved_agent | |
| | | involved_patient | |
| | | involved_instrument | |

Figure 3. *Relations in DanNet*

We have already seen how involved_agent can be used to link senses. Notice also the more fine-grained range of distinctions that have been added to the meronymy~holonymy relation. As we saw earlier, the traditional meronymy relation relates to typical parts, so a cabin might have a roof as a part. However, a log-cabin will have logs as a part, in the sense that it is made of logs, a slightly different kind of part relation. Similarly, a congregation may have a minister as a "part" in the sense that the minister is a member of the congregation; and England has London as a "part" in the sense that it is a location within the larger location. These distinctions have important implications for inferencing: a human user will reflexively accommodate

them but a computer must be provided with the information explicitly and systematically.

These changes to the original WordNet relation set are essentially extensions and elaborations. However, DanNet also addresses a fundamental problem with the original relation which provides the backbone to Princeton WordNet: hyponymy. To be useful for inferencing, the hyponymy relation should: (i) hold of subkinds (a dog is a kind of animal; a cat is a kind of animal), where (ii) co-hyponyms are mutually exclusive (if something is a dog then it is not a cat and vice versa). Pedersen et al. (2009: 277) point out that the traditional use of hyponymy covers a broader range of relations than this, in a way which is problematic for NLP applications: so-called, **hyponymy overload**. Consider the following examples:

(14) *oliemaleri* 'oil painting', *blomstermaleri* 'flower painting', *fidusmaleri* 'pseudo-art', *akvarel* 'water colour', *marinebillede* 'seascape', *klatmaleri* 'daub'

Each of these terms is a hyponym of *maleri* 'painting'. However, if one thinks in terms of mutually-exclusive subkinds, two candidate pairs emerge:

(15) *oliemaleri* 'oil painting' vs *akvarel* 'water colour'

- subkinds of painting distinguished by the paint used

(16) *blomstermaleri* 'flower painting' vs *marinebillede* 'seascape'

- subkinds of painting distinguished by the subject depicted

It is perfectly possible to have an oil painting which is also a flower painting – and it is perfectly possible that that item is also a "daub". Notice that the two mutually exclusive pairs are exclusive along a particular dimension: paintings classified by the kind of paint used; paintings classified by the subject depicted. Pedersen et al. therefore adopt a proposal by Huang et al. (2008) which allows hyponyms defined by a particular dimension of description to be grouped together: *oliemaleri* 'oil painting' and *akvarel* 'water colour' are said to be **paranyms**, terms associated with the same dimension of description. The paranym relation allows co-hyponyms to be clustered into mutually-exclusive subsets.

The terms *fidusmaleri* 'pseudo-art' and *klatmaleri* 'daub' remain problems, however, because any painting can in fact be a daub: this

term does not really describe a subkind as such but rather a subjective evaluation of an item; a daub may indeed be a kind of painting but any painting can be termed a daub if the speaker assesses its quality to be at a certain level. Terms like *klatmaleri* 'daub' therefore cut across the hyponyms of *maleri* 'painting': they are orthogonal to the classification. DanNet therefore allows the hyponym relation to be tagged with a feature ORTHO which indicates that the term represents an evaluation that can apply to any "co-hyponym" of the term.

Even this relatively brief discussion illustrates well the challenges that face the construction of a wordnet which is to be sufficiently richly and systematically elaborated to be used in advanced NLP applications. We will now turn to a third resource, quite unlike the two that we have reviewed so far, which is developed around a very different aspect of sense organisation.

## 4 SALDO

SALDO[3] is "a free full-scale modern Swedish semantic and morphological lexical resource intended primarily for use in language technology applications" (Borin & Forsberg 2009: 7). It is based on a much looser associative relation that we typically find in wordnets – especially as the relation is not sensitive to part of speech. In fact there is only one obligatory relation in SALDO (mother) and one optional relation (father). The mother will be a more central concept, i.e. semantically and/or morphologically less complex, probably more frequent, stylistically more unmarked, and acquired earlier in first and second language acquisition. The father will be a differentiating term (often a domain-specifier) (Borin & Forsberg 2009: 7f). For example, the noun *sol* 'sun' has as a mother the verb *lysa* 'shine'; the father of *sol* 'sun' is *himmel* 'sky', i.e. the domain or context in which the shining takes place (Borin & Forsberg 2009: 10). The simplest way to grasp the essential intuition upon which SALDO´s semantic classification is built is to imagine how you would define a word if you were dealing with someone with very limited vocabulary. You might attempt to indicate what the sun was by saying that it was the thing in the sky which shines. Shining is the most salient characteristic of the sun and the sky is the place that one needs to look to find it.

3 http://spraakbanken.gu.se/eng/saldo

Where a wordnet has a hyponymy hierarchy, SALDO has a centrality hierarchy based on motherhood. So, *sol* 'sun' has a number of siblings that share the same mother, *lysa* 'shine':

(17) verbs: *inform, sparkle, shine, twinkle, shimmer, lustre, flash, glitter, glimmer, glisten, gleam, flimmer, blink, illuminate*; nouns: *light, star, moon, lantern, lamp, comet, flash, candle, light house;* adjectives: *shining, fluorescent, light/bright*

Some of these are full siblings that also share the same father, *himmel* 'sky':

(18) *comet, moon, star*

At the core of SALDO are the roots of these hierarchies: 51 lexical primitives on which all other items depend (Borin & Forsberg 2009: 9, their Figure 1).

(19) *all* 'all', *annan* 'other', *använda* 'use', *att* 'that', *bara* 'only', *bra* 'good', *genom* 'through', *den* 'it', *fort* 'fast', *framme* 'arrived', *färg* 'color', *för2* 'for', *förbi* 'gone/past', *före* 'before', *en2* 'a/one', *göra* 'do', *ha* 'have', *hur* 'how', *hända* 'happen', *i2* 'in', *ja* 'yes', *just* 'just', *kunna* 'be able', *ljud* 'sound', *ljus* 'light', *med* 'with', *men* 'but', *mycken* 'much', *måste* 'must', *namn* 'name', *natur* 'nature', *när* 'when', *och* 'and', *om* 'if', *om2* 'about', *på* 'on', *rak* 'straight', *röra* 'move', *säga* 'say', *tal* 'speech', *till* 'to', *tänka* 'think', *vad* 'what', *var* 'where', *vara* 'be', *varm* 'warm', *vem* 'who', *veta* 'know', *vid* 'by', *vilja* 'want', *öppen* 'open'

This way of looking at the semantic relations between words is obviously very different from the wordnets. One striking difference, when considering the roots of the hierarchies, is that in WordNet we find abstract terms like "entity" which are added to draw together the forest of more lexically articulated and conceptually substantive trees beneath, whereas in SALDO we find highly frequent and often substantive terms such as "light" and "warm" and "say". This is because SALDO is driven to a large extent by conceptual saliency and centrality and to that extent it is reminiscent of the core vocabulary in Wierzbicka and Goddard's Natural Semantic Metalanguage (NSM)[4] (Wierzbicka 1996; Goddard 2008), with which Borin & Forsberg (2009: 8f) compare their work.

4 http://www.une.edu.au/bcss/linguistics/nsm/semantics-in-brief.php

NSM was developed in support of a program of "reductive paraphrase", in which the meaning of complex expressions is given using simple terms. The simple terms express irreducible fundamental concepts which have exponents in all languages. NSM is therefore intended as a kind of universal conceptual interlingua. Like SALDO, the primitive terms of NSM are descriptively substantive and relatively high frequency; of the 51 lexical primitives of SALDO and 61 semantic primitives of NSM, there are 17 shared terms, including: good, do, think, want, when, where, not, if. It proves to be significant, however, that NSM aims at a set of universal paraphrase terms which can be used for building sense definitions of lexical items in all languages, whereas SALDO (SALDO *Instruktion*, p. 10) aims at "så homogena och intuitivt tilltalande horisontella lexemklasser som möjligt"[5] for Swedish, in which the small lexical groupings emerge organically from the internal properties of the Swedish vocabulary system, rather than being imposed externally from a preconceived typology ("Större strukturer i lexikonet växer fram organiskt, utan kontroll 'uppifrån' eller 'utifrån'."[6]) One nice example of this is the relative centrality of comparative *like*. It is a central term in NSM because the relation of comparison is understood as a primitive conceptual relation. It is, however, four steps from the core of SALDO. Another example discussed by Borin & Forsberg (2009: 9) concerns antonymy. In SALDO, antonyms can be related by a mother-child relation: in SALDO, the mother of *dålig* 'bad' is *bra* 'good'; the father of *dålig* 'bad' is *motsats* 'opposite'. So in SALDO, *bra* 'good' is treated as a primitive concept and *dålig* 'bad' derived with respect to it by opposition or contrast; in NSM, *good* and *bad* are treated as primitive evaluative terms which can be used to paraphrase classes of more complex expressions.

Although SALDO and NSM differ radically from the wordnets in the kinds of terms that we find at the roots of their hierarchies, they nevertheless show significant differences related to their contrasting attitudes to universal conceptual structure versus language-particular lexical organisation.

---

5 "a horizontal grouping of lexemes which is as homogeneous and intuitively appealing as possible" (my translation).

6 "Larger structures in the lexicon develop organically, without imposition 'from above' or 'from outside'." (my translation).

## 5 Conclusion

This paper began by setting up a contrast between the demands placed on the traditional dictionary for human use and the lexical resource for NLP use. As the human user brings a vast amount of world and cultural knowledge to the task of dictionary use, supplemented by robust common sense reasoning skills, the dictionary creator can assume all sorts of semantic information as understood; as a computer brings nothing to the lexical semantic resource, independent of the algorithms it has been programmed with, the creator of an NLP resource must include a rich set of information in a systematic and explicit manner and in a format which is suitable for algorithmic manipulation. It is not surprising then to find the creators of each of these resources treading the delicate line between the modelling of linguistic organisation and of conceptual organisation.

As the final discussion concerning the differences between SALDO and NSM show, there is also a tension between potentially universal properties of linguistic organisation and the idiosyncratic properties of particular languages. NSM aims at a universal paraphrase language for the conceptual primitives underlying lexical organisation in human languages; SALDO is emphatically monolingual in its approach. The tension between universal and particular is built into WordNet: at the root of the WordNet hierarchies are abstract terms such as "entity" which serve to root the forest of hyponymy hierarchies beneath them and which are likely to be shared by wordnets for other languages; but the bulk of the relational information represented is potentially idiosyncratic and reflected in the distribution of lexical gaps and the elaboration of hyponymy distinctions further down the tree. Nevertheless, the Princeton WordNet was developed as an analysis of English lexical semantic organisation and as such is a monolingual resource. Similarly, DanNet was explicitly monolingual in its methodology, basing its structure on a monolingual corpus-based dictionary, rather than translation from the Princeton WordNet. This monolingual emphasis is shared by both Icelandic resources presented in this volume, which seek to characterise the lexical semantic organisation of Icelandic in its own terms, without importing a structure from resources developed for other languages (e.g. by translation of WordNet or DanNet).

Another important characteristic shared by all three of the resour-

ces surveyed here is that they are manually constructed. This places an enormous burden on project resources in terms of time, money and manpower. For a small community such as Iceland, this is a critical issue (Rögnvaldsson et al. 2009). In this respect, the two Icelandic projects differ in approach but both provide reason for cautious optimism. Jón Hilmar Jónsson´s *Íslenskt orðanet* adopts a manual methodology and yet despite the practical constraints that this imposes has achieved impressive progress in developing a monolingual sense-oriented resource for Icelandic; Anna Björk Nikulásdóttir´s *Íslenskur merkingarbrunnur* is developing a variety of semi-automatic methods to extract lexical semantic relations from text corpora, which is currently showing promising results. It is to be hoped that the contrasting methodologies (semi-automatic and manual) will prove to be complementary and allow the two projects to collaborate effectively in the development of robust lexical semantic resources for Icelandic.

## References

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science* 34: 222–254.

Borin, L., & Forsberg, M. 2009. All in the Family: A Comparison of SALDO and WordNet. In: B.S. Pedersen, A. Braasch, S. Nimb, and R. Vatvedt Fjeld, (Eds.). *Proceedings of the Workshop "Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies" at 17th Nordic Conference on Computational Linguistics (NODALIDA) 14-16th May 2009*, NEALT Proceedings Series Volume 7, pp. 7–12. Odense, Denmark.

Cruse, D. A. 1991. *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. 2002. Hyponymy and its varieties. In: R. Green, C. A. Bean, & S. H. Myaeng (Eds.). *The semantics of relationships: An interdisciplinary perspective, information science and knowledge management*, pp. 2–21. Springer.

DDO = *Den Danske Ordbog* (´The Danish dictionary´) 1–6. 2003–5. Eds.: E. Hjorth, K. Kristensen, et al. Copenhagen: Gyldendal and Society for Danish Language and Literature.

Fellbaum, C. (Ed.) 1998a. *WordNet. An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C. 1998b. Introduction. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. 1–19. MIT Press, Cambridge, MA.

Fellbaum, C., & Miller, G.A. 1990. Folk psychology or semantic entailment? A reply to Rips and Conrad. *The Psychological Review* 97: 565–570.

Huang, C., Hsiao, P., Su, I., & Ke, X. 2008. Paranymy: Enriching ontological knowledge in WordNets. In*: Proceedings of the fourth global WordNet conference*, pp. 221–228. Szeged, Hungary.

Goddard, C. (Ed.) 2008. *Cross-Linguistic Semantics*. Amsterdam: John Benjamins.

Jónsson, J.H. 2008. Í áttina að samfelldri orðabók – nokkrir megindrættir í Íslensku orðaneti. *Orð og tunga* 10: 29–45.

Jónsson, J.H. 2009a. Ordforbindelser: Grunnelementer i ordboken? *Lexico-Nordica* 16: 161–179.

Jónsson, J.H. 2009b. Lemmatisation of Multi-word Lexical Units: Motivation and Benefits. In: H. Bergenholtz, S. Nielsen & S. Tarp, (Eds.). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, pp. 165–194. Bern: Peter Lang.

Jónsson, J.H. 2009c. Lexicographic description: An onomasiological approach on the basis of phraseology. In: S. Nielsen & S. Tarp, (Eds.). *Lexicography in the 21st Century. In honour of Henning Bergenholtz,* pp. 257–280. Amsterdam: John Benjamins Publishing Company.

Miller, G.A. 1998a. Foreword. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. xv–xxii. Cambridge, MA: MIT Press.

Miller, G.A. 1998b. Nouns in Wordnet. In: C. Fellbaum (Ed.). *WordNet. An Electronic Lexical Database*, pp. 23–46. Cambridge, MA: MIT Press.

Nikulásdóttir, A. & Whelpton, M. 2010a. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17: 141–161.

Nikulásdóttir, A. & Whelpton, M. 2010b. Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In: *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (Workshop 22 of 7th Language Resources and Evaluation Conference)*, pp. 33–39. Valletta, Malta.

Nikulásdóttir, A. & Whelpton, M. 2009. Automatic extraction of semantic relations for less-resourced languages. In: B.S. Pedersen, A. Braasch, S. Nimb, and R. Vatvedt Fjeld, (Eds.). *Proceedings of the Workshop "Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies" at 17th Nordic Conference on Computational Linguistics (NODALIDA) 14–16th May 2009*, NEALT Proceedings Series Volume 7, pp. 1–6. Odense, Denmark.

Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L. & Lorentzen, H. 2009. DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. *Language Resources and Evaluation* 43: 269–299.

NSM. http://www.une.edu.au/bcss/linguistics/nsm/semantics-in-brief.php. (9th June 2011).

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Rögnvaldsson, R., Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Nikulásdóttir, A., Whelpton, M., & Ingason, A.K. 2009. Icelandic Language Resources and Technology: Status and Prospects. In: R. Domeij, K. Kosken-

niemi, S. Krauwer, B. Maegaard, E. Rögnvaldsson, and K. de Smedt, (Eds.). *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, NEALT Proceedings Series Volume 5, pp. 27–32. Odense, Denmark.

SALDO *Instruktion*. Web resource: https://svn.spraakdata.gu.se/repos/sblex/pub/saldo_instruktion.pdf (8.11.2011).

Wierzbicka, A. 1996. *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Úlfarsdóttir, Þ. 2006. Málfræðileg mörkun orðasambanda. *Orð og tunga* 8: 117–144.

## Útdráttur

Orðabækur eru gerðar fyrir fólk en hins vegar eru mörg rafræn orðfræðileg málsöfn sett saman með tölvur í huga. Þeim er ætlað að geyma upplýsingar um form, notkun og merkingu orða á þann hátt að tölvur geti greint mannlegt mál á markvissan hátt, til þess að draga fram upplýsingar úr textum eða tali og til þess að draga ályktanir af þeim upplýsingum sem þannig er aflað. Ganga má út frá því að lifandi notendur viti ýmislegt fyrirfram, annaðhvort af skynsemi sinni eða almennri þekkingu, en aftur á móti hefur tölva enga fyrirfram gefna vitneskju. Í greininni er fjallað um ýmiss konar dæmigerðar merkingarfræðilegar upplýsingar í svonefndum orðanetum eins og WordNet (fyrir ensku) og DanNet (fyrir dönsku), en einnig í orðfræðilegum gagnasöfnum sem eru í grundvallaratriðum annarrar gerðar eins og SALDO (fyrir sænsku).

## Lykilorð

merkingarfræði, merking orða, málgreining, orðanet, merkingarvensl

## Keywords

lexical semantics, natural language processing, wordnets, semantic relations

*Dr. Matthew Whelpton*
*Faculty of Foreign Languages, Literature and Linguistics*
*University of Iceland*
*whelpton@hi.is*

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] *lopa-peysu* 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; það – 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):

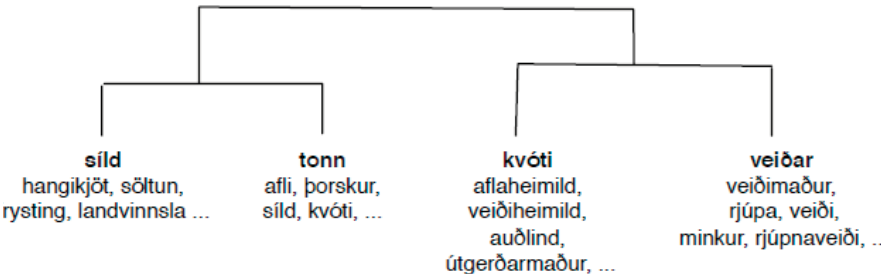
(10) **tonn,** afli, þorskur, síld, kvóti, veiðar, skip, togari, ýsa, loðna, kolmunni, króna, heildarafli, milljón, aflaverðmæti, útgerð, ufsi, steinbítur, vertíð, verðmæti, fisktegund, löndun, [...]

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.

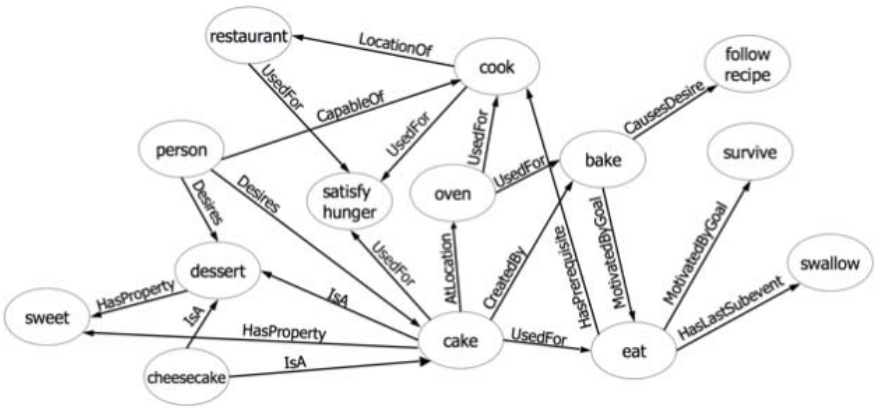
Formgerð íslenska merkingarbrunnsins hefur ekki verið greind að fullu en eins og sjá má hér að ofan eru merkingarupplýsingarnar um einstök orð margvíslegar og ekki alltaf nákvæmlega skilgreinanlegar. Ólíklegt er einnig að merkingarbrunnurinn myndi heildstætt net orða. Frekar má búast við að orð innan 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

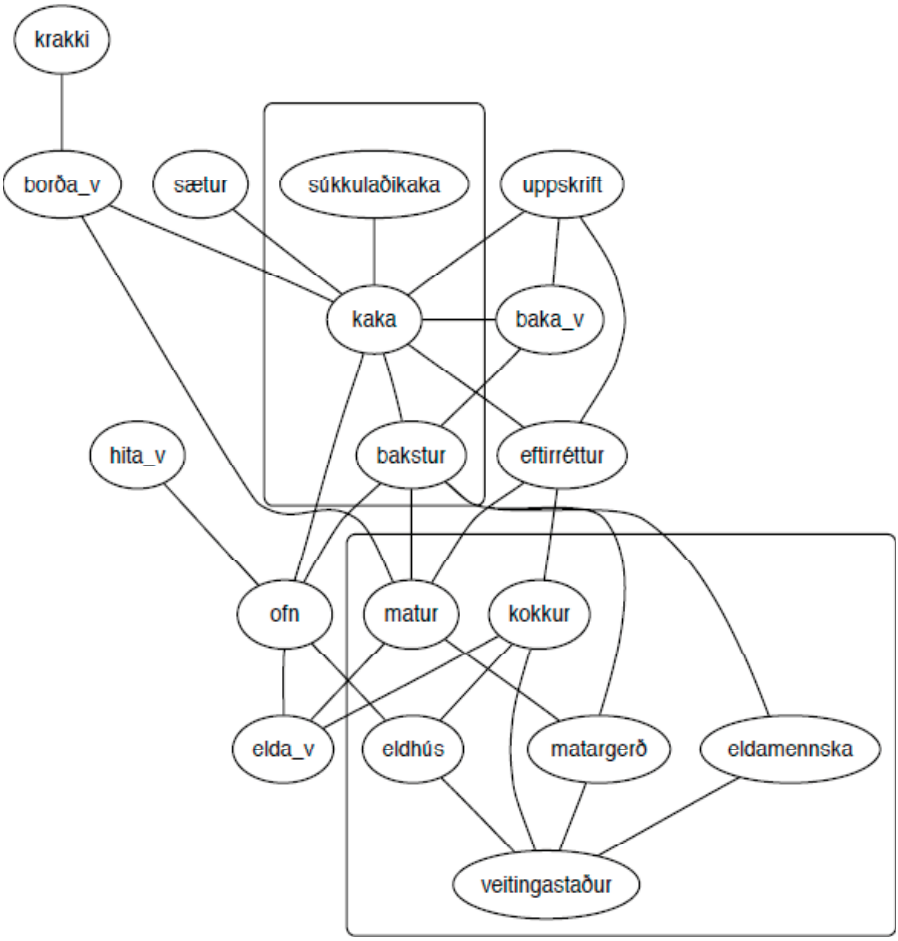
inniheldur 21 merkingarvensl og að auki ein vensl sem kallast *ConceptuallyRelatedTo* sem eru ekki nánar skilgreind. Á mynd 2 má sjá vensl eins og *UsedFor, HasProperty* og *IsA*. Venslin *HasProperty* eru sambærileg við venslin eiginleiki í merkingarbrunninum og *IsA* eru yfirheitavensl. Merkingarbrunnurinn 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 leiðréttingmun 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 orðabóka-notkun.

## 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 Co-ordination 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*

Jón Hilmar Jónsson

# Að fanga orðaforðann: orðanet í þágu orðabókar

## 1 Inngangur

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

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

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

Orð og tunga 14 (2012), 39–65. 

## 2 Áskorun um endurmat orðabókarlýsingar

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

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

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

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

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

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

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

## 3 Forsendur og efniviður

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

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

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

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

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

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

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

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

## 4 Viðfangsefni og markmið

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

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

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

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

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

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

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

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

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

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

## 5 Endurmótun flettulistans, einræðing og fleiryrtar flettur

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

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

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

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

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

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

## 5.1 Fleiryrtar sagnaflettur

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

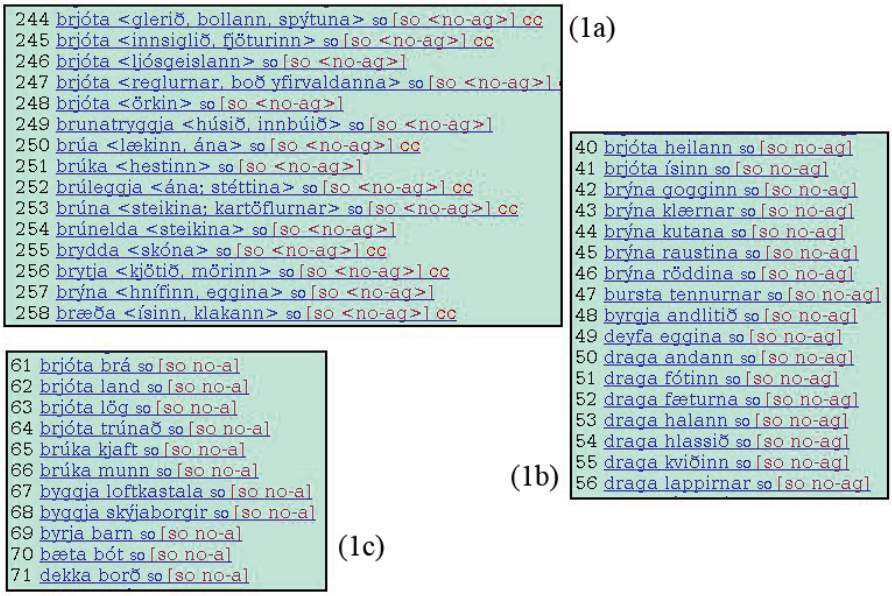
Með tilliti til flokkunar er þó mikilvægara að athuga hvernig tiltekin mörk koma fram í heild meðal flettnanna, óháð einstökum sögnum. Á mynd 1 koma fram stafrófsraðaðir bútar úr flettulista þriggja marka. Í (1a) er nafnliðurinn breytilegur og þar beinist athyglin að fallstjórn og merkingareinkennum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

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

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

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

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




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

## 5.2 Nafnliðaflettur

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

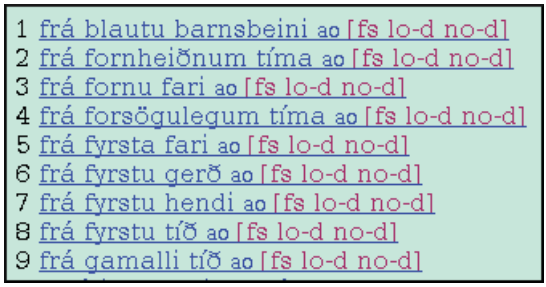
*og mannleg samskipti*, *afbrýðisemi og illt umtal*, *gróið land og bithagar*. Gildi þeirra felst hér að nokkru leyti í því að skerpa merkingarleg einkenni þeirra flettna sem þær tengjast en um leið búa þær yfir merkingareinkennum sem beina þeim í samheitasambönd og stærri merkingarflokka. Aðrir nafnliðir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

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

## 5.3 Atviksliðaflettur

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

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




Mynd 2: *Mörkunarstrengir fleiryrtra atviksliðaflettna draga fram orðbundin og setningarleg munstur, sem oft endurspegla merkingarleg vensl.*

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

## 5.4 Fleiryrtar lýsingarorðaflettur

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

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

(7) 1 vera sofandi (sofa) [so lo-n]
2 vera sofandi (vera sljór/viðutan) [so lo-n]

(8) 1 vera klár (vera skýr/skarpur) [so lo-n]
2 vera klár (vera undirbúinn) [so lo-n]

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

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

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

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

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

# 6 Gagnaefni, gagnategundir og merkingargreining

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

## 6.1 Flettan *hlátur*

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

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

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

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

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

## 6.2 Merkingarflokkun og samheiti

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

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

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

## 6.3 Orðapör sem gagnategund

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

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

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

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

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

## 6.4 Mat á merkingarskyldleika

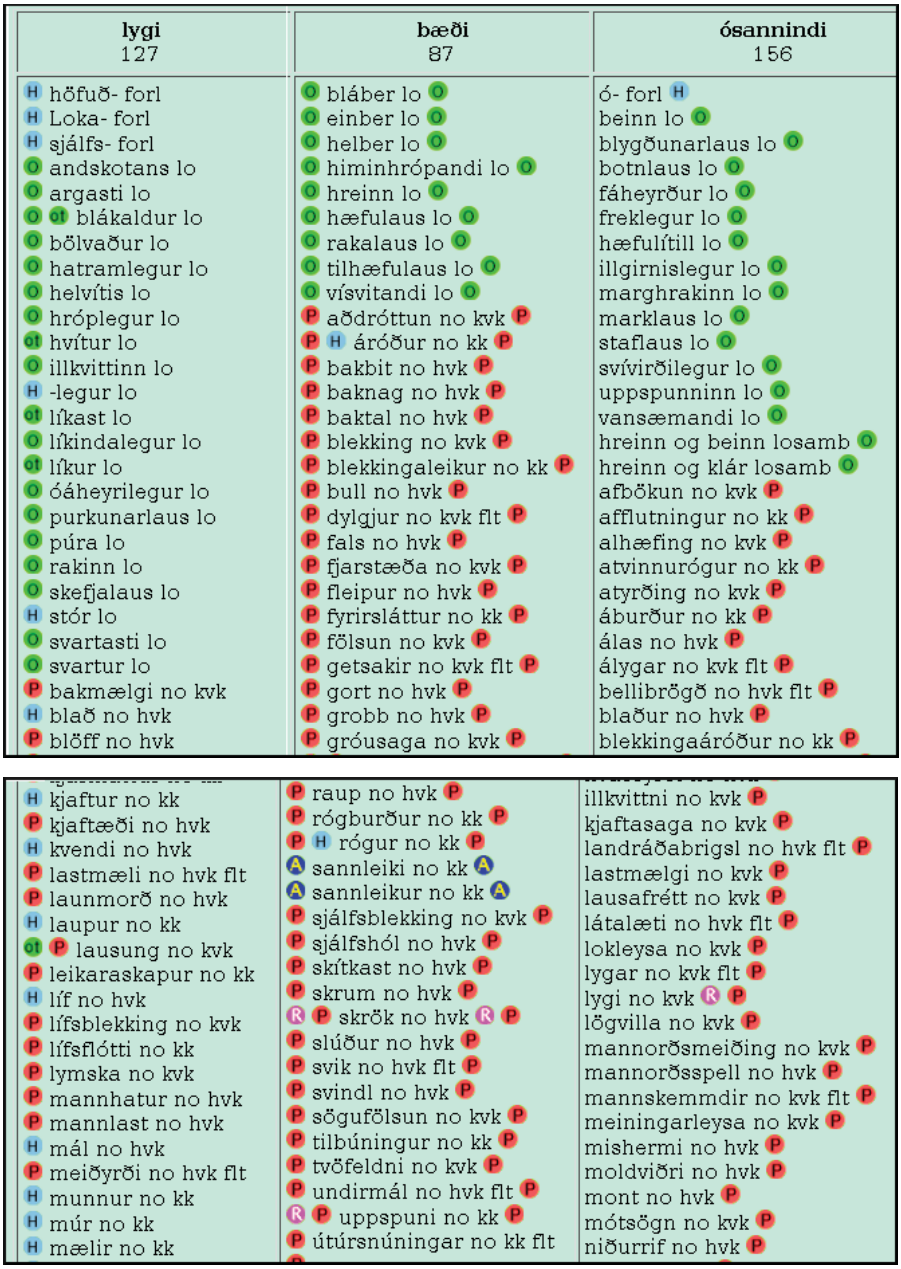
Gerð orðanetsins hvílir eins og áður segir á þeirri meginhugmynd að merkingarvensl megi rekja út frá venslum orða í orðasamböndum og samsetningum. Þá er nærtækt að líta svo á að styrkur merkingarvenslanna, merkingarleg nálægð eða skyldleikastig flettnanna hverrar gagnvart annarri, komi fram í því hversu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

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

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

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

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



| lygi<br>127 | bæði<br>87 | ósannindi<br>156 |
|---|---|---|
| H höfuð- forl | O bláber lo O | ó- forl H |
| H Loka- forl | O einber lo O | beinn lo O |
| H sjálfs- forl | O helber lo O | blygðunarlaus lo O |
| O andskotans lo | O himinhrópandi lo O | botnlaus lo O |
| O argasti lo | O hreinn lo O | fáheyrður lo O |
| O ot blákaldur lo | O hæfulaus lo O | freklegur lo O |
| O bölvaður lo | O rakalaus lo O | hæfulítill lo O |
| O hatramlegur lo | O tilhæfulaus lo O | illgirnislegur lo O |
| O helvítis lo | O vísvitandi lo O | marghrakinn lo O |
| O hróplegur lo | P aðdróttun no kvk P | marklaus lo O |
| ot hvítur lo | P H áróður no kk P | staflaus lo O |
| O illkvittinn lo | P bakbit no hvk P | svívirðilegur lo O |
| H -legur lo | P baknag no hvk P | uppspunninn lo O |
| ot líkast lo | P baktal no hvk P | vansæmandi lo O |
| O líkindalegur lo | P blekking no kvk P | hreinn og beinn losamb O |
| ot líkur lo | P blekkingaleikur no kk P | hreinn og klár losamb O |
| O óáheyrilegur lo | P bull no hvk P | afbökun no kvk P |
| O purkunarlaus lo | P dylgjur no kvk flt P | afflutningur no kk P |
| O púra lo | P fals no hvk P | alhæfing no kvk P |
| O rakinn lo | P fjarstæða no kvk P | atvinnurógur no kk P |
| O skefjalaus lo | P fleipur no hvk P | atyrðing no kvk P |
| H stór lo | P fyrirsláttur no kk P | áburður no kk P |
| O svartasti lo | P fölsun no kvk P | álas no hvk P |
| O svartur lo | P getsakir no kvk flt P | álygar no kvk flt P |
| P bakmælgi no kvk | P gort no hvk P | bellibrögð no hvk flt P |
| H blað no hvk | P grobb no hvk P | blaður no hvk P |
| P blöff no hvk | P gróusaga no kvk P | blekkingaáróður no kk P |

| | | |
|---|---|---|
| H kjaftur no kk | P raup no hvk P | illkvittni no kvk P |
| P kjaftæði no hvk | P rógburður no kk P | kjaftasaga no kvk P |
| H kvendi no hvk | P H rógur no kk P | landráðabrigsl no hvk flt P |
| P lastmæli no hvk flt | A sannleiki no kk A | lastmælgi no kvk P |
| P launmorð no hvk | A sannleikur no kk A | lausafrétt no kvk P |
| H laupur no kk | P sjálfsblekking no kvk P | látalæti no hvk flt P |
| ot P lausung no kvk | P sjálfshól no hvk P | lokleysa no kvk P |
| P leikaraskapur no kk | P skítkast no hvk P | lygar no kvk flt P |
| H líf no hvk | P skrum no hvk P | lygi no kvk R P |
| P lífsblekking no kvk | R P skrök no hvk R P | lögvilla no kvk P |
| P lífsflótti no kk | P slúður no hvk P | mannorðsmeiðing no kvk P |
| P lymska no kvk | P svik no hvk flt P | mannorðsspell no hvk P |
| P mannhatur no hvk | P svindl no hvk P | mannskemmdir no kvk flt P |
| P mannlast no hvk | P sögufölsun no kvk P | meiningarleysa no kvk P |
| H mál no hvk | P tilbúningur no kk P | mishermi no hvk P |
| P meiðyrði no hvk flt | P tvöfeldni no kvk P | moldviðri no hvk P |
| H munnur no kk | P undirmál no hvk flt P | mont no hvk P |
| H múr no kk | R P uppspuni no kk P | mótsögn no kvk P |
| H mælir no kk | P útúrsnúningar no kk flt | niðurrif no hvk P |

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

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

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

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

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

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

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

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

## 7 Birtingarform og vefsíðan ordanet.is

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

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

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

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

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

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


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

# 8 Niðurlag

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

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

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

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

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

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

## Ritaskrá

Anna Björk Nikulásdóttir og Matthew Whelpton. 2010. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17: 141–161.

Dornseiff, Franz. 2004. *Der deutsche Wortschatz nach Sachgruppen.* 8., völlig neu bearbeitete und mit einem vollständigen alphabetischen Zugriffsregister versehene Auflage von Uwe Quasthoff. Berlin: Walter de Gruyter.

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

## Abstract

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

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

## Lykilorð

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

## Keywords

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

*Jón Hilmar Jónsson*
*Stofnun Árna Magnússonar í íslenskum fræðum / Háskóli Íslands*
*Neshaga 16*
*107 Reykjavík*
*jhj@hi.is*

Anna Helga Hannesdóttir

# Orðfræðirit frá fyrri tíð

*Magnús Ólafsson of Laufás: Specimen lexici runici and Glossarium priscæ linguæ danicæ*. Ritstjórar Anthony Faulkes og Gunnlaugur Ingólfsson. Orðfræðirit fyrri alda V. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum og London: Viking society for northern research, University College. 2010. (xlvi + 492 bls.) ISBN 978-9979-654-08-7 (ísl. útgáfa)/978-0-903521-80-2 (bresk útgáfa).

## 1 Inngangur

Í ritröðinni *Orðfræðirit fyrri alda* er nú 5. bindið komið út. Það felur í sér tvö rit sem bæði urðu til um miðja 17. öld: orðabókina *Specimen lexici runici* og handritið DG 55. Orðabókin var gefin út í Kaupmannahöfn 1650 og þar er því um endurútgáfu að ræða. Handritið DG 55, sem geymir „Glossarium priscæ linguæ danicæ“, hefur aftur á móti ekki verið gefið út áður og birtist hér í fyrsta sinn. Að útgáfunni standa Stofnun Árna Magnússonar í íslenskum fræðum og Viking Society for Northern Research. Ritstjórar eru þeir Anthony Faulkes og Gunnlaugur Ingólfsson.

Í orðabókum þeim sem gefnar hafa verið út í ritröðinni hingað til hefur áhersla verið lögð á að fjalla um einkenni þessara gömlu orðabóka og gildi þeirra fyrir íslenska mál- og orðsögu. Útgáfurnar eru m.a. ætlaðar sem „framlag til íslenskra orðfræðirannsókna“ (Orðfræðirit IV:[vii]). Sjónarmiðið hefur því einkum verið orðfræðilegt. Í útgáfu þeirri sem hér er til umfjöllunar er markmiðið þó aðeins annað en það sem áður hefur verið ríkjandi. Ritstjórarnir láta sér ekki

Orð og tunga 14 (2012), 67–75. 

nægja að gera þessi gömlu ritverk aðgengileg fyrir nýjar kynslóðir fræðimanna sem hafa áhuga á íslenskum orðaforða 17. aldarinnar. Öllu fremur vilja þeir taka við þar sem hin íslenska endurreisn 17. aldar lét staðar numið við að varpa ljósi á orðaforðann í íslenskum miðaldabókmenntum. Við það bætist sjónarmið sem nánast varðar handritafræði og fílólógíu.

Allt ritstjórnarefni er á ensku. Hin nýja útgáfa er því aðgengileg alþjóðlegum fræðimönnum án kunnáttu í íslensku, á sama hátt og skrif 17. aldar fræðimanna á latínu.

Hér verður farið yfir útgáfuna 2010 í stórum dráttum. Einnig verður í stuttu máli fjallað um *Specimen lexici runici* 1650 sem orðabók og hún tengd norrænni orðabókahefð fyrri alda.

## 2 Efni útgáfunnar 2010

*Specimen lexici runici* (SLR) og handritið DG 55 eru gefin út í mjúku bandi, á þykkum glanspappír í sama stóra, tveggja blaða broti og orðabókin 1650. Útgáfunni fylgir athugasemd ritstjóra („Note on this Edition"). Þar er gerð grein fyrir verkaskiptingu ritstjóranna. Anthony Faulkes samdi inngang og sá um afritun og vinnslu handrits þess sem hér er gefið út í fyrsta sinn, DG 55. Hann er höfundur að meginparti skýringa og ábyrgur fyrir hönnun útgáfunnar allrar. Gunnlaugur Ingólfsson útbjó íslenska orðalistann í lok verksins og lagði einnig sitt af mörkum við skýringarnar. Í formálanum er líka gerð grein fyrir táknkerfi því sem notað er í útgáfunni.

Útgáfan hefst á inngangi („Introduction") í tveim köflum: „The History of the Glossary" og „The Sources". Þessum köflum fylgir síðan yfirlit yfir heimildir og skammstafanir: „Bibliographical references and abbreviations" og innganginum lýkur með „Index of manuscripts". Þá tekur við hin ljósritaða útgáfa af SLR jafnhliða samsvarandi efni í „Glossarium priscæ lingvæ danicæ" (DG 55). Að þessum kafla loknum er kaflinn „Notes", sem geymir athugasemdir Faulkes, fyrst við heimildaskrána í SLR: „Syllabus Autorum, qvorum in hoc lexico testimonia citantur" og þar næst við flettiorðin í verkunum tveimur. Þá fylgir stafrétt útgáfa á handritinu DG 55 og verkinu lýkur með skrá yfir íslensku orðin sem koma fyrir í verkunum báðum. Útgáfan fyrir utan innganginn nær yfir 492 blaðsíður.

Strax við efnisyfirlitið vakna tvær spurningar sem einungis er svar-

að óbeint í útgáfunni. Í fyrsta lagi: Hvert er eiginlega aðalmarkmið útgáfunnar? og í öðru lagi: Við sögu hvaða „Glossary“ er átt í inngangskaflanum?

## 3 Inngangur ritstjóra

Inngangi Anthony Faulkes (s. vii–xlvi) er skipt í tvo kafla: annars vegar er rakinn aðdragandi *Specimen lexici runici* og tengsl ritsins við DG 55 og önnur samtímaverk hins vegar er gerð grein fyrir heimildum að dæmum og skýringum sem koma fyrir í orðabókinni og handritinu.

### 3.1 „The Glossary“: Uppruni og saga

Vitað er að danski vísindamaðurinn og fornfræðingurinn Ole Worm átti frumkvæðið að útgáfu orðabókarinnar SLR, sem eignuð er Magnúsi Ólafssyni, presti í Laufási í Eyjafirði. Viðfangsefni orðabókarinnar og hlutverk Magnúsar kemur fram í titli orðabókarinnar: „Specimen lexici runici, obscuriorum qvarundam vocum, qvæ in priscis occurrunt Historiis & Poëtis Danicis, enodationem exhibens. Collectum à Dn. Magno Olavio Pastore Laufasiensi in Islandia doctissimo.“

Skýrt er frá hvernig Worm fór þess á leit við séra Magnús 1635 að hann tæki saman lista yfir orðaforðann í fornum íslenskum kveðskap (*Specimen* 2010:viii). Þrátt fyrir áhuga og þekkingu í fornum fræðum var íslenskukunnáttu Ole Worms vægast sagt ábótavant. Telur Faulkes að Magnús hafi þegar verið byrjaður á verki af svipuðu tagi því þegar hann lést 1636 var orðalistinn langt kominn. Svo virðist sem fóstursonur hans og eftirmaður í Laufási, Jón Magnússon, hafi bætt í uppkast Magnúsar og sent Ole Worm, án þess þó að ljúka verkinu að fullu. Afrit Jóns og uppkast Magnúsar eru nú glötuð.

Þar með hefst leitin að „The Glossary“. Eins og Faulkes skýrir frá styðst SLR (1650) ekki eingöngu við verk prestanna. Í háskólabókasafninu í Uppsölum er varðveitt handritið DG 55, „Glossarivm Priscæ Lingvæ Danicæ collectum a Magno Olai Jslando Pastore Lavfasiensi; anno M. DC. XXXVI“. Handritið geymir afrit af verkum þeirra Magnúsar og Jóns. Gerð er grein fyrir uppruna þess og tengslum bæði við hin glötuðu handrit Laufásprestanna og við SLR eins og hún birtist 1650. DG 55 er ýtarlega lýst. Gætir viss misræmis í efni SLR og

handritsins bæði hvað varðar einstaka orð og tilvitnanir í heimildir. Ritstjóri leitast ákaft við að komast að raun um hvaðan þetta efni sé sótt og hvenær því hafi verið bætt við. Sama á við um efni í SLR sem ekki er í DG 55.

Fleiri komu við sögu undirbúnings SLR en Laufásprestarnir tveir og upphafsmaður DG 55, sem talinn er vera Stephan J. Stephanius. Þar á meðal er, eins og Faulkes hefur áður sýnt fram á, Guðmundur Andrésson, höfundur *Lexicon islandicum* sem gefin var út 1683 (og 1999 sem *Orðfræðirit fyrri alda* IV). Bent er á samnýtingu á efni í þessum verkum og rök eru leidd að því að Guðmundur hafi aukið við efni í SLR og jafnframt notað það efni og jafnvel annað efni úr SLR í handrit sitt að *Lexicon islandicum*. Þar fyrir utan gerir Faulkes ráð fyrir að Worm sjálfur hafi bætt við eða látið bæta við efni í orðabókina, meðal annars úr skrifum Brynjólfs Sveinssonar biskups (*Specimen* 2010:xv).

Kaflanum lýkur með að gerð er stuttlega grein fyrir mikilvægi orðabókarinnar SLR. Hún er fyrsta orðabókin sem birt var yfir íslenska tungu. Þótt það sé fyrst og fremst fornmálið sem fjallað er um er þar líka talsvert af 17. aldar íslensku að finna. Fyrir utan að vera þýðingarmikil sem vitnisburður um heimildir að miðaldatextum, sem nú eru sumar glataðar, er hún einnig áhugaverð frá orðabókasögulegu sjónarmiði (*Specimen* 2010:xxi f.). Faulkes greinir í stuttu máli frá orðabókagerð á 17. öld án þess þó að tengja orðabókina hinni vaknandi norrænu orðabókahefð né heldur hinni miklu starfsemi sem þá var í Skandinavíu varðandi lestur, túlkun og útgáfu á íslenskum miðaldaverkum. Það er ekki orðabókafræði sem er sérgrein Faulkes.

## 3.2 Heimildir orða og dæma

Heimildir þær sem getið er í SLR eru raktar. Hér gerir Faulkes grein fyrir fjölda tilvitnana í miðaldaritin. Hann nefnir 270 tilvitnanir í Grettlu, rúmlega 100 í Eglu o.s.frv. Einnig eru tilgreind dæmi úr orðabókinni þar sem engra heimilda er getið og skorið er úr um uppruna þeirra í handritunum. Ekki er eingöngu látið nægja að geta þess í hvaða rit íslenska efnið sé sótt, því leitast er við að finna handrit það sem notast hefur verið við. Þar með er bætt nýrri vitneskju við það sem nú er þekkt um sögu einstakra handrita. Þar sem sum þessara handrita eru nú glötuð eru tilvitnanirnar í þau ótvíræður vottur um að þau hafi verið til á 17. öld. Faulkes telur til dæmis að handrit það

að Egils sögu sem notað var við SLR geti kastað nýju ljósi á stemma Möðruvallabókar.

Hvað varðar íslenskar miðaldabókmenntir, fílólógíu og handritafræði er Faulkes á heimavelli. Yfirlit hans um heimildir efnisins í SLR og DG 55 er mikilvægt framlag bæði hvað varðar viðhorf fornfræðinga 17. aldarinnar til íslenskra fræða og aðgang þeirra að handritum. Alkunn er sú meinloka Worms að fornmálið norræna hafi í elstu heimildum verið ritað með rúnum. Í bréfaskiptunum við Magnús Ólafsson ítrekar Worm fyrirspurnir sínar um forna rúnatexta. Magnús harmar að hann kunni ekki að lesa rúnaskrift, að hann eigi engar bækur skráðar með þannig letri og hann hafi reyndar aldrei séð þannig texta (Breve 1965:130). Öll bréfaskipti Ole Worms voru þýdd á dönsku og gefin út 1965–68. Þar má ef til vill finna heimildir um orðabókaverkefni hans sem ekki koma fram í útgáfu Jakobs Benediktssonar 1948 á bréfum hans, en við þá útgáfu styðst Faulkes (*Specimen* 2010:vii).

## 4 Útgáfa ritanna tveggja

Hinni nýju útgáfu orðabókarinnar SLR er þannig hagað að á hverri opnu er birt ljósrituð síða úr SLR (1650) jafnhliða endurritun af samsvarandi efni úr handritinu DG 55 (bls. 1–299).

Misjafnt er hversu vel ritin samsvara hvort öðru. Ljósritið er haft til viðmiðunar þannig að efni úr handritinu er svo hagað að þar sem orð vantar miðað við flettiorðin í SLR er höfð tóm lína og við orð sem fyrirfinnast í handritinu en ekki í SLR er tákn, [†] (daggarður), á spássíu. Sem dæmi má taka blaðsíðu 50 í SLR: (*hækilbiugur–happ*). Af þeim 16 flettiorðum sem eru á síðunni í SLR eru 15 líka í DG 55. Nokkur munur er þó á orðaforðanum: orðið *hæll* kemur fyrir í tveimur merkingum í SLR en vantar alveg í DG 55. Þar er aftur á móti orðið *handganga* sem ekki er í SLR. Hér er endurritun handritsins látin fylgja hinum 144 ljósprentuðu blaðsíðum orðabókarinnar.

### 4.1 Athugasemdir ritstjóra

Aftan við orðabókatextann fylgja athugasemdir við hvert einasta flettiorð sem kemur fyrir í verkunum tveimur: SLR og DG 55 (bls.

301–386). Þar er latneskum skýringum verkanna sleppt en allt íslenskt efni endurtekið og skýrt frá hvaðan hvert orð og hver merking sem getið er sé sótt. Sama á við um dæmin og orðasamböndin. Gerð hefur verið grein fyrir heimildunum í inngangi ritstjóra (sjá hér að ofan), hverri fyrir sig. Hér eru það flettiorðin í stafrófsröð sem mynda inngang í heimildirnar. Orð sem ekki eru í SLR eru einnig hér merkt með daggarði og orð sem einungis er þar að finna eru merkt með stjörnu.

## 4.2 Stafrétt útgáfa „Glossarium priscæ linguæ Danicæ, collectum a Magno Olai Jslando, Pastore Lavfasiensi; Anno M. DC. XXXVI.“

Handritið DG 55 er hér gefið út eins og það liggur fyrir (bls. 387–468). Textanum er ekki skipt í dálka heldur eru fletturnar látnar ná yfir síðuna. Flettiorðin og allt íslenskt og danskt efni er skáletrað en latneskt efni prentað með beinu letri. Síður í handritinu eru gefnar upp á spássíu. Í stafköflunum G–S eru flest öll flettiorðin skráð á sama hátt og í SLR, það er að segja fyrst með rúnaletir og síðan með latnesku letri. Víða er íslenskt efni þýtt á dönsku. Athugasemdir og leiðréttingar fylgja neðanmáls.

## 4.3 Orðalisti

Verkinu lýkur með skrá í þremur dálkum yfir íslensku orðin í *Specimen lexici Runici* (bls. 469–492). Orðin eru færð í nútímabúning og vísað er á blaðsíðu í SLR og í flettu eða flettur þar sem þau koma fyrir. Einnig er orðflokkur orðsins gefinn upp. Mörg þeirra orða sem eingöngu koma fyrir í DG 55 eru einnig í listanum en þó ekki öll. Óljóst er hvað ræður: *prýðilega* (prijdilega) og *sæng* vantar en *auðmaður* og *auðmenni* eru með.

# 5 Frágangur

Hér er að mörgu leyti um myndarlegt verk að ræða. Efnið er þó ekki

sérlega aðgengilegt enda er tilgangur útgáfunnar að vissu leyti óljós. Viðfangsefnið virðist í rauninni vera hvorugt ritanna, SLR eða DG 55, fyrir sig heldur virðist með „the Glossary“ vera átt við mengi heimilda og efnis sem á einhvern hátt tengist orðabókinni. Útgáfa orðabókar í ritröðinni *Orðfræðirit fyrri alda* gæti í sjálfu sér verið vísbending um að efnið sé unnið frá orðfræðilegum sjónarhóli. Svo er þó ekki í þessu tilfelli, því sem orðfræðilegar heimildir eru verkin látin tala fyrir sig sjálf. Í staðinn liggur áherslan á hinu textafræðilega og fílólógíska sjónarhorni. Horft er aftur í tímann, á miðaldamálið, og ekki fram á við á upphaf orðfræðilegra lýsinga á íslensku.

Um gæði verksins sem textafræðilegs og fílólógísks framlags er ég lítt dómbær, og lofsyrði á veikum grunni eru kannski ekki mikils virði. Þrátt fyrir það tel ég hér mikið verk unnið við að fylla í eyður í vitneskju okkar um orðaforðann í íslenskum fornbókmenntum og einnig um hvaða miðaldahandrit voru fyrir hendi á fyrri hluta 17. aldar. Þar að auki er mikill kostur að nú skuli þessi verk vera handbær þeim sem áhuga hafa á fyrstu orðabókinni um íslenska tungu.

## 6 Viðauki: um *Specimen lexici runici* 1650

Eins og gerð er grein fyrir í inngangi útgáfunnar er *Specimen lexici runici* að mörgu leyti dæmigerð fyrir orðabækur 17. aldarinnar. Þetta varðar einkum sundurleitar upplýsingar í flettugreinunum og ófullkomna prentlist. Orðabókin er þó að sumu leyti sérstæð. Auk þess að vera fyrsta orðabókin sem fjallar um íslenskt mál er hún fremst í flokki orðabóka á Norðurlöndum um að skýra frá málfræðilegum eiginleikum flettiorðsins. Nafnorðum, lýsingarorðum og sögnum fylgja oft, en þó ekki alltaf, upplýsingar um orðflokk og oft er nánar getið um orðmyndina ef hún er önnur en grunnmyndin. Einnig kemur fyrir að tiltekin beygingarmynd orðsins sé gefin upp og rædd. Orðaforðinn er að miklu leyti nafnorð og er kyn þeirra tilgreint. Nokkuð er um lýsingarorð en sagnorð og smáorð eru tiltölulega fá. Orðabókin geymir þannig upplýsingar sem eru áhugaverðar fyrir þann sem vill kynna sér fyrstu tilraunir til málfræðilegrar lýsingar íslenskra orða.

Vitað er að þrátt fyrir áhuga sinn á íslenskum fornritum hafði Ole Worm litla eða enga kunnáttu í íslensku. Þessi danski læknir og vísindamaður var einn atkvæðamesti fornfræðingur Norðurlanda á fyrri hluta 17. aldar og hann var frumkvöðull útgáfunnar 1650. Hann

safnaði viðamiklum upplýsingum um rúnasteina og aðrar norrænar fornleifar og gaf 1643 út sex binda rit með yfirliti um forn, norræn minnismerki: *Danicorum monumentorum libri sex*. Í *Specimen lexici runici* sameinast hið almenna álit lærðra skandinava samtíðarinnar að íslenskan væri lykillinn að sameiginlegri arfleifð frá gullöld norrænnar menningar og sú fullvissa Ole Worms að elstu norræn rit hefðu upprunalega verið skráð með rúnaskrift. Á þessum tímum endurreisnar var litið á íslenskar miðaldabókmenntir sem vitnisburð um glæsilega menningu norðursins; menningu sem hvergi gaf hinni klassísku evrópsku menningu eftir. Vandinn var þó sá að málið í þessum óþrjótandi uppsprettum fróðleiks var torskilið fyrir skandinavíska fræðimenn.

Það er því ekki íslenska samtíðarinnar sem lýsingin í SLR á við. Eins og kemur fram í bréfi Ole Worms til Magnúsar Ólafssonar í Laufási, þar sem hann lætur fyrst í ljós óskir um orðalista, eru það gömul orð og orðtök sem hann hefur í huga:

> Gid der var en, som kunde opstille en Ordfortegnelse over gamle og digteriske Ord og Talemaader; det vilde være et meget nyttigt Arbejde, og højst nødvendigt for at frelse det gamle Sprog fra Undergang. (Breve 1965:343)

Þegar SLR kom út 1650 var hún ekki fyrst og fremst til þess fallin, eins og orðabókum yfirleitt er ætlað, að greiða fyrir samskiptum manna sem hafa takmarkaða kunnáttu í tungumáli hvor annars. Sem slík mun hún ekki hafa komið að miklu gagni. Þetta á hún sameiginlegt með fyrstu íslensku orðabókunum yfirleitt, þeim var ætlað að koma til móts við þarfir „erlendra fræðimanna sem áhuga höfðu á íslenskum fornritum“ (Guðrún Kvaran 2009:54, sjá einnig Anna Helga Hannesdóttir 2004:105).

Orðabókin á sér hliðstæðu í sænskri orðabókasögu 17. aldar. Í upphafi aldarinnar var Johannes Bureus þjóðminjavörður og þjóðarbókavörður Svía. Hann vann að fornrannsóknum á svipaðan hátt og Worm og fylgdust þeir með verkum hvor annars. Bureus bókfesti rúnaristur og fornan orðaforða og gaf einnig út miðaldatexta. Seinna á öldinni lét Olof Verelius gefa út íslenskar miðaldasögur á íslensku, latínu og sænsku. Einnig vann hann að orðabókinni *Index lingvae veteris scythoscandicae sive gothicae* sem kom út 1691, níu árum eftir dauða hans. Hann var, eins og Worm, fullviss um hlutverk íslenskunnar sem sam-

nefnara hinna norrænu mála og að íslenska og sænska væru eitt og sama tungumál. Í orðabók hans ægir saman orðum úr íslenskum og sænskum miðaldaheimildum í þeim tilgangi að sýna fram á gullna fortíð norrænnar, einkum sænskrar, menningar.

Orðabókahefð þeirri sem þróast út frá hinni íslensku endurreisn í Danmörku og Svíþjóð á 17. öld hefur verið lýst sem „patriotisk lexikografi“ (Ralph 2001:302 ff.). Orðabækurnar voru liður í útgáfu norrænna gullaldarbókmennta og þeim var ætlað að „säkerställa ett gammalt fosterländskt ordförråds fortbestånd, så att det inte skulle gå förlorat sedan det räddats till eftervärlden genom återupptäckten av den isländska litteraturen“ (Ralph 2001:307). Hér á talsvert eftir að rannsaka og því mikill fengur í hinni nýju útgáfu fyrir þann sem gerir fyrstu íslensku orðabækurnar að viðfangsefni nýrra rannsókna.

## Heimildir

Anna Helga Hannesdóttir. 2004. Ordboken som språklig mötesplats. Í: *Språkhistoria och flerspråkighet*. Föredragen vid ett internationellt symposium i Uppsala 17–19 januari 2003. Utg. Lennart Elmevik. Acta academiae regiae Gustavi Adolphi LXXXVII, bls. 103–114.

Breve = *Breve fra og til Ole Worm* I. 1965. Oversat af H. D. Schepelern. Udg. af Det Danske Sprog og Litteraturselskab. København: Munksgaard.

Guðrún Kvaran. 2009. Enginn lifir orðalaust. Fáein atriði úr sögu íslensks orðaforða. Í: *Orð og tunga* 11: 45-63.

Orðfræðirit [fyrri alda] IV. = *Lexicon islandicum. Orðabók Guðmundar Andréssonar*. 1999. Útg.: Gunnlaugur Ingólfsson og Jakob Benediktsson. Orðfræðirit fyrri alda IV. Reykjavík: Orðabók Háskólans.

Ralph, Bo. 2001. Orden i ordning. Den historiska framväxten av en lexikografisk tradition i Sverige. Í: *Nordiska studier i lexikografi* 5. Rapport från Konferens om lexikografi i Norden, Göteborg 26–29 maj 1999. Red. Gellerstam et al. Göteborgs universitet: Meijerbergs arkiv för svensk ordforskning, bls. 282–322.

*Anna Helga Hannesdóttir*
*Institutionen för svenska språket*
*Göteborgs universitet*
*anna.hannesdottir@svenska.gu.se*

Ásgrímur Angantýsson

# Handbók um íslensku

*Handbók um íslensku*. Hagnýtur leiðarvísir um íslenskt mál, málnotkun, stafsetningu og ritun. Ritstjóri Jóhannes B. Sigtryggsson. JPV útgáfa, Reykjavík. 2011. (401 bls.) ISBN 978-9935-11-172-2.

## 1 Inngangur

Stofnun Árna Magnússonar í íslenskum fræðum stendur að útgáfu *Handbókar um íslensku* (HUÍ) með stuðningi Þjóðhátíðarsjóðs. Henni er „ætlað að vera handhægt rit þar sem hægt er að leita svara við margs konar spurningum sem vakna við ritun" og lögð er áhersla á „þau atriði í málnotkun, stafsetningu og ritun sem fólk á helst erfitt með" (bls. 5). Aftan á bókarkápu er enn fremur sagt að HUÍ sé „aðgengilegt uppsláttarrit". Í reglugerð um Stofnun Árna Magnússonar í íslenskum fræðum (861/2008) segir: „Stofnunin miðlar þekkingu á íslenskri tungu. Málfarsráðgjöf og leiðbeiningar hennar miða að eflingu og varðveislu íslenskrar tungu í ræðu og riti og skulu byggðar á fræðilegum grundvelli." Markhópurinn er samkvæmt formála sá sístækkandi hópur sem fæst við ýmiss konar skrif og markaðssetning bókarinnar bendir til þess að hún sé ætluð mjög breiðum hópi lesenda. Þar sem leiðbeiningar um málnotkun er víða að finna í kennslubókum og handbókum hlýtur hugmyndin með þessari útgáfu að vera sú að draga saman það markverðasta í þeim efnum, setja það skilmerkilega fram og byggja umfjöllunina á rannsóknum eins og kostur er.

Markmiðið með þessum ritdómi er annars vegar að leggja mat á hvernig til hefur tekist með útgáfu *Handbókar um íslensku* og benda á

Orð og tunga 14 (2012), 77–82. 

leiðir til að gera betur ef hún verður gefin út aftur og hins vegar að leiðbeina lesendum um notkun bókarinnar. Meginniðurstaðan er sú að ritið sé skref í átt að vönduðu yfirlitsverki á þessu sviði en þarfnist engu að síður rækilegrar endurskoðunar. Annar kafli ritdómsins fjallar um skipulag og frágang og þar er því haldið fram að stokka þurfi upp efnisskipan bókarinnar. Þriðji kafli fjallar um efnistök og meðal niðurstaðna þar er að umfjöllun um ýmsar mikilvægar málnýjungar verði útundan. Í fjórða kafla eru niðurstöður dregnar saman.

## 2 Skipulag og frágangur

*Handbók um íslensku* er einkar glæsileg í útliti. Kápan er falleg og textinn vel frágenginn. Efnisskipan orkar hins vegar tvímælis og bitnar því miður á notagildi bókarinnar. Ekki virðist hafa verið tekin ákvörðun um hvort bókin ætti að vera uppflettirit, greinasafn eða blanda af hvoru tveggja þar sem skilin væru þó skýr. Fyrir vikið er efninu raðað niður á nokkuð handahófskenndan hátt og hætt er við að margvíslegur fróðleikur fari fyrir ofan garð og neðan. Þetta er bagalegt því að handbækur verða að vera aðgengilegar og notendavænar.

Meginmál bókarinnar skiptist í tvo hluta. Sá fyrri heitir *Málnotkun, stafsetning og ritun* (bls. 9–282) en sá síðari *Um íslenskt mál* (bls. 283–378) og er köflum raðað í stafrófsröð í hvorum hluta fyrir sig. Auk þess er formáli (bls. 5–8), heimildaskrá (bls. 379–387) og atriðisorðaskrá (bls. 388–401). Efnisyfirlit meginmáls er innan á bókarkápu fremst og aftast en þar hefði nauðsynlega þurft að vekja athygli á atriðisorðaskránni því að hún er gagnlegur lykill að efni bókarinnar. Fyrri hlutinn virðist eiga að endurspegla „það sem fólk á helst erfitt með“ og í síðari hlutanum eru „fróðleikskaflar um ýmis svið tungunnar“ eins og það er orðað í formálanum. Sumir kaflar í fyrri hlutanum gætu þó vel staðið sem sjálfstæðir leskaflar, t.d. kaflarnir *Gott mál* og *Leiðbeiningar um gott mál: hóflega formlegt ritað mál*. Þar eru líka sérstakir kaflar um suma orðflokka en ekki aðra, t.d. langur kafli um fornöfn en enginn sambærilegur kafli um nafnorð. Raunar má velta fyrir sér hvort málnotendur eigi yfirhöfuð „erfitt með“ tiltekna orðflokka.

Í fyrri hlutanum ægir saman mjög almennum og þröngum kaflaheitum. Annars vegar eru t.d. kaflar sem heita *Beygingar*, *Fornöfn*, *Gott mál* og *Réttritun og uppruni orða* og hins vegar *Spurningarmerki*, j *inni í orði* og *Strik og bönd*. Hér hefði verið heillavænlegra að hafa annaðhvort flettur í stafrófsröð með markvissum skýringum og dæm-

um eða vel skipulagða inngangs- og leskafla með yfirheitum á borð við *Stafsetning og greinarmerki*, *Mismunandi tegundir ritsmíða* og *Vandmeðfarin málfarsatriði* og skipta þeim svo í undirkafla með lýsandi heitum; *Stór stafur og lítill*, *j inni í orði*; *Bréf*, *Umsóknir*; *Vandmeðfarin orðasambönd* o.s.frv. Atriðisorðaskráin á þó að auðvelda lesandanum að fletta upp atriðum sem vefjast fyrir honum en það er óhentugt að þurfa að leita mikið þar. Kaflinn *Leiðbeiningar um gott mál: hóflega formlegt ritað mál* týnist hálfpartinn í uppflettihlutanum. Betur færi á að hafa hann sem eins konar inngang að bókinni því að þar er fjallað um mikilvægar forsendur málfarsleiðbeininga af því tagi sem gefnar eru í bókinni.

Í seinni hlutanum er ýmislegt nýtt og fróðlegt en efnisvalið er nokkuð sundurleitt og ekki alltaf ljóst hvaða tilgangi það þjónar í riti af þessu tagi. Hugtakalisti í málfræði (bls. 285–291) missir t.d. marks vegna þess að dæmin vantar og *Stafsetningarorðabókin* stendur varla undir sérstökum kafla (bls. 365–366). Sú ráðstöfun að skrá heimildir í lok einstakra kafla og svo aftur í lok bókar er eflaust til þæginda fyrir lesandann eins og efnið liggur fyrir en sennilega væri það óþarft ef því væri skipað niður á markvissari hátt, t.d. í flettur og leskafla (sbr. *Handbók um málfræði* eftir Höskuld Þráinsson).

Hugsanleg leið til þess að greina betur á milli þess sem er nýtt í þessari bók og þess sem er tekið saman úr ýmsum handbókum og hjálpargögnum um íslenskt mál væri að hafa fróðleik almenns efnis í fyrri hlutanum og frumsamdar greinar og ítarefni í seinni hlutanum. Með endurskipulagningu af því tagi mætti gera ritið mun aðgengilegra og einnig auka gildi þess fyrir þá sem eru kunnugir öðrum ritum á þessu sviði.

## 3 Efnistök

Í *Handbók um íslensku* er víða notað orðalag á borð við *Ekki er rétt að segja X heldur á að nota Y* (t.d. í tengslum við beygingu frændsemisorða á bls. 20). Hér er að óþörfu fjallað um algeng máltilbrigði sem málvillur. Nærtækara væri að segja að í formlegu máli sé mælt með X frekar en Y þótt Y sé oft notað í daglegu máli. Leiðbeiningar af þessu tagi stangast raunar á við umfjöllunina um hugtökin rétt mál, rangt mál, gott mál, vont mál, málsnið o.fl. (bls. 79–82). Ef forsendur þess kafla lægju til grundvallar allri bókinni og kæmu skýrt fram í inngangi væri minni hætta á ósamræmi af þessu tagi.

Í HUÍ er umræða um „þágufallssýki“, notkun dvalarhorfs/framvinduhorfs og nýju þolmyndina/formgerðina vægast sagt snubbótt og ekkert vitnað í heimildir um efnið þótt full ástæða sé til. Allt eru þetta formgerðir sem málvöndunarmenn hafa fett fingur út í og málfræðingar hafa fjallað talsvert um á undanförnum árum (sjá t.d. Ástu Svavarsdóttur 1982, Jóhannes Gísla Jónsson og Þórhall Eyþórsson 2003 og Höskuld Þráinsson 2005). Leiðbeiningar um meðferð gagnverkandi fornafna (*hvor/hver annan*) og deilifornafna (*sinn hvor/hver*) eru vel unnar og byggðar á traustum heimildum. Hefðbundin notkun þessara fornafna virðist þó á hröðu undanhaldi í nútímamáli (sbr. Höskuld Þráinsson 2011). Það er málpólitísk spurning hversu miklum kröftum eigi að eyða í að endurvekja tilbrigði af þessu tagi.

Málfarsleiðbeiningar í HUÍ eiga að vera byggðar á fræðilegum grunni og því er eðlileg krafa að þær séu rökstuddar. Það er að sjálfsögðu víða gert í bókinni en ekki alltaf. Fremst í meginmáli (bls. 11–12) er t.d. listi yfir nokkur algeng dæmi um forsetningarnar *að* og *af* í ýmsum samböndum en engar skýringar gefnar. Hér væri rakið að nefna tengsl við dvöl og hreyfingu og að sögnin í setningunni skipti máli (*Það **er** gaman **að** þessu* / *Margir **hafa** gaman **af** þessu*) (sbr. t.d. Jón G. Friðjónsson 2003). Ef málfarsleiðbeiningar eiga að vera til gagns verður að höfða til skilnings.

Eitt viðfangsefna HUÍ er að leiðbeina um stafsetningu og greinarmerkjasetningu. Í *Auglýsingu um íslenska stafsetningu* (nr. 132/1974 með innfelldum breytingum 261/1977) eru reglur um stóran staf og lítinn mjög óljósar á köflum, einkum hvað varðar valfrelsi í heitum stofnana (t.d. *M/menntamálaráðuneyti*). Bent hefur verið á leiðir til að túlka þær á skýrari hátt og auðvelda málnotendum að fara eftir þeim (sjá umræðu hjá Margréti Guðmundsdóttur 2000, Margréti Jónsdóttur 2006 og Höskuldi Þráinssyni 2009). Því miður er engin tilraun gerð til þess að greiða úr þessum flækjum í HUÍ og ekkert vitnað í umræðu málfræðinga um stafsetningarreglurnar. Í reglum um eitt orð og tvö, þar sem einnig er visst valfrelsi, er hins vegar stundum tekin afstaða þótt forsendurnar séu ekki ljósar (bls. 37). Í því efni er raunar ósamræmi vegna þess að á bls. 15 er mælt með rithættinum *innan bæjar* en á bls. 37 er mælt með því að láta áherslu ráða. Þá er mælt með að skrifa heiti fornsagna í aðskildum orðum (*Brennu-Njáls saga* o.s.frv.) „samkvæmt hefð“ (bls. 34). Þessi ráðlegging fer þó í bága við anda stafsetningarreglnanna og það viðmið að orðáhersla ráði.

Ítarleg umræða um málfræðihugtök er að sjálfsögðu utan ramma HUÍ en gera verður þá kröfu að skilgreiningar séu skýrt orðaðar og

í samræmi við stöðu þekkingar á íslenskri málfræði. Á bls. 72 og 73 segir: „Forsetningar verða að atviksorðum þegar ekkert fallorð fer á eftir“. Orðalag af þessu tagi er kunnuglegt úr kennslubókum um íslenska málfræði (sbr. Björn Guðfinnsson 1958:81, 84–85) en stenst ekki nánari skoðun. Nærtækara er að líta svo á að smáorð á borð við *að* séu ýmist forsetningar eða samtengingar og að það ráðist af stöðu þeirra í setningum (sjá t.d. umræðu hjá Höskuldi Þráinssyni 2005:110–113). Umræða um „fleiryrtar samtengingar“ (bls. 171–172) er sama marki brennd. Þar endurómar hugtak úr úreltri skólamálfræði því að færð hafa verið rök fyrir því að svokallaðar fleiryrtar aukatengingar séu í raun settar saman úr atviksliðum og einyrtum aukatengingum (sbr. Halldór Ármann Sigurðsson 1981). Þá segir í HUÍ að orðin *sem* og *er* séu „stundum talin til fornafna” en margir vilji þó „heldur flokka þau með samtengingum“ (bls. 44) eins og fræðilegur ágreiningur sé um málið. Skemmst er frá því að segja að fyrri greiningin byggist eingöngu á skólabóka- og latínuhefð en sú síðari á málfræðilegum rökum (sbr. Höskuld Þráinsson 1980).

Þeir sem fást við skrif á opinberum vettvangi, t.d. fjölmiðlamenn, reka sig yfirleitt fljótt á að ýmis máltilbrigði sem þeim eru töm njóta lítillar virðingar í málsamfélaginu og kjósa því yfirleitt að sneiða hjá þeim. Ef HUÍ á að leiðbeina þeim sem eru í slíkum sporum þyrfti að taka saman miklu rækilegra yfirlit um „ambögur“ ásamt skýringum og athugasemdum (sbr. Ara Pál Kristinsson 1998, Árna Böðvarsson 1992, Morgunblaðsþætti Gísla Jónssonar og Jóns G. Friðjónssonar um íslenskt mál o.s.frv.). Sennilega yrði slík samantekt þó efni í heila bók.

## 4 Lokaorð

*Handbók um íslensku* veitir svör við „margs konar spurningum sem vakna við ritun” og er leiðbeinandi um margt það „í málnotkun, stafsetningu og ritun sem fólk á helst erfitt með” (bls. 5) en hún er ekki nógu „aðgengilegt uppsláttarrit“, svo að vitnað sé í orðalag á bókarkápu, vegna þess að efnisskipan er of ómarkviss. Leiðbeiningarnar miða eflaust á sinn hátt að „eflingu og varðveislu íslenskrar tungu í ræðu og riti” eins og gert er ráð fyrir í reglum um Stofnun Árna Magnússonar í íslenskum fræðum en nokkuð vantar upp á að þær séu alltaf „byggðar á fræðilegum grundvelli“ eins og kveðið er á um í sömu reglum; a.m.k. þyrfti stundum að segja lesendum betur frá því hvernig niðurstöður eru fengnar og af hverju mælt er með

einu tilbrigði frekar en öðru. Útgáfa bókar af þessu tagi vekur líka spurningar um íslenskt staðalmál. Hvernig er það og hver ákvað að þannig skyldi það vera? Viljum við festa það í sessi? Hvaða baráttumál á að setja á oddinn? Þessar spurningar þarfnast umræðu og í endurskoðaðri útgáfu þyrfti helst að gera skýra grein fyrir þeim markmiðum og málpólitísku sjónarmiðum sem liggja að baki.

## Heimildir

Ari Páll Kristinsson. 1998. *Handbók um málfar í talmiðlum*. Reykjavík: Málvísindastofnun Háskóla Íslands.

Árni Böðvarsson. 1992. *Íslenskt málfar*. Reykjavík: Almenna bókafélagið.

Ásta Svavarsdóttir. 1982. „Þágufallssýki.“ Breytingar á fallnotkun í frumlagssæti ópersónulegra setninga. *Íslenskt mál* 4: 19–62.

Björn Guðfinnsson. 1958. *Íslenzk málfræði*. 5. útgáfa. Eiríkur Hreinn Finnbogason sá um útgáfuna. Reykjavík: Ísafold.

Halldór Ármann Sigurðsson. 1981. Fleiryrtar aukatengingar? *Íslenskt mál* 3: 59–76.

Höskuldur Þráinsson. 1980. Tilvísunarfornöfn? *Íslenskt mál* 2: 53–96.

Höskuldur Þráinsson. 1995. *Handbók um málfræði*. Reykjavík: Námsgagnastofnun.

Höskuldur Þráinsson. 2009. Um stóran og lítinn staf. Einföld hjálparregla og dæmi um gagnsemi hennar. *Íslenskt mál* 31: 133–148.

Höskuldur Þráinsson (ritstj.). 2005. *Setningar. Handbók um setningafræði.* Íslensk tunga III. Reykjavík: Almenna bókafélagið.

Höskuldur Þráinsson (ritstj). 2012. Væntanlegt. *Tilbrigði í íslenskri setningagerð. Yfirlit yfir aðferðir og helstu niðurstöður.* Reykjavík: Háskólaútgáfan.

Jóhannes Gísli Jónsson og Þórhallur Eyþórsson. 2003. Breytingar á frumlagsfalli í íslensku. *Íslenskt mál* 25: 7–40.

Jón G. Friðjónsson. 2003. Íslenskt mál, 6. þáttur. *Morgunblaðið*, 12. júlí.

Margrét Guðmundsdóttir. 2000. Af þjáningum prófarkalesara. *Íslenskt mál* 22: 151–157.

Margrét Jónsdóttir. 2006. *Stafsetningarorðabókin*. Ritdómur. *Íslenskt mál* 28: 185–203.

Sigríður Sigurjónsdóttir og Joan Maling. 2001. Það var hrint mér á leiðinni í skólann: Þolmynd eða ekki þolmynd? *Íslenskt mál* 23: 123–180.

*Ásgrímur Angantýsson*
*Hug- og félagsvísindadeild*
*Háskólinn á Akureyri*
*asgrimur@unak.is*

Veturliði G. Óskarsson

# Rit um aðkomuorð á Norðurlöndum

Ritröðin *Moderne importord i språka i Norden* I–XI (Oslo 2003–2009) er afrakstur samnefnds rannsóknarverkefnis. Aðalritstjóri er Helge Sandøy prófessor í Björgvin og var hann jafnframt í forsvari fyrir verkefninu. Út eru komin tólf bindi þegar þetta er ritað en áætlað er að bindin verði alls fimmtán. Auk þeirra verka sem hér verður fjallað um er innan tíðar von á bók eftir Jógvan í Lon Jakobsen um viðhorf í Færeyjum. Enn fremur er stefnt að því að út komi bók um viðhorf í finnskumælandi hluta Finnlands eftir Saija Tamminen. Loks er í vændum heildaryfirlit og lokaskýrsla í einu bindi eftir Helge Sandøy og Tore Kristiansen.

Meginmarkmið verkefnisins voru annars vegar þau að bera saman afdrif *aðkomuorða* (erlendra orða, tökuorða) sem borist hafa inn í Norðurlandamálin, þ.m.t. finnsku, eftir seinni heimsstyrjöld (1945) og hvernig orðin hafa aðlagast ritmáli og talmáli, og hins vegar að varpa ljósi á viðhorf málnotenda til slíkra aðkomuorða og annarra erlendra áhrifa á tungumálið (sjá heimasíðu verkefnisins, http://folk.uib.no/hnohs/Moderne importord i spraka i Norden.html). Þau tólf bindi sem út eru komin eru vel yfir 2000 bls. að lengd og hafa að geyma meira en 70 mismunandi greinar, bókakafla og sérrit. Ekki er vegur að gera grein fyrir svo miklu efni í stuttri umsögn og verður hér að mestu staldrað við það sem snýr að íslensku.

> Helge Sandøy (ritstj.). 2003. *Med 'bil' i Norden i 100 år. Ordlaging og tilpassing av utalandske ord*. Moderne importord i språka i Norden I. Oslo: Novus forlag. 153 bls. ISBN 82-7099-380-8.

Þetta fyrsta bindi ritraðarinnar er ráðstefnurit með 17 greinum um ný orð og aðkomuorð og aðlögun þeirra í norrænum málum. Um

Orð og tunga 14 (2012), 83–89. 

íslensku fjalla Guðrún Kvaran og Ásta Svavarsdóttir. Grein Guðrúnar nefnist „Typer af nye ord i islandsk“ (bls. 33–41) og fjallar höfundur þar um nýyrði í íslensku og íslensku nýyrðahefðina og gerir grein fyrir orðmyndunaraðferðum. Grein Ástu nefnist „Tilpasning af importord i islandsk“ (bls. 75–81) og fjallar hún þar almennt um aðlögun aðkomuorða í íslensku. Báðar greinarnar gera í stuttu máli prýðilega grein fyrir helstu grundvallaratriðum í nýyrðasmíð og viðtöku erlendra orða í íslensku.

> Helge Sandøy og Jan-Ola Östmann (ritstj.). 2004. *"Det främmande" i nordisk språkpolitik. Om normering av utländska ord*. Moderne importord i språka i Norden II. Oslo: Novus forlag. 275 bls. ISBN 82-7099-395-6.

Átta greinar um opinber eða ríkjandi málpólitísk viðhorf til erlendra áhrifa á Norðurlandamálin. Greinarnar voru samdar að tilhlutan norrænu málnefndanna. Um íslensku fjallar Ari Páll Kristinsson í kaflanum „Offisiell normering av importord i islandsk“ (bls. 30–70). Kaflinn er greinargott yfirlit um helstu þætti sem lúta að ríkjandi viðhorfum gagnvart aðkomuorðum í íslensku á 150 ára tímabili, 1850–2000, með yfirliti um forsöguna fram á miðja 19. öld. Höfundur fjallar almennt um nýyrði og nýyrðastefnuna, um aðlögun aðkomuorða og rithátt þeirra, um starf orðanefnda og um orðabækur og þau viðhorf til orða af erlendum uppruna sem fram koma í þeim.

> Bente Selback og Helge Sandøy (ritstj.). 2007. *Fire dagar i nordiske aviser. Ei jamføring av påverknaden i ordforrådet i sju språksamfunn*. Moderne importord i språka i Norden III. Oslo: Novus forlag. 173 bls. ISBN 978-82-7099-472-4.

Tíu bókarkaflar um aðkomuorð í dagblöðum á Norðurlöndum eftir Bente Selback. Um íslensku er fjallað á bls. 25–36. Meginniðurstöðurnar eru þær að aðkomuorð eru mun færri en í dagblöðum annars staðar á Norðurlöndum. Þeim hefur þó fjölgað talsvert frá 1975 til 2000, að einu sviði undanteknu, sem eru íþróttafréttir en þar fækkaði aðkomuorðum um 80% á milli matsáranna. Umtalsvert færri aðkomuorð er að finna í ritstjórnarefni en í auglýsingum. Flest aðkomuorðin má rekja til ensku. Til grundvallar umfjölluninni liggja textar eins tölublaðs fimm íslenskra dagblaða frá árinu 1975 og þriggja frá árinu 2000. Elín Bára Magnúsdóttir orðtók textana og er textasafnið nálega 300.000 orð.

Tore Kristiansen og Lars S. Vikør (ritstj.). 2006. *Nordiske språkholdningar. Ei meiningsmåling*. Moderne importord i språka i Norden IV. Oslo: Novus forlag. 248 bls. ISBN 978-82-7099-439-7.

Níu bókarkaflar, auk tveggja viðauka, um viðhorf til notkunar á enskum orðum í Norðurlandamálum. Efnið byggist á símakönnun sem gerð var árið 2002. Úrtakið var um 500–1000 manns í hverju landi (801 á Íslandi) og voru sömu spurningar, alls níu, lagðar fyrir alla aðspurða. Úr þeim er unnið með megindlegum aðferðum. Um íslensku fjallar Kristján Árnason í kaflanum „Island" (bls. 17–39). Meðal þess sem fram kemur er að helmingur íslensku þátttakendanna í rannsókninni notaði ensku daglega og er það umtalsvert meira en meðal aðspurðra í hinum löndunum. Íslensku þátttakendurnir voru eigi að síður neikvæðastir gagnvart notkun ensku, einkum sem vinnustaðarmáls. Yngstu þátttakendurnir voru þó mun jákvæðari gagnvart ensku en þeir eldri. Þá kemur einnig fram að hærri launum og meiri menntun fylgir í flestum tilfellum meiri notkun ensku, einkum talmáls.

Tore Kristiansen (ritstj.). 2006. *Nordiske sprogholdninger. En masketest*. Moderne importord i språka i Norden V. Oslo: Novus forlag. 183 bls. ISBN 978-82-7099-448-9.

Níu bókarkaflar um dulin eða ómeðvituð viðhorf íbúa á Norðurlöndum til enskra áhrifa. Notuð var sú aðferð sem nefnist „matched guise test" (grímupróf), þar sem þátttakandi í rannsókn metur sama mælanda tvisvar án þess að vita af því sjálfur. Um íslensku skrifar Halldóra Björt Ewen í kaflanum „Island" (bls. 33–48). Meðhöfundur er Tore Kristiansen. Þátttakendur í íslenska hlutanum voru um 350. Þeir hlustuðu á fimm hljóðupptökur með lesnum fréttatextum, sem að mestu leyti voru samhljóða, og voru beðnir að meta hve vel þeir teldu lesarana henta til að gegna starfi útvarpsfréttamanns. Skyldu þeir leggja mat á það hversu metnaðarfullir, aðlaðandi, greindir, traustvekjandi, duglegir, sjálfstæðir, áhugaverðir og afslappaðir lesararnir væru og raða þeim svo með hliðsjón af því. Í tveimur af upptökunum var sami lesari. Í annarri var komið fyrir nokkrum enskum orðum í textanum en í hinni voru á sömu stöðum höfð íslensk orð. Rannsóknin miðast við þessa tvo texta. Meginniðurstöðurnar eru þær að mat þátttakenda á lesaranum var mun jákvæðara þegar hann las „hreina" textann en þegar hann las þann sem hafði að geyma aðkomuorð. Það kom þó í ljós að grímupróf af þessu tagi er erfitt að leggja fyrir Íslendinga og Færeyinga því að raunverulegt markmið rannsóknarinnar reyndist

erfitt að fela fyrir þeim. Það átti ekki við um þátttakendur annars staðar á Norðurlöndum.

> Guðrún Kvaran (ritstj.). 2007. *Udenlandske eller hjemlige ord? En undersøgelse af sprogene i Norden*. Moderne importord i språka i Norden VI. Oslo: Novus forlag. 188 bls. ISBN 978-82-7099-474-8.

Sjö bókarkaflar sem einkum fjalla um innlend samheiti aðkomuorða („afløsningsord"), oftast nýyrði. Markmið þessa hluta rannsóknarinnar var fyrst og fremst að kanna afstöðuna á milli aðkomuorða og samsvarandi „heimaorða" með tilliti til notkunar og hlutfallslegrar tíðni. Rannsóknin er að hluta byggð á sama efni og greint var í 3. bindi ritraðarinnar, þ.e. texta dagblaða frá árunum 1975 og 2000, en að auki var leitað í gagnasöfn. Guðrún Kvaran fylgir ritinu úr hlaði með greinargerð um verkefnið (bls. 9–18) og fjallar síðan um íslenska þáttinn í kaflanum „Importord og afløsningsord i islandsk" (bls. 19–48). Að lokinni greinargerð um bakgrunn málhreinsunar og íslenskrar málstefnu og um efnisval fyrir rannsóknina og greiningu efnis fjallar höfundur um 40 orð sem valin voru sérstaklega fyrir norrænu heildarrannsóknina af fjórum efnissviðum (tölvur, matur og matvæli, boltaíþróttir og dægurtónlist ungmenna) og hvaða orð eru notuð í hverju tilviki í íslensku. Enn fremur gerir höfundur grein fyrir sérathugun á 138 orðum sem valin voru til nánari athugunar. Lokahluti bókarkaflans fjallar um þær orðmyndunaraðferðir sem viðhafðar voru við myndun orðanna af efnissviðunum fjórum.

> Pia Jarvad og Helge Sandøy (ritstj.). 2007. Stuntman *og andre importord i Norden. Om udtale og bøjning*. Moderne importord i språka i Norden VII. Oslo: Novus forlag. 221 bls. ISBN 978-82-7099-484-7.

Átta bókarkaflar um aðlögun orða af erlendum uppruna með tilliti til framburðar og beygingar í Norðurlandamálum. Um íslensku fjallar Ásta Svavarsdóttir í kaflanum „*Djúsið* eller *djúsinn*? Om tilpasning af moderne importord i islandsk talesprog" (bls. 27–51). Hún greinir þar frá rannsókn á 20 málbreytum, fimm beygingarlegum og fimmtán hljóðfræðilegum. Rannsóknin byggist á 43 viðtölum sem tekin voru árið 2004. Í viðtölunum var merkingu 50 orða lýst fyrir þátttakendum og þeir beðnir að giska á hvaða orð væri um að ræða. Þar með er komið í veg fyrir að ritháttur orðs eða framburður og notkun spyrjanda hafi áhrif á niðurstöðuna. Í flestum tilfellum giskuðu þátttakendur á það orð sem leitað var eftir. Þar var ætíð um að ræða orð sem borist hafa úr ensku í íslensku eftir 1945. Meginniðurstaðan er sú að aðkomuorð af

þessum toga laga sig yfirleitt vel að íslensku máli og á það einkanlega við um framburð þeirra. Hljóðfræðibreyturnar sem rannsakaðar voru koma langoftast fram með því gildi sem svarar til eðlilegs íslensks framburðar að tveimur undanskildum sem ekki sýna jafnskýra aðlögun. Það eru samhljóðaklasar í upphafi orða eins og *chilla* og langt/stutt sérhljóð á undan tveimur samhljóðum í orðum á borð við *roastbeef*, en hvort tveggja er þó mun oftar lagað að íslensku hljóðkerfi en ekki. Hið sama á við um beygingarlegar breytur að undanskilinni sambeygingu lýsingarorða af enskum toga, en þau eru iðulega höfð óbeygð.

> Helge Omdal og Helge Sandøy (ritstj.). 2008. *Nasjonal eller internasjonal skrivemåte? Om importord i seks nordiske samfunn*. Moderne importord i språka i Norden VIII. Oslo: Novus forlag. 187 bls. ISBN 978-82-7099-490-8.

Sjö bókarkaflar sem fjalla um aðlögun nýlegra aðkomuorða (þ.e. sem tekin hafa verið upp eftir seinni heimsstyrjöld), fyrst og fremst af engilsaxneskum uppruna, að norrænu ritmálunum. Um íslensku skrifar Ásta Svavarsdóttir í kaflanum „"Staffið er megakúl." Om tilpasning af moderne importord i islandsk skriftsprog" (bls. 21–48). Höfundur greinir þar 458 íslensk orð úr sömu athugun á dagblaðatextum og þeirri sem fjallað var um í 3. bindi ritraðarinnar, aukinni með viðbótargögnum úr gagnagrunni *Morgunblaðsins*. Af þessum orðum eru 292 úr ensku. Greiningin byggist á 45 mismunandi stafsetningar- og beygingarlegum breytum, og miðast greiningin á stafsetningaraðlögun við orð af enskum uppruna en hin beygingarlega við öll orðin sem athuguð voru. Meðal þess sem fram kemur er að langflest aðkomuorðin eru annaðhvort beygingarlega aðlögðuð eða hlutlaus með tilliti til þeirra breytna sem kannaðar voru, og einungis 2% sýna augljós merki um erlend beygingareinkenni. Öllu fleiri orð sýna einhver erlend merki í rithætti sínum en þó er nálega helmingurinn lagaður að íslenskum ritvenjum og fjórðungur í viðbót er hlutlaus með tilliti til breytna sem voru kannaðar; minna en þriðjungur orðanna er því ritaður andstætt íslenskum ritvenjum.

Síðustu fjögur rit þessarar ritraðar, sem fjallað er um hér, eru eigindlegar sérrannsóknir sem fjalla um sænsku, dönsku, íslensku og Finnlandssænsku.

Catharina Nyström Höög. 2005. *Teamwork*? *Man kan lika gärna samarbeta*! *Svenska åsikter om importord*. Moderne importord i språka i Norden IX. Oslo: Novus forlag. 190 bls. ISBN 978-82-7099-411-3.

Jacob Thøgersen. 2007. *Det er meget godt som det er ... er det ikke? En undersøgelse af danskernes holdninger til engelsk*. Moderne importord i språka i Norden X. Oslo: Novus forlag. 270 bls. ISBN 978-82-7099-479-3.

Hanna Óladóttir. 2009. Shake, sjeik *eller* mjólkurhristingur? *Islandske holdninger til engelsk språkpåvirkning*. Moderne importord i språka i Norden XI. Oslo: Novus forlag. 149 bls. ISBN 978-82-7099-567-7.

Leila Mattfolk. 2011. *Attityder till det globala i det lokala. Finlandssvenskar om importord.* Moderne importord i språka i Norden XII. Oslo: Novus forlag. 237 bls. ISBN 978-82-7099-653-7.

Í þessum fjórum ritum er dregið saman mikið efni um viðhorf Svía, Dana, Íslendinga og sænskumælandi Finna til tungumálsins og til erlendra áhrifa og aðkomuorða, einkum úr ensku. Sænska og íslenska rannsóknin byggjast á djúpviðtölum við 24 einstaklinga hvor rannsókn, sú finnska á 36 viðtölum og sú danska á viðtölum við 47 einstaklinga. Öll viðtölin voru tekin árin 2002–2003. Þátttakendur voru valdir samkvæmt hugmyndum um lífsstíl, sem er vinnulíkan ættað úr félagsfræði og markaðssálfræði, og þeim skipt í fjóra hópa með hliðsjón af því sem til einföldunar mætti kalla stjórnendur og undirmenn. Skiptingin miðast annars vegar við einkenni vinnustaðarins, þ.e. hvort um er að ræða hefðbundið framleiðslufyrirtæki eða nútímalegt þjónustufyrirtæki, og hins vegar við stöðu fólks á vinnustað, þ.e. yfirmenn og millistjórnendur andspænis undirmönnum. Talið er að þessir þættir hafi áhrif á, eða endurspegli, að einhverju leyti lífsstíl fólks og sjálfsmynd og þá um leið viðhorf þess. Í reynd eru fyrirtækin valin fyrst og síðan þátttakendur innan þeirra. Reynt var að hafa meðalaldur sem næstan 35 árum. (Sjá nánar Höög 2005:41 o.áfr., Thøgersen 2007:20 o.áfr., Hanna Óladóttir 2009:33 o.áfr., Mattfolk 2011:37 o.áfr.)

Greinargerð Hönnu Óladóttur hefst á sögulegum inngangi sem er prýðileg viðbót við umfjallanir Guðrúnar Kvaran og Ástu Svavarsdóttur í 1. bindi ritraðarinnar og Ara Páls Kristinssonar í 2. bindi. Meginkafli ritsins er greining 24 einstaklingsviðtala. Er þar fjallað um notkun ensku og viðhorf til hennar, íslenskt mál og viðhorf til

þess og notkun nýyrða og aðkomuorða. Meðal annars kemur fram að stjórnendur noti ensku heldur meira í starfi en undirmenn en lítill eða enginn munur sé á enskunotkun í frítíma viðmælenda. Viðmælendur telja að Íslendingar eigi að tala íslensku sín á milli og eru fremur neikvæðir gagnvart notkun ensku sem vinnustaðarmáls. Háskólakennsla mætti, að þeirra mati, fara fram á ensku þar sem slíkt er nauðsynlegt. Almennt eru þeir jákvæðir gagnvart aukinni áherslu á enskukennslu en lítill áhugi er fyrir því að skipta úr ensku yfir í íslensku sem notendamál í tölvukerfum sem þeir nota. Margir viðmælendanna líta á íslensku sem helsta sameiningartákn þjóðarinnar og margir þeirra eru óánægðir með fjölda aðkomuorða í málinu. Þeir eru eigi að síður fremur jákvæðir gagnvart aðlögun erlendra orða. Almennt eru þeir mjög jákvæðir gagnvart nýyrðum en nota þau minna en efni standa til. Ein helsta ástæðan er sú að nýyrðin koma of seint, notendur eru orðnir vanir erlendu orðunum sem nýyrðin eiga að leysa af hólmi.

– o –

Ritröðin sem hér var greint frá er mikið verk og einstakt í sinni röð. Með henni er skráð merkileg vitneskja um áhrif enskrar tungu á norræn mál og finnsku, um stöðu þeirra áhrifa í upphafi 21. aldar og um þróun þeirra um rúmrar hálfrar aldar skeið. Ekki verður hægt að fjalla um þetta efni á komandi árum, né almennt um áhrif heimsmálsins á önnur mál, án þess að hafa niðurstöður *Moderne importord i språka i Norden* til hliðsjónar. Ég vil óska ritstjórum og höfundum til hamingju með afraksturinn.

*Veturliði G. Óskarsson*
*Uppsalaháskóla*
*veturlidi.oskarsson@nordiska.uu.se*

# Bókafregnir

## Nýjar íslensk-erlendar orðabækur

*Íslensk-spænsk orðabók. Diccionario Islandés-español*. Ritstj.: Guðrún H. Tulinius, Margrét Jónsdóttir, Sigrún Á. Eiríksdóttir, Teodore Manrique Antón og Viola Miglio. Reykjavík: Forlagið. 2011. (xi + 701 bls.) ISBN 978-9979-53-554-6.

Á síðasta ári kom út ný íslensk-spænsk orðabók. Hún kemur í kjölfar spænsk-íslenskrar orðabókar sem kom út fyrir fáeinum árum og var unnin af sama ritstjórnarhópi (*Spænsk-íslensk orðabók*, Mál og menning 2007). Með þessari nýju bók er því lokið ákveðnu verki sem þær Guðrún H. Tulinius og Margrét Jónsdóttir gerðu m.a. grein fyrir í þessu tímariti fyrir nokkrum árum (sjá *Orð og tunga* 8 (2006), bls. 153-155).

Uppflettiorð í bókinni eru um 27 þúsund talsins eftir því sem segir á bókarkápu og þeim fylgja ríflega 13 þúsund orðasambönd og máldæmi. Hún er því á stærð við algengar skólaorðabækur rétt eins og forveri hennar enda er hún ekki síst ætluð nemendum á ýmsum skólastigum. Í samræmi við það hefur áhersla verið lögð á orðaforða daglegs máls í samtímanum, orð sem tengjast ákveðnum sérsviðum og fagorð úr tilteknum fræðigreinum. Við val á sértækum orðaforða hefur verið tekið mið af samskiptaþörfum íslensku- og spænskumælandi notenda nú á dögum, með því t.d. að leggja áherslu á orð sem tengjast ferðamennsku, sjávarútvegi og menntun, viðskiptum, lögfræði og upplýsingatækni. Í greinunum eru gefnar hefðbundnar upplýsingar um málfræði, þ.e. orðflokk og kyn íslensku uppflettiorðanna ásamt völdum beygingarmyndum eða endingum, og um málnotkun, einkum merkingarsvið og/eða málsnið. Framan við orðabókina eru, auk formála, stuttar en greinargóðar leiðbeiningar um notkun hennar og yfirlit yfir skammstafanir og styttingar.

Islex *orðabókin*. (án ártals) Íslensk-dönsk/norsk/sænsk veforðabók. Aðalritstjóri: Þórdís Úlfarsdóttir. Reykjavík/Kaupmannahöfn/Bergen /Gautaborg: Stofnun Árna Magnússonar í íslenskum fræðum, Det Danske Sprog- og Litteraturselskab, Institutt for lingvistiske, litterære og estetiske studier & Institutionen för svenska språket. Vefaðgangur: http://islex.hi.is/ (danskt, norskt og sænskt viðmót: http://islex.dk/, http://islex.no/, http://islex.se).

Islex orðabókin er veforðabók milli íslensku annars vegar og dönsku, sænsku og norsku – bæði bókmáls og nýnorsku – hins vegar. Hún var formlega opnuð á degi íslenskrar tungu þann 16. nóvember 2011 og er nú öllum aðgengileg án endurgjalds. Islex orðabókin byggist á nýjum íslenskum orðabókarstofni sem miðast fyrst og fremst við nútímamál. Hún hefur að geyma um 50 þúsund íslensk uppflettiorð með jafnheitum eða skýringum á dönsku, norsku og sænsku, mörg hver með fleiri en eina merkingu. Í orðabókinni er jafnframt mikill fjöldi orðasambanda og notkunardæma sem líka eru þýdd og talsverður fjöldi fastra orðasambanda hefur fengið stöðu sjálfstæðra uppflettiorða þannig að notendur hafa beinan leitaraðgang að þeim. Víða eru skýringarmyndir við fletturnar og framburður orða er gefinn sem hljóðdæmi. Orðflokkur uppflettiorðanna er tilgreindur svo og kyn nafnorða en beyging þeirra er gefin með beinni tengingu við sérstakt gagnasafn með beygingu íslenskra orða, *Beygingarlýsingu íslensks nútímamáls*, þar sem sjá má beygingu þeirra í heild. Mikið er líka um millivísanir milli skyldra orða. Verkinu er ætlað að svara brýnni þörf fyrir handhægar orðabækur milli íslensku og hinna norrænu málanna. Það er einkum ætlað nemendum og kennurum á öllum skólastigum, þýðendum úr og á íslensku og þeim sem þurfa að setja saman texta á einhverju norðurlandamálanna.

Islex er samstarfsverkefni rannsóknar- og háskólastofnana á Íslandi, í Danmörku, Noregi og Svíþjóð. Það er kostað of opinberum fjárveitingum í þátttökulöndunum auk styrkja úr íslenskum og norrænum sjóðum. Alls hafa um 30 sérfræðingar og þýðendur unnið að verkefninu á árunum 2006-2011. Aðalritstjóri verksins er Þórdís Úlfarsdóttir og Halldóra Jónsdóttir er verkefnisstjóri. Verkið er enn í vinnslu og notendur munu t.d. taka eftir að sums staðar eru þýðingar á öll málin enn ekki tiltækar þótt þeim fari stöðugt fjölgandi. Stefnt er að því að fleiri mál bætist við síðar og þegar er hafin vinna við færeyskan hluta Islex orðabókarinnar.

## Orðabækur handa börnum

*Barnaorðabók. Ensk-íslensk, íslensk-ensk*. Íslensk þýðing og staðfæring: Nanna Rögnvaldardóttir. Myndskreytingar: Jon Ranheimsæter. Reykjavík: Mál og menning. 2008. (187 bls.) ISBN 978-9979-3-2960-2.

*Íslensk barnaorðabók*. Ritstjórar: Ingrid Markan og Laufey Leifsdóttir. Myndir: Anna Cynthia Leplar. Reykjavík: Mál og menning. 2010. (249 bls.) ISBN 978-9979-3-3192-6.

Á undanförnum árum hafa komið út tvær orðabækur sem ætlaðar eru börnum á aldrinum 8-12 ára. Þær eru í sama broti, ámóta stórar og um margt líkar að ytri gerð, t.d. eru þær báðar ríkulega myndskreyttar. Sú eldri er tvímála orðabók milli íslensku og ensku. Hún er þýdd og staðfærð eftir danskri bók, *Min første røde ordbog* (Gyldendal 2006). Skipulag hennar er þannig að

á eftir notkunarleiðbeiningum (1 bls.) og efnisyfirliti kemur ensk-íslenskur orðalisti (u.þ.b. 60 bls.) með stuttum og einföldum orðabókagreinum. Þær eru flestar aðeins enskt uppflettiorð og íslenskt jafnheiti þess, stundum með skýringarmynd. Í einstaka greinum eru þó gefin stutt notkunardæmi til frekari skýringar, t.d. þegar um fleiri en eina merkingu er að ræða og með orðum sem ekki hafa skýra merkingartilvísun heldur tákna e.k. vensl innan setningar. Beinar málfræðilegar upplýsingar eru aftur á móti litlar sem engar. Á eftir þessum hluta fylgja u.þ.b. 30 opnur þar sem orðaforði á tilteknum merkingarsviðum er sýndur myndrænt þannig að á hverri opnu er mynd eða myndir sem tengjast tilteknu sviði eða þema og eru orð (sem að jafnaði eru einnig þýdd í orðalistanum) tengd við viðeigandi myndhluta. Sem dæmi um þemu má nefna „Nature“ (dýr, plöntur o.fl.), „The body“ (líkamshlutar), „The Living Room“ (húsgögn, húsbúnaður) og „Time“ (árstíðir, vikudagar, veður o.fl.). Síðustu opnurnar í þessum hluta tengjast tilteknum orðflokkum – sögnum og forsetningum, andstæðum, ólíkum myndum sagnarinnar *to be* og algengu orðalagi („Common Phrases“). Síðasti hluti bókarinnar (rúmlega 50 síður) geymir svo íslensk-enskan orðalista sem speglar fyrsta hlutann.

*Íslensk barnaorðabók* er ný einmála orðabók yfir íslensku. Hún er greinilega öðrum þræði ætluð til þess að þjálfa börn í notkun orðabóka og búa þau þannig undir að notfæra sér almennar orðabækur síðar. Orðabókagreinarnar eru byggðar upp og settar fram eins og tíðkast í hefðbundnum orðabókum með upplýsingum um orðflokk og beygingu uppflettiorðanna, orðskýringu sem skipt er í merkingarliði þar sem við á og notkunardæmum þar sem orðin eru sýnd í samhengi. Skammstöfunum er aftur á móti stillt í hóf, t.d. eru beygingarmyndir sýndar í heilu lagi en ekki einungis með endingum. Skýringarnar eru heilar setningar og greinilega er leitast við að hafa þær einfaldar og ljósar. Mikið er af myndum í bókinni og þær virðast bæði gegna skreytingar- og skýringarhlutverki. Framan við orðabókarhlutann er ítarlegur inngangur (10 bls.) með leiðbeiningum um notkun bókarinnar þar sem bygging flettugreinanna og einstakir hlutar þeirra eru skýrðir nákvæmlega. Þar er jafnframt formáli um bókina og skammstafanaskrá. Aftan við orðabókarhlutann eru svokallaðar þemasíður á tíu opnum. Þær eru af ýmsu tagi, t.d. er ein opna helguð málsháttum, önnur orðtökum og sú þriðja andheitapörum og er hvert og eitt skýrt með mynd. Ein opna geymir heiti á árstíðum og mánaðaheiti, þ.á m. yfirlit yfir gömul íslensk mánaðaheiti, og önnur er lögð undir norræna goðafræði. Hlutverk orðabókarinnar sem kennslugagns fær stuðning af verkefnum fyrir mismunandi aldursflokka sem nálgast má á vefsíðu útgáfunnar (sjá http://www.forlagid.is/?p=583254).

## Þrjú afmælis og minningarrit

*Nefningar*. Greinar eftir Svavar Sigmundsson gefnar út í tilefni af sjötugsafmæli hans 7. september 2009. Ritnefnd: Ari Páll Kristinsson, Guðrún

Kvaran og Hallgrímur J. Ámundason. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum. Rit 73. 2009. (xiv + 458 bls.) ISBN 978-9979-654-09-4.

*Úr fórum orðabókarmanns*. Greinasafn Ásgeirs Blöndals Magnússonar gefið út í aldarminningu hans 2. nóvember 2009. Ritnefnd: Ágústa Þorbergsdóttir, Gunnlaugur Ingólfsson og Jónína Hafsteinsdóttir. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum. Rit 76. 2010. (xii + 283 bls.) ISBN 978-9979-654-11-7.

*Fjöruskeljar*. Afmælisrit til heiðurs Jónínu Hafsteinsdóttur sjötugri 29. mars 2011. Ritnefnd: Guðrún Kvaran, Hallgrímur J. Ámundason og Svavar Sigmundsson. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum. Rit 81. 2011. (xii + 256 bls.) ISBN 978-9979-654-18-6.

Þrjú greinasöfn um orðaforða, nafnfræði og málfræði hafa nýlega verið gefin út í ritröð Stofnunar Árna Magnússonar í íslenskum fræðum í tilefni tímamóta hjá starfsmönnum stofnunarinnar eða forvera hennar. Þetta eru afmælisrit til Svavars Sigmundssonar og Jónínu Hafsteinsdóttur og rit í aldarminningu Ásgeirs Blöndals Magnússonar (1909–1987).

Í afmælisriti Jónínu, *Fjöruskeljum*, eru tuttugu og tvær greinar skrifaðar af félögum hennar og samstarfsfólki sérstaklega fyrir þetta rit. Þær fjalla allar á einn eða annan hátt um örnefni eða önnur staðarnöfn. Meðal höfunda eru fræðimenn á svið náttúruvísinda jafnt sem hug- og félagsvísinda og sjónarhornið er því talsvert fjölbreytilegt. Margar greinanna fjalla um einstök örnefni eða örnefni á tilteknum svæðum til sjávar og sveita. Þar má nefna grein Kjartans Ólafssonar um Ræningjahól og grein Valgarðs Egilssonar um Sólarfjall, grein Ævars Petersens um örnefni í Mánáreyjum, Hauks Jóhannessonar um örnefni í Kolbeinsvík og grein eftir Gunnlaug Ingólfsson um heiti á húsum og öðrum kennileitum í Kleppsholtinu í Reykjavík um miðbik 20. aldar. Aðrar greinar fjalla um tiltekna gerð eða flokka örnefna svo sem grein Hrefnu Sigríðar Bjartmarsdóttur, *Nykur í þjóðtrú og örnefnum*. Fleiri greinar koma inn á þjóðtrú í tengslum við örnefni, þ.á m. grein eftir Bjarna Harðarson, *Síðar urðu þar reimleikar*, um þjóðsögu tengda svonefndan *Stelpusteini* í Biskupstungum. Einnig má nefna greinar sem snerta örnefni sem vitnisburð um söguleg og menningarleg fyrirbæri, t.d. grein Birnu Lárusdóttur, *Þúfnabanar, kjarnorka og netabolir*, og grein Bjarna F. Einarssonar um rostunga við Ísland. Svavar Sigmundsson fjallar um aldur örnefna og Ari Páll Kristinsson skrifar grein um rithátt ríkjaheita. Í heild gefur ritið fjölbreytta og áhugaverða innsýn í örnefnafræði og vitnar um það hvernig hún snertir ólík fræðasvið.

Hin greinasöfnin tvö eiga það sammerkt að þar er birt úrval greina eftir þá menn sjálfa sem ritin eru tileinkuð og að greinarnar hafa allar birst áður í bókum eða tímaritum. Greinasafn Svavars er efnislega skylt afmælisriti Jónínu því langflestar þeirra þrjátíu og fimm greina sem þar eru prentaðar fjalla um nafnfræði og örnefnafræði. Meðal viðfangsefna eru íslensk örnefni

með tiltekin formleg einkenni (t.d. nöfn sem enda á *–tún* eða *–staðir*), örnefni af ákveðnum uppruna (t.d. staðarnöfn sprottin af mannanöfnum), staðbundin einkenni örnefna og nöfn á stjórnsýslueiningum á Íslandi. Einnig eru í ritinu yfirlitsgreinar, þ.á m. greinar sem fela í sér samanburð við örnefni annars staðar, t.d. í Skotlandi og Orkneyjum, og greinar um ákveðin fyrirbæri s.s. *Nafngiftir erlendra sjómanna á íslenskum stöðum*. Nokkrar greinar fjalla um önnur nöfn en örnefni, t.d. nöfn í skáldverkum Halldórs Laxness og íslensk ættar- og millinöfn. Þótt nafnfræði hafi augljóslega verið í brennidepli við val á greinum eru líka nokkrar greinar um önnur efni í ritinu, þ.á m. um tiltekin orð, um viðskeytið *–ari*, um hljóðdvöl í íslensku og um íslensk málfræðiheiti á 19. öld. Greinunum í bókinni er raðað í aldursröð; sú elsta var skrifuð 1968, sú yngsta 2004. Þær hafa áður birst í tímaritum, greinasöfnum, afmælisritum og ráðstefnuritum innan lands og utan. Um þriðjungur þeirra er skrifaður á íslensku en hinar eru ýmist á ensku, þýsku eða einhverju Norðurlandamálanna. Þótt safnið sé mikið að vöxtum, rúmar fjögur hundrað blaðsíður auk skráa, sýnir ritaskrá Svavars (1962–2009), sem birt er í lok bókarinnar, að það geymir einungis hluta þess sem hann hefur birt. Fyrir utan ritaskrána er stutt skammstafanaskrá og ítarleg nafnaskrá í lok bókarinnar. Heimildir fylgja hins vegar hverri einstakri grein, ýmist neðanmáls eða í sérstakri heimildaskrá, auk upplýsinga um það hvar hún birtist fyrst og stundum fylgir henni útdráttur á öðru máli en því sem hún er skrifuð á.

Greinasafn Ásgeirs Blöndals Magnússonar skiptist í tvo hluta. Sá fyrri nefnist *Orðfræði* og þar eru tólf tiltölulega stuttar greinar um orð og orðaforða. Flestar þeirra skrifaði Ásgeir sem fasta pistla í íslensk málfræðitímarit, *Íslenska tungu* á árunum 1960–1965 undir yfirskriftinni „Úr fórum Orðabókarinnar“ og í þættinum „Orð af orði“ í fyrstu árgöngum *Íslensks máls*. Auk þess eru í þessum hluta tveir mjög ítarlegir ritdómar um etýmólógískar orðabækur Holthausen (birtur 1950) og de Vries (birtur í þremur hlutum 1957–1962). Í síðari hluta safnsins, sem hefur yfirskriftina *Málfræði*, eru sex greinar frá árunum 1953–1981, hver um sig um tiltekið efni í íslenskri eða norrænni málsögu. Flestar þeirra fjalla um þróun hljóðkerfis og framburðar en þarna er einnig grein um endurtekningarsagnir með t-viðskeyti og önnur um þróun orðaforðans. Aftan við greinarnar er skammstafanaskrá, sameiginleg ritaskrá fyrir allar greinarnar og ítarleg skrá um íslensk orð sem fjallað er um í greinunum. Gunnlaugur Ingólfsson, einn ritnefndarmeðlima, skrifar inngang að greinasafninu. Þar segir hann frá ævi og störfum Ásgeirs og fjallar um greinarnar sem birtar eru í safninu en í því mun vera dregið saman megnið af því sem Ásgeir ritaði um íslenskt mál og málfræði fyrir utan höfuðrit hans, *Íslenska orðsifjabók* (1989), sem gefin var út að honum látnum.

## Ný útgáfa mannanafnabókar

> Guðrún Kvaran. *Nöfn Íslendinga*. Ný útgáfa. Reykjavík: Forlagið. 2011. (662 bls.) ISBN 978-9979-53-546-1.

Ný útgáfa af bókinni *Nöfn Íslendinga* leit dagsins ljós 2011. Hún er endurgerð bókarinnar sem fyrst var gefin út 1991 og samin af Guðrúnu Kvaran og Sigurði Jónssyni frá Arnarvatni, en Guðrún er ein skrifuð fyrir nýju útgáfunni. Samkvæmt formála höfundar hefur allt verkið verið endurskoðað og aukið talsvert en markmið þess og búningur er í stórum dráttum óbreytt. Hlutverk þess er „að safna saman á einn stað sem flestum þeirra nafna sem vitað er að Íslendingar hafa borið“ og er einkum stuðst við manntöl og þjóðskrá. Verkið hefst á ítarlegum inngangi (næstum 30 síður) þar sem fyrst er gerð grein fyrir nýju útgáfunni en síðan fjallað um heimildir um mannanöfn, gerð grein fyrir mismunandi tegundum nafna – eiginnöfnum, millinöfnum og kenninöfnum – með tilvísun til mannanafnalaga, sagt frá mannanafnanefnd og mannanafnaskrá og fjallað um það sem hefur haft áhrif á nafnaval á ýmsum tímum. Þá er fjallað um myndun eiginnafna með viðskeytum og viðliðum og um gælunöfn. Síðustu kaflar inngangsins geyma ýmsar tölulegar upplýsingar um mannanöfn og tíðni þeirra svo og um tvínefni. Meginhluti bókarinnar er í orðabókarformi. Þar er nöfnunum, þ.e.a.s. eiginnöfnum og millinöfnum, raðað í stafrófsröð og nokkur fróðleikur um hvert nafn, alls um 6000 flettur. Greinarnar eru nokkuð mislangar en í þeim er a.m.k. gerð grein fyrir formlegum einkennum hvers nafns (ritmynd(um), kyni og beygingu), sagt frá sögu þess og útbreiðslu og fjallað um uppruna þess og/eða gerð. Þau nöfn sem hafa verið samþykkt á mannanafna- eða millinafnaskrá eru merkt sérstaklega. Jafnvel þótt nöfnum hafi fjölgað mikið frá eldri útgáfu og bókin lengst um nærri 50 síður gera breytt letur og þynnri pappír það að verkum að nýja bókin er talsvert þynnri og handhægari en sú fyrri.

## Norræn rit um orðabókafræði

> Loránd-Levente Pálfi. *Leksikon over ordbøger og leksika*. 2. udgave. Under medvirken af Johnny Finnssøn Lindholm. Kaupmannahöfn: Frydenlund. 2011. (640 bls.) ISBN 87-7887-976-0.

Þetta rit gefur yfirlit yfir danskar orðabækur, orðasöfn og alfræðirit sem komið hafa út eftir 1990, hvort sem það er á bók eða í rafrænni útgáfu. Samkvæmt formála höfundar er því einkum ætlað að létta notendum upplýsingaleit með því að benda á nýleg dönsk uppsláttarrit, bæði orðabækur af öllu tagi og ýmiss konar sértæk orðasöfn, t.d. á tilteknum sérfræðisviðum. Bókin skiptist í þrjá hluta auk ítarlegs formála með notkunarleiðbeiningum og eftirmála. Fyrsti hlutinn er ætlaður til efnisleitar eftir lykilorðum og nefnist „Primær opslagsdel“. Mörg lykilorðanna eru heiti á tungumálum (t.d. ís-

lenska eða franska) og vísa þá einkum á tiltækar orðabækur milli dönsku og viðkomandi máls en meðal þeirra eru líka efnisorð (t.d. kryddjurtir, spil eða málvísindi) sem vísa yfirleitt til sértækra orðasafna um viðkomandi fyrirbæri eða á umræddu sviði. Mikið er um millivísanir á milli greina, t.d er í flettunni „islandsk" gerð ræki leg grein fyrir orðabókum milli dönsku og íslensku en þar er jafnframt vísað í allmargar greinar þar sem íslenska kemur við sögu („arbejdsmarked; arkivarbejde; bliss; botanik" o.s.frv.). Í greinunum eru verk á viðkomandi sviði ekki einungis tilgreind heldur er þeim lýst allrækilega (stærð, einkennum, útgáfuformi o.fl.). Þar er einnig að finna tilvísanir til umsagna og dóma um viðkomandi uppflettirit eftir því sem það á við. Í öðrum hluta verksins sem ber yfirskriftina „Bibliografi" er númeruð skrá yfir öll þau verk sem tilgreind eru, alls næstum 2000 titla. Þeim er raðað í stafrófsröð og í skránni eru nákvæmar bókfræðilegar upplýsingar um hvert verk. Í þessum hluta er líka skrá yfir aðrar heimildir, fyrst og fremst umsagnir og dóma um fjölda uppsláttarrita. Þriðji og síðasti hlutinn geymir nafnaskrá yfir höfunda og ritstjóra með tilvísun til númeruðu skrárinnar á undan. Þótt þetta verk snúist fyrst og fremst um dönsk uppsláttarrit er þar að finna upplýsingar um fjölda verka frá síðustu 20 árum sem snerta íslensku, bæði tvímála orðabækur milli dönsku og íslensku og ýmis sértæk rit, oft fjölmála, þar sem íslenska kemur við sögu, t.d. ýmis norræn orðasöfn og orðalista.

> *Nordiska studier i lexikografi* 10. Rapport från Konferensen om lexikografi i Norden, Tammerfors 3.-5. juni 2009. Ritstj.: Harry Lönnroth & Kristina Nikula. Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 11. Tammerfors: NFL, Tammerfors universitet och Språkrådet i Norge. 2010. (529 bls.) ISBN978-951-44-8143-7.

Þetta er ráðstefnurit frá 10. ráðstefnu norræna orðabókafræðifélagsins, NFL, sem haldin var í Tammerfors í Finnlandi í júní 2009. Þar eru birtar greinar sem unnar eru upp úr 38 af þeim 48 fyrirlestrum sem haldnir voru á ráðstefnunni, þeirra á meðal greinar eftir báða boðsfyrirlesarana á ráðstefnunni. Reinhard Hartmann frá Háskólanum í Exeter fjallar um orðabókafræði sem háskólagrein og framfarir í orðabókarannsóknum í greininni „Has Lexicography Arrived as an Academic Discipline? Reviewing Progress in Dictionary Research during the Last Three Decades" og Henning Bergenholtz frá Háskólanum í Árósum fjallar um hlutverk og eðli orðabóka í grein sinni „En ordbog er en brugergenstand, en god ordbog er indrettet til sit specielle formål". Aðrar greinar fjalla um ýmis efni sem varða orðabókagerð og orðabókafræði. Fjórir íslenskir höfundar eiga greinar í ritinu og þær má taka sem dæmi um fjölbreytileika viðfangsefnanna. Ari Páll Kristinsson fjallar um tungumálaheiti í norrænum einmála orðabókum og hvernig staða málanna getur endarspeglast í gerð orðskýringanna, Guðrún Kvaran ritar grein um ýmiss konar vandkvæði og álitamál tengd orðavali í íslenska tökuorðabók. Þá skrifa Jón Hilmar Jónsson, Anna Helga Hannesdóttir og Sofia

Tingsell saman grein þar sem þau velta fyrir sér framsetningu ýmiss konar fastra orðasambanda í rafrænni tvímála orðabók á grundvelli merkingar og hlutverks fremur en forms. Anna Helga skrifar líka grein með Bo Ralph þar sem þau fjalla um sögulega þróun í sænskri orðabókahefð og um þau mörk sem tíðkast hefur að draga milli málfræðilegrar og orðfræðilegrar lýsingar á tungumálum.

# Ellefta ráðstefnan um norræna orðabókargerð í Lundi 24.–27. maí 2011

Ráðstefna norrænna orðabókarfræðinga, *11 konferensen om lexikografi i Norden*, var haldin á vegum orðabókarritstjórnar sænsku akademíunnar, norræna orðabókafræðifélagsins (NFL) og norska málráðsins í Lundi í Svíþjóð dagana 24.– 27. maí 2011. Ráðstefnuna sóttu um 150 manns, flestir frá Norðurlöndunum en einnig frá Lettlandi, Litháen, Tékklandi, Þýskalandi, Hollandi, Belgíu, Frakklandi og Bandaríkjunum.

Á ráðstefnunni voru haldin 65 erindi í þremur málstofum og þar var fjallað um ýmsa þætti orðabókargerðar. Í megindráttum skiptist efni fyrirlestranna í umfjöllun um sögu orðabókarfræði og eldri orðabækur, um rafræna útgáfu prentaðra orðabóka eða orðabókarverka sem stofnað var til með útgáfu á bók í huga, og loks um orðabókargerð þar sem frá upphafi var stefnt að rafrænum aðgangi, þ.e. að veflægum gagnagrunnum. Undir síðasta liðinn falla fyrirlestrar um nýjan hugbúnað og aðferðir, auk kynninga á nýjum orðabókarverkum. Ráðstefnurit er væntanlegt þar sem erindin munu birtast.

Ráðstefnuna sóttu allmargir Íslendingar, flestir starfsmenn orðfræðisviðs Stofnunar Árna Magnússonar í íslenskum fræðum (SÁ). Níu erindi voru flutt sem tengdust íslenskum orðabókarverkum. Umfjöllun um Islex var þar áberandi enda var markmiðið að kynna verkið fyrir fyrirhugaða opnun þess, á degi íslenskrar tungu 16. nóvember 2011. Þessi erindi voru flutt af ritstjórum verksins frá Íslandi, Danmörku, Noregi og Svíþjóð og samstarfsfólki þeirra. Erindi tengd Islex voru þessi:

- Halldóra Jónsdóttir og Þórdís Úlfarsdóttir, SÁ: *Islex - Resultat af nordisk sprogsamarbejde.*
- Anna Helga Hannesdóttir, Háskólanum í Gautaborg, Margunn Rauset, Háskólanum í Bergen og Aldís Sigurðardóttir, DSL í Kaupmannahöfn: *En-, två- eller flerspråkig ordbok?*
- Håkan Jansson, Háskólanum í Gautaborg: *Parallellkorpusen som resurs i lexikografiskt arbete*.
- Ylva Hellerud, Háskólanum í Gautaborg: *Översättaren som lexikograf.*
- Kristín Bjarnadóttir, SÁ: *Breaking away from tradition: Linking a database of inflection to an electronic dictionary.*

Önnur íslensk erindi voru þessi:

- Guðrún Kvaran, SÁ: *Ordbogsmanuskripter og et historiskt ordbogsarbejde.*
- Gunnlaugur Ingólfsson, SÁ: *De første islandske ordbøger.*
- Jón Hilmar Jónsson, SÁ: *Adverb og adverbialer: en forsømt ordklasse i ordbøkene.*
- Helgi Haraldsson, prófessor emeritus, Háskólanum í Osló: *Islandsk-tsjekkisk /Tsjekkisk-islandsk ordbok. Orientering.*

Dvölin í Lundi var ánægjuleg, í góðum félagsskap og góðu veðrið, a.m.k. í minningunni. Þá var eftirminnilegt að sjá prúðbúna doktoranta streyma í útskrift við Háskólann í Lundi í næsta húsi við ráðstefnustaðinn, í síðkjólum og kjólfötum, við drynjandi fallbyssuskot. Andrúmsloftið við virðulegan gamlan háskóla birtist þar í öllu sínu veldi.

## Ráðstefnur sumarið 2012

### Norræn nafnfræðiráðstefna

Dagana 6.–9. júní 2012 verður fimmtánda norræna nafnfræðingaráðstefnan haldin í Askov í Danmörku undir yfirskriftinni „Navne og skel – skellet mellem navne“ (Nöfn og skil – skilin milli nafna). Nafnfræðideild norrænu rannsóknarstofnunarinnar við Kaupmannahafnarháskóla stendur að ráðstefnunni. Í kynningu á þema hennar benda skipuleggjendur á að ýmiss konar skil eða mörk, bæði hlutlæg og huglæg, komi við sögu í nafnarannsóknum og að efnið geti jafnt snert örnefnarannsóknir, mannanafnarannsóknir, rannsóknir á nöfnum í borgarumhverfi og á hvers kyns nöfnum af öðru tagi. Allar upplýsingar um ráðstefnuna má nálgast á vefsíðu hennar, http://nfi.ku.dk/navnekongres2012.

### Alþjóðleg ráðstefna um sögulega orðabókafræði og orðfræði

Dagana 25.– 28. júlí 2012 verður 6. alþjóðlega ráðstefnan um sögulega orðabókafræði og orðfræði (ICHLL) haldin í Jena í Þýskalandi. Hún er skipulögð af indóevrópskudeild Friedrich-Schiller háskólans í Jena og saxnesku vísindaakademíunni í Leipzig. Fyrirlestrar, sem fluttir verða á ensku, þýsku eða frönsku, munu fjalla um ýmsar hliðar sögulegrar orðabókafræði, rannsóknir á sögulegum orðabókum og sögulega orðfræði. Frestur til að skila útdráttum er liðinn en enn er hægt að skrá sig á ráðstefnuna. Ítarlegar upp-

lýsingar má nálgast á vefsíðunni http://www.saw-leipzig.de/news/ichll?set_language=en.

## Euralex 2012

Í sumar verður fimmtánda alþjóðlega Euralex-ráðstefnan haldin í Osló. Slíkar ráðstefnur eru haldnar á tveggja ára fresti í nafni evrópsku orðabókafræðisamtakanna Euralex (European Association for Lexicography) og að þessu sinn er Háskólinn í Osló gestgjafi. Ráðstefnan, sem fer fram dagana 7.–11. ágúst, er skipulögð af málvísinda- og norrænudeild háskólans (Institutt for lingvistiske og nordiske studier) og norska málráðinu undir forystu Ruth Vatvedt-Fjeld. Gert er ráð fyrir að ráðstefnan spanni öll helstu svið orðabókafræði en þau efni sem verða í brennidepli á ráðstefnunni að þessu sinni eru:

- Orðabókafræði og þjóðarímynd
- Frumbyggjamál og orðabókafræði
- Orðabókagerð á grundvelli málheilda
- Orðabókafræði í máltækni
- Fjölmála orðabækur
- Orðabókafræði og kenningar í merkingarfræði
- Íðorð, sérhæft mál og orðabókafræði

Auk þess er gert ráð fyrir kynningum á orðabókafræðilegum og orðfræðilegum verkefnum og einnig verður rúm fyrir önnur umræðuefni. Auk fyrirlestra verður boðið upp á veggspjaldakynningar og sýningar . Frestur til að skila tillögum að erindum eða veggspjöldum er þegar liðinn en enn er hægt að skrá sig til þátttöku í ráðstefnunni. Allar upplýsingar um hana eru á vefsíðunni http://www.hf.uio.no/iln/forskning/aktuelt/arrangementer/konferanser-seminarer/2012/euralex/ sem líka má nálgast gegnum vef Euralex, http://www.euralex.org/,

## Norræna málnefndaþingið

Dagana 30. og 31. ágúst 2012 heldur samstarfsnet norrænu málnefndanna hið árlega þing sitt í Osló. Þema þess verður textun kvikmynda og sjónvarpsefnis. Þingið er öllum opið en dagskrá þess hefur enn ekki verið birt. Frekari upplýsingar munu m.a. birtast á vefsíðu *Nordisk sprogkoordination*, http://nordisksprogkoordination.org/konferencer-1.