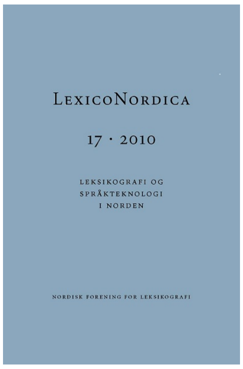

LexicoNordica

17 · 2010

leksikografi og språkteknologi i norden

nordisk forening for leksikografi

Lexicon Acquisition through Noun Clustering

Anna Björk Nikulásdóttir & Matthew Whelpton

This paper describes an experiment with clustering of Icelandic nouns based on semantic relatedness. This work is part of a larger project aiming at semi-automatically constructing a semantic database for Icelandic language technology. The harvested semantic clusters also provide valuable information for traditional lexicography.

1. Introduction

Semantic resources are already an established part of natural language processing (NLP) applications for dominant languages. Following the Princeton WordNet (Fellbaum 1998) for English, many other languages have created their own WordNet-like resources (cf. http://www.globalwordnet.org). However, for less-resourced languages like Icelandic, the situation is much less favourable. Icelandic language technology (LT) has really only existed for about a decade (Rögnvaldsson et al. 2009) and despite a rich lexicographic tradition there have until now been no specially LT-oriented semantic resources. Fortunately, over the last decade, the prerequisites for the application of (semi-)automatic methods in developing such semantic resources have now been created: a Part-of-Speech-tagger, a shallow parser and a lemmatizer (Loftsson 2008; Loftsson and Rögnvaldsson 2007; Ingason et al. 2008).

In 2007, a pilot study was run to extract semantic relations from an Icelandic dictionary (Nikulásdóttir and Whelpton 2009; Nikulásdóttir 2007a; Nikulásdóttir 2007b); following the success of this study and parallel developments in the field, a work-package for the creation of a database of semantic relations was incorporated into a major new project in Icelandic LT[1]. One central aim of the project is to experiment with known methods for the extraction of semantic relations and investigate how well they can be applied to Icelandic, given two significant factors: a) Icelandic is a highly inflected language; b) there are as yet no large corpora for the language. Most of the research in this area has focused on English which differs from Icelandic in both respects. To as great an extent as possible, we aim to exploit and develop methodologies which will be generally viable for other less-resourced languages with the support of open source tools.

The methods for the extraction of semantic relations or other semantic information can be divided into a) pattern-based methods and b) statistical methods. Pattern-based methods make use of syntactic and lexico-syntactic patterns as introduced by Hearst (1992), whereas statistical methods investigate statistical properties of language data. Following hybrid methodologies developed in recent years (Pantel and Pennachiotti 2008; Cimiano 2006; Cederberg and Widdows 2003) we intend to exploit and to combine methods from both approaches (Nikulásdóttir and Whelpton 2010).

In this paper we describe one statistical method for the extraction of semantic information, namely clustering on the basis of semantic relatedness. In sections 2 and 3 we discuss semantic relatedness and clustering in general. In section 4 we describe an experiment with the clustering of Icelandic nouns and how the results can be utilized for construction of a semantic database, as well as for lemma acquisition in traditional lexicography.

[1] In 2009, the project *Viable Language Technology beyond English – Icelandic as a test case* received a three year Grant of Excellence from the Icelandic Research Fund *RANNÍS*

2. Semantic relatedness

Relations between words or concepts in semantic databases in the style of the Princeton WordNet (Fellbaum 1998) are predominantly classical semantic relations like hypernymy and synonymy. An extension of the set of classical relations is evolving in different resources: DanNet (Pedersen et al. 2009) e.g. uses the CONCERNS relation to express a general topic-relatedness of two concepts, (*goal* CONCERNS *sport*) and Boyd-Graber et al. (2006) have conducted experiments in enriching WordNet with a directed, weighted "evoking" relation. The evoking relation describes how strongly one concept evokes another one, e.g. *car* evokes *road.*

The Swedish SALDO resource (Borin and Forsberg 2009) is not designed along the lines of WordNet, but rather uses loosely characterized associative relations as its structuring principle.

We believe that a semantic database for NLP applications could profit from such loosely characterized relations alongside the classical semantic relations. One way to harvest such relations is to use measures of *semantic similarity* or *semantic relatedness*. The definition of semantic similarity has been rather vague (Manning and Schütze 1999:296), but it is now generally accepted that semantic similarity should be distinguished from semantic relatedness, which is a more general relation. Some scholars (cf. Resnik 1995:448) have treated the two kinds of relation as orthogonal to each other: so *car* and *wheel* are related by a specific classical relation – meronymy; and yet they are not similar to each other in the way that, say, *car* and *bicycle* are. We follow Budanitsky and Hirst (2001:29) in treating semantic relatedness as an umbrella term for a range of semantic relations including not only semantic similarity (*car~bicycle*) but also "lexical relationships such as meronymy (*car-wheel*) and antonymy (*hot-cold*), or [...] any kind of functional relationship or frequent association (*pencil-paper, penguin-*

Antarctica)." A more thorough discussion of semantic similarity and semantic relatedness can be found e.g. in Zesch and Gurevych (2009) and Turney (2006).

For the automatic extraction of semantic information, pattern-based methods (cf. Hearst 1992) are commonly used for the extraction of classical semantic relations such as hypernymy and meronymy. These methods use reliable lexico-syntactic patterns, like NP_1 *such as* NP_2 ((, NP)* (and NP_n))* and extraction rules. In this example a description of the respective rule would be "if NP$_1$ is followed by *such as* and NP$_2$ (and possibly an enumeration of nominal phrases), then NP$_1$ represents a hypernym of NP$_2$ (and the other NPs, if present). Given for example the sentence: *sports such as soccer, handball and basketball* ... one would extract (*soccer* HYPERNYM *sport*), (*handball* HYPERNYM *sport*) and (*basketball* HYPERNYM *sport*).

There are two main approaches to the automatic computation of the less well-defined semantic relatedness: a) to use knowledge sources like WordNet and Wikipedia, where paths between concepts or the glosses/definitions build the basis for measures of relatedness (Zesch and Gurevych 2009; Pedersen et al. 2004; Budanitsky and Hirst 2001; Resnik 1995); and b) to apply distributional methods to text corpora (Bullinaria 2008; Cimiano 2006; Weeds 2003; Cederberg and Widdows 2003; Manning and Schütze 1999). In this paper we follow the second of these approaches and describe an experiment using the distributional method of semantic clustering, based on co-occurrences of nouns and common content-bearing words (mainly nouns, verbs, adjectives). As is described in Section 3, the notion of "co-occurrence" used here is purely collocational (i.e. the co-occurrence of the target word with common content-bearing words); however, we also note the success of Cimiano (2006) and Weeds (2003) in measuring distributional similarity with respect to grammatical functions, especially with respect to direct objects. Application of this method for Icelandic awaits future work.

Such general semantic relatedness is important because it can be used as a "confidence measure" to validate specific semantic relations extracted with other methods (such as the pattern-based methods mentioned earlier). As an example, Cederberg and Widdows (2003) use general semantic relatedness to rank their hyponymy results, extracted using lexico-syntactic patterns. In doing so, they achieved a 30% reduction in error. We have already conducted initial experiments with this method, and intend to implement it on a broader basis (Nikulásdóttir and Whelpton 2010).

3. Distributional similarity and clustering

The fundamental assumption underlying distributional methods for computing semantic relatedness is that the semantic properties of a word determine the context they appear in. Thus, words appearing in a similar context are likely to be semantically related. The basic method therefore involves compiling distributional information on a set of *target words* with respect to some uniform definition of *context*.

3.1. The list of target words

The list of target words is generated using a text corpus. At the moment a balanced PoS-tagged, lemmatized corpus, *Mörkuð íslensk málheild* (MIM), is being developed at the Árni Magnússon Institute for Icelandic Studies (Helgadóttir 2004). The planned size of this corpus is about 25 million tokens, a reasonable size but still not especially large. For our present studies we use a subset of a preliminary version of this corpus (hereafter, SubMIM) containing about 8.8 million tokens, including punctuation marks etc. The source of this data is mainly newspaper texts (Morgunblaðið, a selection from the years 2000-2007), but further texts come from a public science web portal at the University of Iceland (http://visindavefur.is), reports from Icelandic ministries, and from a medical Journal (Læknablaðið).

The tagging and lemmatizing was performed using the PoS-tagger *IceTagger* and the lemmatizer *Lemmald*, both included in the open source IceNLP-toolkit.[2]

The list of target words in our experiment is composed of those nouns which occur at least 18 times in SubMIM, approximately 11,500 nouns. To reduce noise, we removed the 100 most frequent nouns from the list, leaving 11,400 nouns. It should be noted that the original lists consisted of automatically lemmatized word forms. Incorrect lemmata were deleted but not corrected, and thus the lists do not mirror 100% the frequency relations in the corpus and the analysis does not account for wrongly lemmatized words. With the correction of the lemmatization these lists will change.

3.2. Defining context – words and windows

Semantic similarity is computed for a set of target words with respect to some uniform definition of context. There are a number of ways of defining "context". In our case, we assessed the co-occurrence of our target words with respect to a list of high frequency content-bearing words. Our initial plan was simply to use a general frequency list for Icelandic, purged of stop words[3] (1000 words total). However, this list includes words not present in the 2,000 most frequent words from SubMIM. It was therefore decided to replace those words which occurred on the general list but not on the SubMIM list with words from the SubMIM list, giving a hybrid 1,000 word list. To reduce noise, the 100 most frequent words were then removed, giving a final list of 900 words.

[2] http://sourceforge.net/projects/icenlp

[3] Another approach is taken by Bullinaria (2008), who does not remove stop words from the list of context words.

Information was then collected for every target noun on how often it occurred within a 25 word window of each of these 900 context words: i.e. the number of times a context word occurs within 12 words before or 12 words after a target noun. This information was represented in a matrix with the target nouns labeling the rows and the common content words labeling the columns. Each cell therefore contains the number of times a target noun (row) co-occurs with a content word (column). (See Table 1 below.)

Significant parameters which affect the characteristics of this co-occurrence matrix, and hence the overall results, include (cf. also Bullinaria 2008): the size and quality of the corpus being used and the information represented in the corpus, i.e. our corpus is part-of-speech tagged and lemmatized allowing us to extract the noun lemmata; the choice of the context words (the compilation of our list of 900 frequent content-bearing words represents in itself a range of significant choices); and the size of the co-occurrence window (e.g. co-occurrence could mean simple adjacency or a window of increasing length, up to perhaps 100 words).

To give a simple example of co-occurrence from our corpus, the target words *þvottahús* 'laundry room' and *baðherbergi* 'bathroom' co-occur with the context word *íbúð* 'flat, apartment' within a window of 25 words:

> ... hafa sérgeymslur inni í **íbúð**unum, bað með aðstöðu fyrir þvottavél, auk sameiginlegs <u>þvottahús</u>s og sameiginlegrar geymslu ...

> ... **íbúð**irnar afhendast fullbúnar án gólfefna en <u>baðherbergi</u> er flísalagt ...

In processing these snippets the program increments a counter for both target words in the cell representing the context word *íbúð*.

If *þvottahús* and *baðherbergi* generally share similar contexts in the corpus, i.e. both mostly appear near the same context words, their similarity value will be high[4]. Note how in the (fictive) co-occurrence matrix shown in Table 1 the top three target words share similar context, whereas *literature* and *cod* have different distributions:

	cw 1	*cw 2*	*cw 3*	*cw 4*	*cw 5*	*cw 6*
dining room	7	0	5	10	0	0
bathroom	11	0	9	9	0	0
laundry room	8	0	9	11	0	0
literature	0	8	0	0	0	0
cod	0	0	0	0	14	23

Table 1: An example of a co-occurrence matrix (cw = context word)

3.3. Similarity measures

To assess semantic relatedness, the rows in the co-occurrence matrix (representing the distribution of individual target nouns) must be compared. To increase the information content of the co-occurrence counts, the raw count data has to be transformed, e.g. by a logarithmic value (for different measures see e.g. Manning and Schütze 1999; Sahlgren 2006; Bullinaria 2008). The comparison of two rows follows through a *similarity measure.* The most common similarity measure used in measuring semantic relatedness is the so-called *cosine* similarity measure (cf. Manning and Schütze 1999). This measure gives a similarity value between 0 and 1, such that words with very different distribution have a similarity value closer to 0 but those with a very similar distribution have a similarity value closer to 1. Parameters which have an affect

[4] These snippets also contain other target words, but for illustrative reasons only two are discussed in the example – likewise some words can be both context and target words, and in some approaches this holds for every word.

on results in this case concern the methods for transforming raw count data and the choice of similarity measure.

In summary, it is possible to process a corpus and to compute the similarity of the distribution of words of interest. Similarity values gained in this way are seen as an estimate of the semantic relatedness between words.

3.4. Clustering methods

The similarity information described so far concerns the similarity of individual word pairs, such as *laundry room* and *bathroom*. However, such information across all the nouns in a corpus can be used to produce clusters of nouns, where clustering is based on their relative similarity to each other. From a lexical semantic point of view, the ideal cluster would fall under a superordinate concept or semantic domain. An important challenge for clustering is that it is not known in advance how many such superordinate concepts or domains are evoked by nouns in the corpus.

A clustering algorithm must therefore group the words solely on the basis of their similarity values; identification and labeling of the superordinate domains which result from successful clustering must be performed manually. Continuing with the example of *laundry room* and *bathroom,* the algorithm might group these words together with *dining room, bedroom, child's bedroom, entrance, garage, wardrobe, parquet,* etc. It is, however, the task of a human assessor to label this cluster, for example with the concept HOUSE.

As with the other measures mentioned above, many clustering algorithms exist. In our experiment described in the next section we used the *k*-means clustering algorithm, with some adjustments. The interested reader can find good descriptions of *k*-means and other algorithms for example in Manning and Schütze (1999) and Duda et al. (2001).

4. Experiment: Clustering of nouns in Icelandic

In the following we describe an assessment of an experiment clustering Icelandic nouns according to semantic relatedness. We use two different statistics based on co-occurrence counts of words and cluster the data using adjusted *k*-means. First, an overview of the general approach and results is given, and then sections 4.2 and 4.3 describe manual assessment of selected clusters, an expert validation and a comparison to the *Icelandic Dictionary* respectively.

4.1. General results

We use two different measures to transform the raw co-occurrence counts in the co-occurrence matrix containing rows of target words and columns of context words: a logarithmic measure (Manning and Schütze 1999:302) and Positive Pointwise Mutual Information, PPMI (Bullinaria and Levy 2007)[5]. As a result two distinct matrices are produced, a) the *log* co-occurrence matrix and b) the PPMI co-occurrence matrix. These matrices form the input for the clustering algorithm, where the cosine similarity measure is used to compute similarity of distributions (i.e. rows in the matrix). The results of the clustering based on the *log* co-occurrence matrix show 79 clusters, each containing from 9 to 192 words, whereas the PPMI matrix results in 76 clusters with 4 to 198 words. The cluster content is automatically ranked according to how close a word is to the word considered being “in the middle” of the cluster.

[5] PPMI is a statistical method based on conditional and overall probabilities, used to increase the information content of raw co-occurrence data. We avoid technical details here but interested readers can find a more thorough account in the reference provided.

From a shallow screening of the clusters, it seems that 60 clusters from the *log* matrix can be characterized by a subsuming concept and 59 clusters from the PPMI matrix. As the clustering is completely unguided – i.e. no pre-defined categories are given – the resulting clusters have different ontological status: scientific domains (BIOCHEMISTRY, BIOLOGY), concrete things (HOUSE, VEHICLE), domains containing mostly proper nouns (FOOTBALLERS, MUSIC/MUSICIANS) domains from public discourse (POLITICS, FINANCES/BUSINESS), etc.

We will now report on a preliminary manual assessment of these clusters. Lacking a gold standard to evaluate these results, we selected one domain for manual assessment, the domain of finances and business. In the following we describe the examination of these clusters.

4.2. Expert validation

For the expert validation, two employees of a bank were asked to rate relatedness of words to the domain of finances and business.

4.2.1. Clusters for expert validation

The clusters used in this part of the assessment come from the partition based on the *log* transformation measure. This partition includes five clusters related to the domain of finances and business, each of them containing from 74 up to 173 words, all in all 555 words. We selected the 50 top words from each cluster for the validation.

4.2.2. Validation by finance professionals

To evaluate to what extent the words in the clusters are related to the domain of finance and business, two employees from a bank were asked to rate the words from the clusters described in *Section*

4.2.1[6]. Four scores were possible: 3 – very related to the domain; 2 – fairly related to the domain; 1 – not particularly related to the domain; 0 – not at all related to the domain. This method produced an interesting result with respect to domain-specificity: the bankers were both so focused on the banking domain that words obviously related to the business domain, like *samsteypa* 'conglomerate' and *sölufyrirtæki* 'a selling company', were rated as 0. Even company names (such as the well-known bank, *Glitnir*) were mostly rated with 0, sometimes with 1. Given this domain-bias, only the results from the two clusters directly related to banking proved to be useful and they are the focus of this assessment.

The inter-annotator agreement with respect to the four scoring categories was 65%; however, grouping the positive scores (very/fairly related) and the negative scores (little/not-at-all related) increased inter-annotator agreement to 80%.

score	banker 1	banker 2
3	58	55
2	22	16
1	16	25
0	4	4

Table 2: Rating results from two employees of a bank, rating words in two clusters, each containing 50 words.

As shown in Table 2, according to *banker 1,* 80% of the words in the two clusters are very or fairly related to the banking domain and 71% according to *banker 2*. Taking only the inter-annotator agreement into account, 64% of the words in the two clusters are very or rather related to the domain. This degree of agreement is not especially high, though one has to bear in mind, that "there exist numerous equally valid alternative ways" of doing manual categorization (Bullinaria 2008:5).

[6] We want to thank the two employees at *Íslandsbanki* for their assessment.

4.3. Comparison with the Icelandic Dictionary (ÍO)

The purpose of the comparison with ÍO is twofold: to find out a) whether a data-driven method like this adds anything to the information already in the dictionary and b) whether the classification of those words present in the dictionary matches the clustering results[7]. This part of the assessment also concerns the domain of finance and business.

4.3.1. The Icelandic Dictionary (ÍO)

ÍO originates in the first general monolingual Icelandic dictionary from 1963: *Íslenzk orðabók handa skólum og almenningi* 'An Icelandic Dictionary for Schools and the General Public'. All later editions and revisions build on this version and no general modernisation has taken place.

The following assessment concerns the coverage of lemmata in ÍO. We will therefore report briefly on the lemma list and definition vocabulary. A substantial part of the lemmata in the first version (1963) comes from a bilingual Icelandic-Danish dictionary from 1920-1924: *Íslensk-dönsk orðabók* (Kvaran 1998). Even in the second edition of ÍO (1983), there is a bias towards dated vocabulary, but we are not aware of whether the planned correction of this bias back towards contemporary usage (Árnason 1998) has been performed. Árnason also notes that ÍO has inconsistent definition texts and unclear objectives for the selection of lemmata. It has furthermore been determined that c. 42% of the definition vocabulary did not form lemmata in the 1983 edition (Bjarnadóttir 1998). There is a lexicographical rule stating that derivatives and compounds built by productive word formation rules where meaning and form are predictable from the parts need not become

[7] We started the comparison also using *Stóra orðabókin um íslenska málnotkun* 'The Large Dictionary of Icelandic Language Use' (Jón Hilmar Jónsson 2005), but the coverage of the finance/business domain words was too small to be of use.

lemmata in a standard dictionary. However, this rule has not been consistently followed in ÍO (cf. Bjarnadóttir 1998:39):

eiturbikar bikar með eitri í ***poison goblet*** *a goblet containing poison*
eplakaka kaka með eplum í ***apple cake*** *a cake containing apples*

Yet even taking into account this rule (and other lexicographical lemma selection rules), Bjarnadóttir concludes that about 6,300 lexemes, or 12% of the definition vocabulary, are missing from the lemma list. It is thus apparent that the construction of the lemma list in ÍO has not followed strict guidelines, and the reasons for a word being included or not included in the lemma list are not always clear-cut.

Since the year 2000, ÍO has been accessible on the web (http://snara.is) and that is the version which we use for our experiment.

4.3.2. Clusters for the dictionary comparison

Both partitions (the one based on *log* and the one based on PPMI) include five clusters related to the domain of finances and business. The *log*-clusters have a total of 555 words and the PPMI-clusters a total of 582 words, with 458 words being common to both partitions. As with the clusters for the expert validation, the 50 top words from each cluster were selected. The resulting lists were merged, thus erasing the partition into distinct clusters and removing duplicates, and all proper nouns were deleted. The final list for the comparison with ÍO consists of 260 common nouns.

4.3.3. Lemma coverage

The first test concerns the question of how many words from the test list are lemmata in ÍO. Of the 260 selected words, 147 or 56.5% appear as a lemma in ÍO. Another 17 words, all compounds, are included in one of the compound lists within the definition of many lemmata. This means that 96 words or 36.9% are not listed in the dictionary. Some of these words are productive compounds, often avoided in dictionaries, like compounds with *heildar-* 'total': *heildarvelta* 'total turnover', and *heildarútflutningur* 'total export' (but recall that the dictionary lemma list also contains many productive compounds, see section 4.3.1). Given that our resource is intended for NLP applications, we want our database to include such compounds, as this will save applications from having to decode them individually.

Another possible reason for a word being or not being in the dictionary is frequency. The 260 words in the finance/business list appear from 18 to 134 times in SubMIM. The ten most frequent words are all lemmata in the dictionary, but other high frequency words like *markaðsvirði* 'market value' (81 occurrences), *hönnunarfyrirtæki* 'design company' (80 occurrences), *eignarhaldsfélag* 'holding company' (70 occurrences) and *yfirtökutilboð* 'take-over bid' (70 occurrences) are not. Also low frequency words are either in the dictionary (*smásöluverð* 'retail price') or not (*smásölustig* 'retail level') – both words appear only 18 times in the corpus.

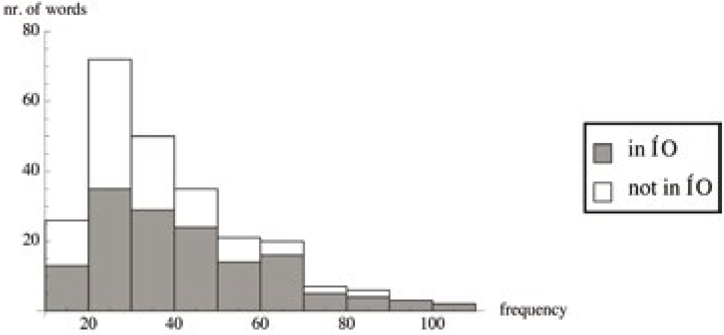

Figure 1: Frequencies of words from the finance/business domain listed and not listed in ÍO (the highest frequency numbers are left out to make the graph better readable)

Figure 1 shows the frequency distribution of the 260 words, separated by occurrence and non-occurrence in ÍO. This suggests that our clustering technique does indeed provide useful results for extending dictionary coverage.

4.3.4. Domain labeling

The second evaluation task regarding ÍO was to compare the classification of the words listed in the dictionary with the cluster domain. ÍO has the domain label *viðskipti/hagfræði* 'business/ economics'. Of the 147 words listed as lemmata, 52 or 35.4% have this domain assignment. Four words are assigned to the domain *stjórnsýsla* 'administration'; *ál* 'aluminium' and *dísilolía* 'diesel' are assigned to *eðlis-/efnafræði* 'physics/chemistry', though they can also be seen as related to economics. Three unrelated words have other domain assignments. That leaves 86 words without any domain assignment: of these, words like *lánsfé* 'loan capital', *afborgun* 'amortization', and *bankareikningur* 'bank account'are all strongly related to the finance domain; however, these 86 words also include items that are completely unrelated to the finance domain, such as *kindakjöt* 'mutton', *vín* 'wine', and *varahlutur* 'spare part'. Once again, our clustering technique does provide useful results for extending the domain information in the dictionary, though for this very reason it makes the dictionary an ineffective reference point for the assessment of cluster quality.

5. Conclusion

We have described one method for extracting semantic information from text. These results will contribute to the development of a semantic database for Icelandic language technology. The methods described here will be used alongside other techniques (cf. Nikulásdóttir and Whelpton 2010), in the belief that a hybrid methodology (Pantel and Pennachiotti 2008; Cimiano 2006; Cederberg and Widdows 2003) will yield the highest quality results from limited resources. The results reported here also suggest that clustering by semantic relatedness can be of great use to traditional lexicography in discovering potential lemma candidates.

References

Árnason, Mörður 1998: Endurútgáfa "Íslenskrar orðabókar". Stefna – staða – horfur. In: *Orð og tunga* 4:1-8.

Bjarnadóttir, Kristín 1998: Um skýringarorðaforðann. In: *Orð og tunga* 4:33-44.

Blöndal, Sigfús 1920-1924: *Íslensk-dönsk orðabók.* Reykjavík.

Borin, Lars and Markus Forsberg 2009: All in the Family: A Comparison of SALDO and WordNet. In: Bolette Sandford Pedersen et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies,* vol. 7 of *NEALT Proceedings Series,* Odense, Denmark, 7-12.

Boyd-Graber, Jordan; Christiane Fellbaum, Daniel Osherson, and Ropert Schapire 2006: Adding Dense, Weighted Connections to WordNet. In: Petr Sojka et al. (eds): *Proceedings of the GWC,* 29-35.

Budanitsky, Alexander and Graeme Hirst 2001: Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of five Measures. In: *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001),* Pittsburgh, PA, 29-34.

Bullinaria, John A. 2008: Semantic Categorization Using Simple Word Co-occurrence Statistics. In: M. Baroni et al. (eds): *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics,* Hamburg: ESSLI, 1-8.

Bullinaria, John A. and Joseph P. Levy 2007: Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. In: *Behavior Research Methods,* 39:510-526.

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

Cimiano, Philipp 2006: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* Springer.

Duda, Richard O.; Peter E. Hart and David G. Stork 2001: *Pattern Classification.* New York, Chichester etc.: John Wiley.

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

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

Helgadóttir, Sigrún 2004: Mörkuð íslensk málheild. In: *Samspil tungu og tækni.* Reykjavík: Ministry of Education, Science and Culture, 65-71.

Íslenzk orðabók handa skólum og almenningi 1963. Árni Böðvarsson (ed). Reykjavík: Menningarsjóður.

Íslensk orðabók handa skólum og almenningi 1983. 2nd ed. Árni Böðvarsson (ed). Reykjavík: Menningarsjóður.

Ingason, Anton Karl; Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson 2008: A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Bengt Nordström et al. (eds): *Advances in Natural Language Processing,* vol. 5221 of *Lecture Notes in Computer Science,* Berlin: Springer, 205-216.

Kvaran, Guðrún 1998: Uppruni orðaforðans í ”Íslenskri orðabók”. In: *Orð og tunga* 4:9-16.

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

Loftsson, Hrafn and Eiríkur Rögnvaldsson 2007: IceParser: An Incremental Finite-State Parser for Icelandic. In: Joakim Nivre et al. (eds): *Proceedings of the 16th Nordic Conference on Computational Linguistics (NODALIDA),* 128-135.

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

Nikulásdóttir, Anna Björk and Matthew Whelpton 2010: Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In: *Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, Malta.

Nikulásdóttir, Anna Björk and Matthew Whelpton 2009: Automatic Extraction of Semantic Relations for Less-Resourced Languages. In: Bolette Sandford Pedersen et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies,* vol. 7 of *NEALT Proceedings Series,* Odense, Denmark, 1-6.

Nikulásdóttir, Anna Björk 2007a: *Automatische Extrahierung von semantischen Relationen aus einem einsprachigen isländischen Wörterbuch* [Automatic Extraction of Semantic Relations from a Monolingual Icelandic Dictionary], Master's thesis, University of Heidelberg.

Nikulásdóttir, Anna Björk 2007b: Sjálfvirk greining merkingarvensla í *Íslenskri orðabók*. In: *Orð og tunga* 9:5-24.

Pantel, Patrick and Marco Pennacchiotti 2008: Automatically Harvesting and Ontologizing Semantic Relations. In: Paul Buitelaar and Philipp Cimiano (eds): *Ontology Learning and Population: Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning from Text.* IOS Press.

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

Pedersen, Ted; Siddharth Patwardhan, and Jason Michelizzi 2004: WordNet::Similarity – Measuring the Relatedness of Concepts. In: *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04),* San Jose, CA, 1024-1025.

Resnik, Philip 1995: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95),* San Mateo, CA. San Francisco: Morgan Kaufmann, 448-453.

Rögnvaldsson, Eiríkur; Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna Björk Nikulásdóttir, Matthew Whelpton, and Anton Karl Ingason 2009: Icelandic Language Resources and Technology: Status and Prospects. In: Rickard Domeij et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources,* vol. 5 of *NEALT Proceedings Series*, Odense, Denmark, 27-32.

Sahlgren, Magnus 2006: The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High Dimensional Vector Spaces. PhD thesis, Stockholm University.

Turney, Peter 2006: Similarity of Semantic Relations. In: *Computational Linguistics* 32(3), 379-416.

Weeds, Julie 2003: *Measures and Applications of Lexical Distributional Similarity.* Ph.D. thesis, University of Sussex.

Zesch, Torsten and Iryna Gurevych 2009: Wisdom of Crowds versus Wisdom of Linguistics – Measuring the Semantic Relatedness of Words. In: *Natural Language Engineering,* 16(1):25-59.

Internet references

ÍO = *Íslensk orðabók* (2007). Mörður Árnason (ed). 4th edition. Reykjavík: Edda http://snara.is (March 2010).

Jón Hilmar Jónsson (2005): *Stóra orðabókin um íslenska málnotkun*. Reykjavík: JPV. http://snara.is (March 2010).

Anna Björk Nikulásdóttir
Ph.D. student
University of Iceland
Sæmundargötu 2
IS-101 Reykjavík
abn@hi.is

Matthew Whelpton
Associate Professor
University of Iceland
Sæmundargötu 2
IS-101 Reykjavík
whelpton@hi.is