Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic

Anna B. Nikulásdóttir, Matthew Whelpton

University of Iceland
Reykjavík, Iceland
abn@hi.is, whelpton@hi.is

Abstract

This paper describes work in progress on semi-automatically constructing a semantic database for Icelandic. The focus is on methods for the extraction of semantic relations used to collect material for the database. Established methodologies have largely focused on English. As Icelandic is a less-resourced language with much richer inflection than English, we must make adjustments to these established methods to address linguistic and sparse data problems. As a general principle, we aim to develop methodologies which will be viable for other less-resourced languages, with the support of open source tools.

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. Global Wordnet Association[1]). 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 PoS-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, 2007); following the success of this study[2] 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: 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)[3]. 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 characteristics of the language: (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. It should be noted that a preliminary motivation for our work is the desire to build a database of "native" semantic relations for Icelandic, i.e. to extract information from Icelandic resources, reflecting distributional and collocational properties of Icelandic lexemes (cf. also DanNet, (Pedersen et al., 2009)) rather than to fit the Icelandic data to an external model, for instance by importing the ontological structure of the English WordNet by translation (cf. e.g. (Fernández-Montraveta et al., 2008)). Whether, and to what extent, these two methodologies produce significantly different results remains an open question but our aim is to contribute to the ultimate evaluation of such differences by contributing a native ontology for Icelandic — and to test the extent to which such an aim is achievable for languages like Icelandic with relatively limited resources.

In this paper we describe the context of our work and report our first experiments with the extraction from text of semantic information on nouns. We want to stress that we describe work in progress and that no formal evaluation data is available yet. In section 2 we review the inflectional properties of Icelandic nouns and describe the currently-available corpus. The final design of the Icelandic semantic database has not yet been established but in section 3 we consider several important issues relating to the structure of wordnets. Following the hybrid methodology developed in recent years (Cederberg and Widdows, 2003; Cimiano et al., 2005; Pantel and Pennacchiotti, 2008) we exploit different methods, both pattern-based and statistical, to extract semantic information. We describe those in sections 4 and 5. In section 6 some possibilities for validation and expansion of results are discussed, concluding with an assessment of prospects for future work.

2. A Corpus of Icelandic

Icelandic is a highly inflected language, which makes a PoS-tagged and lemmatized corpus essential for any further automatic processing of Icelandic text. Nouns in Icelandic inflect for case, number and gender, and so does the cliticised definite article[4]. An example of a lemmatized

[1] `http://www.globalwordnet.org`

[2] For reasons of space, we are unable to review this study here. Those interested in a detailed overview of this study and of the relations extracted should consult (Nikulásdóttir and Whelpton, 2009) and (Nikulásdóttir, 2007).

[3] `http://iceblark.wordpress.com`

[4] Icelandic does not have an indefinite article

Kepa Sarasola, Francis M. Tyers, Mikel L. Forcada (eds.)
7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May 2010, p. 33–39

PoS-tagged noun *börnunum* '(to) the children' is:

TOKEN	POS-TAG	LEMMA
börnunum	nhfþg	barn

The PoS-tag stands for noun (n), neuter (h), plural (f), dative (þ), and definite article (g). Additionally, a special tag for proper nouns is included in the tagset for nouns. At the moment a balanced PoS-tagged, lemmatized corpus, 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 8.8 million tokens, including punctuation marks etc. The source of this data is mainly newspaper texts (*Morgunblaðið*), but further texts come from a public science web portal at the University of Iceland (*Vísindavefurinn*[5]), reports from Icelandic ministries, and from a medical Journal (*Læknablaðið*). Two versions of SubMIM are used for different automatic extraction methods: (a) the basic PoS-tagged and lemmatized version is used for the statistical methods and (b) a shallow-parsed version without lemmata is used for the pattern-based methods. The tagging, lemmatizing and parsing was performed using the PoS-tagger *IceTagger*, the lemmatizer *Lemmald*, and the shallow parser *IceParser*, all included in the open source IceNLP-toolkit[6]. The parsed version was parsed without using the option of marking grammatical functions, but another version including those tags will be useful for some further experiments (see section 5). The corpus is fully automatically processed and no manual correction has yet been performed.

Nevertheless, for the foreseeable future, Icelandic LT faces a serious sparse data problem compared to English; although we will supplement SubMIM (and MIM when it is complete) with web-based data, this problem will remain. Work on the lexicon needs very large corpora: even the British National Corpus[7] (BNC) with its 100 million tokens has been shown to be too small for a broad coverage statistical analysis of word occurrences (Kilgarriff and Grefenstette, 2003). We are encouraged by the development in Leipzig, Germany, of a 250 million token corpus of Icelandic (Hallsteinsdóttir et al., 2007), collected from all `.is` domains, and we hope to include this in our dataset. However, this still puts us far behind the corpus of billions of words which is now being developed for English through web crawling (Pomikálek et al., 2009).

3. Relations in wordnets

The primary relations between nouns in WordNet are synonymy and hyponymy (Fellbaum, 1998). Synonymous or near-synonymous words build synsets as labels of concepts. Other relations like hyponymy or meronymy hold between synsets, exceptionally other relations hold between words, like antonymy. This organisation is followed by all wordnets within the EuroWordNet[8] scheme, even if they include more relations, such as for instance "involved agent" (*violin* INVOLVED_AGENT *violinist*) or "role agent" (*passenger* ROLE_AGENT *journey*)[9]. DanNet (Pedersen et al., 2009) extends the EuroWordNet template with for instance the relations "concerns" (*goal* CONCERNS *sport*) and "has_hypernym ortho" (*road-side tree* HAS_HYPERNYM_ORTHO *tree*). All these databases have in common that all links between words and synsets are labelled with a defined relation. However, many NLP applications that use wordnets could benefit from a more dense structure of arcs, including nonclassical relations (Morris and Hirst, 2004; Zesch and Gurevych, 2009). There is indeed a plan to extend WordNet with directed, weighted arcs between synsets, that correspond to the "evoking" relation, i.e. how strongly one synset evokes another one according to a human (Boyd-Graber et al., 2006). One interesting resource in the family of semantic networks is SALDO, a Swedish Associative Thesaurus (Borin and Forsberg, 2009). It is strictly hierarchical, but relies on loosely characterized associative relations rather than classical semantic relations.

The structure of the semantic database for Icelandic LT has not yet been fixed but the aim is that the structure should be as data-driven as possible, i.e. as much as possible of the extracted semantic information should find its way into the database and thus it should not be limited to classical semantic relations.

4. Pattern-based methods for relation extraction

Pattern-based methods for relation extraction have been widely used since the publication of (Hearst, 1992). The essence of these methods is to use seed words known to be in a certain semantic relation to harvest syntactic and/or lexico-syntactic patterns indicating the relation in question. As an example, Hearst discovered several patterns for the extraction of hypernymy with the seed words *England - country*, e.g. the pattern `NP {, NP}* {,} and other NP` (Hearst, 1992, p. 541) from phrases like *England and other countries*. Such patterns are known to be reliable in the extraction of hypernyms (Cimiano et al., 2005) and have also been used for the extraction of meronyms (part-whole relations) (Berland and Charniak, 1999; Girju and Badulescu, 2006). Using few seed words requires a large corpus, since one needs (a) the words to occur several times togehter and (b) ideally in different patterns, but reliable patterns have been shown to be low frequency in English (Cimiano et al., 2005). As an experiment to deal with the sparse data problem we developed a method called validation of most common syntactic patterns. The motivation for this method is twofold: (a) to recognize as many patterns as possible from sparse data (b) without having to predefine the relations that are to be extracted. The method includes four steps: (i) extract syntactic patterns with their actual realizations from the corpus according to some predefined criteria - in this case every noun and prepositional

[5] `http://www.visindavefur.is`
[6] `http://sourceforge.net/projects/icenlp`
[7] `http://www.natcorp.ox.ac.uk`
[8] `http://www.illc.uva.nl/EuroWordNet`
[9] These examples are adapted and translated from DanNet, `http://wordnet.dk/dannet/lang`

phrase in the corpus; (ii) with the help of a special GUI, loosely validate the most common patterns according to possible indication of some semantic relation - in this case every pattern occurring at least 10 times in the corpus; (iii) combine similar patterns using an edit distance algorithm and regular expressions; (iv) extract related words from the corpus. In the following we will discuss each step in more detail.

4.1. Extraction of patterns

For the extraction of patterns the shallow-parsed version of SubMIM was used (see section 2). The aim was then to extract all noun phrases and prepositional phrases that could possibly include semantically related nouns and/or adjectives. Thus the criteria for the extraction of phrases were (a) all noun phrases including more than one noun or at least one noun and at least one adjective, (b) all coordinated noun phrases or adjective phrases, (c) all prepositional phrases including more than one noun, (d) all connected noun and prepositional phrases. Always the longest possible chain of phrases was extracted.
The following example shows how IceParser analyzes the phrase *feitur og kryddaður matur* 'greasy and spicy food', and the pattern extracted from the parser output:

IceParser output:
```
[NP[APs[AP feitur lkensf AP][CP og
c CP][AP kryddaður lkensf AP]APs]
matur nken NP]
```
Extracted pattern:
```
[NP[APs[AP lensf][CP og c][AP
lensf]] nen]
```

Here "NP" denotes a noun phrase, "APs" a sequence of adjective phrases, "AP" an adjective phrase and "CP" a coordinating conjunction. The PoS-tag starting with "l" stands for adjective, the one starting with "n" for noun and "c" stands for conjunction. We ignore the tag for gender so the "k" (for masculine) in the PoS-tags for nouns and adjectives is removed. In general, no words are included in the patterns except conjunctions and prepositions, since they can help identify semantic relations. In this case *og* 'and' is retained.
About 370,000 different patterns were extracted in this way. Of these, about 94,000 occur more than once and about 5,300 more than ten times in SubMIM. Only those patterns occurring more than ten times were validated. The reason for this high number of patterns is the large tagset of Icelandic, which contains about 700 tags. This high granularity is probably not necessary and has in many cases been ignored in the process of merging the patterns. The final tests will show which tags need to be kept in the patterns and which can be ignored.

4.2. Validation of patterns

For the validation of the most common patterns a simple GUI was developed. Selecting a pattern lists all instances of the pattern plus frequency. If some semantic relatedness is salient between nouns and/or adjectives in a majority of the instances, the validator assigns one predefined category to the pattern. The categories were defined after a first look at the most common patterns and they are thought of as a rough partition of the patterns, which are to be tested further through relation extraction. Five of the categories refer to the syntactic structure of a pattern: genitive construction, attributive construction (adjective(s) plus noun), and coordinated nouns, adjectives or proper nouns; and three refer to a semantic relation: superordinate, location and role. If none of these categories apply to a pattern, but the validator thinks it might indicate some semantic relatedness, the category "other" is chosen. Less than the half of all examined patterns (i.e. 2,275 of 5,268 patterns) were classified as possibly indicating a semantic relation.
Since the GUI only takes results from the pattern extraction as input and does not process this input in any way, it is totally language independent. The final version of the GUI will be open source, as is consistent with our general aim of building up a shared set of methodologies and tools for less-resourced languages.

4.3. Merging of patterns

As the number of positively validated patterns was much higher than expected, they must be simplified and merged in some way. The tag for number was removed from the patterns and for pronouns and adjectives all tags except the ones marking the word class were removed. The patterns are then merged and generalized using the minimum edit distance algorithm and then further merged using regular expressions. The method for the computation of minimum distance between patterns and their generalization was adaptded from (Ruiz-Casado et al., 2005). Given for example two patterns expressing a genitive construction `[NP nn][NP ne]` (nominative noun - indefinite genitive noun) and `[NP nn][NP neg]` (nominative noun - definite genitive noun), the distance is 1, due to the single difference of definiteness `ne` vs. `neg`. The generalization algorithm then makes one pattern out of the two: `[NP nn][NP ne|neg]`. A standard regular expression for this pattern would be `[NP nn][NP neg?]`, and often several general patterns can be unified in one regular expression. This has been done for several of the identified pattern categories and shows considerable reduction in the number of patterns. The original number of patterns including genitive constructions was 384, but through the simplifying and merging process they have been reduced to 71 patterns. As stated above, the patterns may be refined after the final testing.

4.4. Extraction of relations

Since we are presenting work in progress, no final evaluation data is available yet. However, we have extracted relations from SubMIM based on patterns marked as *superordinate, coordinated nouns*, and *genitive construction*. Only one pattern was used to extract superordinates or hypernyms, an Icelandic equivalent to one of the patterns introduced by (Hearst, 1992): `NP (,NP)*, and|or other NP` (in Icelandic nine different morphological forms of *other* can be used in this pattern). All in all 369 hypernyms were extracted, which is a rather low number. Cederberg and Widdows (2003) extracted 513

hypo-/hypernym pairs from approximately the first 430,000 words from the BNC, using the six patterns reported by (Hearst, 1998). Finding more patterns in Icelandic indicating hypernymy would possibly increase the number of extracted hypernyms. The results are however reliable in that most word pairs express some kind of a sub-/superordinate relation, although how many are really taxonomic hypo-/hypernym pairs still needs to be evaluated.

Genitive constructions can express many different relations between the involved nouns. The main interest for the extraction of semantic relations is the part-whole relation often expressed by a genitive construction (Berland and Charniak, 1999; Girju and Badulescu, 2006; Pantel and Pennacchiotti, 2008). The problem is that it is impossible to judge from the lexical-syntactic pattern alone whether it expresses a part-whole relation or not. Berland and Charniak (1999) extract parts of a given seed word representing the whole, e.g. *car*. They then filter out words having the suffixes *ness, ing* and *ity*, since those tend to express qualities rather than parts. Finally they use a probability measure to rank their results, which gave them 55% accuracy with the top 50 words. Another approach was taken by (Girju and Badulescu, 2006), where they used WordNet to ontologize the extracted word pairs. With a training set and a learning algorithm, information on the ontological category of each word in the pair was used to deduce the likelihood of a new part-whole relation between the pair.

For the moment we call the general relation expressed by a genitive construction the relation of properties. Even without refining this relation, it can give valuable information. The results already reveal polysemy of terms, not necessarily accounted for in dictionaries: *cod* as "fish" (*the stomach of the fish*) or as a "product" (*the market price of the fish*); *house* as a "building" (*the roof of the house*), as a "property" (*the running of the house*), as a "theatre" (*the house consultant [house=theatre]*), or as a "restaurant/pub" (*the house band*). It is also possible to categorize relations including the action nouns registered as such in existing lexicographic resources, following the filtering of *ness, ing,* and *ity* proposed by Berland and Charniak (1999). The next step in processing this material will involve validation by human assessors and by statistical testing, which could improve results.

Noun-coordination information has been used to collect co-hyponyms (Roark and Charniak, 1998; Caraballo, 1999) and Cederberg and Widdows (2003) use it to extend results from pattern-based hypernym extraction. Then a hyponym is used as a seed word to extract potential co-hyponyms, since coordinated nouns often belong to the same hierarchy level in a hypernym hierarchy. We will discuss the potential use and problems of this pattern in section 6.2.

5. Semantic relatedness and clustering

We adopt a broad conception of semantic relatedness, by which words that belong to the same semantic domain or topic are semantically related; this broad definition is in line with (Manning and Schütze, 1999, p. 296), though they use the term "semantic similarity". A thorough discussion of semantic similarity and semantic relatedness can be found in (Zesch and Gurevych, 2009) and (Turney, 2006). An established way to compute semantic relatedness is to use the cosine similarity measure between two vectors. The vectors can be built by counting cooccurrences of words of interest - in our case nouns - with the most frequent content-bearing words of the language (Manning and Schütze, 1999; Cederberg and Widdows, 2003) or for example according to their cooccurrence with verbs in certain grammatical functions such as subject or direct object (Weeds, 2003; Cimiano, 2006). We intend to exploit both methods and first experiments have been conducted with the former method, based on cooccurrences with frequent content-bearing words. We have used the tagged and lemmatized version of SubMIM, though, for languages without access to such a resource, it is also possible to perform this co-occurrence analysis on a clean corpus (Bullinaria, 2008). The resulting cooccurrence matrix from SubMIM has about 11,300 rows representing nouns, including proper nouns, and 900 columns representing frequent content-bearing words. The content-bearing words were obtained from a word frequency list for Icelandic. This list was filtered for stop words and compared with the most frequent words in SubMIM. Several words from the common frequency list were not high frequency in SubMIM and were exchanged with corpus specific high frequency words. The top 100 words from the resulting list of 1,000 high frequency content-bearing words were then deleted (see (Manning and Schütze, 1999, p. 302)), leaving a list of 900 words used for the cooccurrence analysis. With the correction of the automatic lemmatization and a larger corpus we expect both a larger number of results and an improvement in the computation of semantic relatedness. Nevertheless we used these results in a preliminary experiment on clustering nouns from SubMIM with respect to semantic relatedness. To reduce noise in the clustering data, very frequent and very rare words were eliminated (cf.(Dhillon and Modha, 2001)). Excluding words occurring in collocation with more than 15% of the 900 content-bearing words and words that are not counted more often than 18 times[10] (18 is 2% of the column size of the matrix), the nouns to be clustered were reduced to 7,871 nouns. The first choice in exploring new data with clustering is often the *k*-means algorithm, an elementary but very popular approximation method (Duda et al., 2001, p. 526). In this algorithm a set of initial cluster centers is randomly defined. The data elements are then assigned to the closest center, according to some distance measure - here the cosine similarity measure - and then the center of each cluster is recomputed. The procedure of assigning data elements and recomputing centers is then repeated until some stopping criterion is reached.

The first observation made after running *k*-means on our data was that just like the Euclidean distance measure most often used in *k*-means (Manning and Schütze, 1999, p. 516), clustering using cosine similarity results in singelton clusters. At the same time, several clusters were very large and as one would expect, they normally have poor cluster quality. Cluster quality was computed by summing up the similarity values of all members of a cluster with

[10] With more data this threshold should be set higher, normally higher frequency is needed for lexical statistical analysis (Kilgarriff and Grefenstette, 2003).

Extracted word pair	Pattern	Similarity
líkami - fruma ('body' - 'cell')	genitive construction	0.7435
námskrá - grunnskóli ('curriculum' - 'elementary school')	genitive construction	0.5326
morð - glæpur ('murder' - 'crime')	superordinate	0.5165
þröstur - fugl ('thrush' - 'bird')	superordinate	0.4923
þorskur - botnfiskur ('cod' - 'bottom dweller')	superordinate	0.4921
þróun - líftæknifyrirtæki ('development' - 'biotechnology company')	superordinate	0.4529
frumskógur - málari ('jungle' - 'painter')	genitive construction	0.2676

Table 1: Results from a pattern-based method validated with a semantic relatedness measure

its mean vector and dividing by number of members, thus getting the average similarity value for the cluster (Dhillon and Modha, 2001). To force the algorithm to make clusters of reasonable size, i.e. not too small and not too large, an elementary validation process was implemented. It examines a finished *k*-means partition and deletes all clusters that have less than four members. If a cluster has more than some MAX members, it is split in two clusters. The validation process thus can change the initial *k* number of clusters. After the validation *k*-means is run again. This is repeated until no change is made to the partition in the validation process. Results on SubMIM using MAX=200 and initial number of clusters *k*=32 show 60 clusters containing from 23 to 184 words. We found that 46 of the 60 clusters can be characterized by a subsuming concept, whereas 14 cannot. These concepts have different ontological status so that a one-to-one mapping in an ontology is not possible. There are traditional scientific domains like BIOCHEMISTRY (*hormone, secretion, metabolism*), BIOLOGY, and METEOROLOGY, domains from public discourse like FINANCES (*privatisation, tax environment, monopoly*) POLITICS, and GLOBALISATION, concrete things like HOUSE (*bathroom, master bedroom, laundry room*), VEHICLE, as well as domains containing mostly proper nouns like FOOTBALLERS (*lampard, gerrard, thierry*), MUSIC/MUSICIANS, and PROPER NOUNS in general.

6. Combination of results

In order to improve results gained from different extraction methods, it is reasonable to combine them and so be able to extend and/or validate results. A word pair extracted with one method can be supported or not supported through results from another method. As shown by Cimiano et al. (2005), different pattern-based methods with different resources can give better results than just using one resource and Cederberg and Widdows (2003) use latent semantic analysis and noun coordination information to improve results of automatic hyponymy extraction. Pantel and Pennacchiotti (2008) extend their pattern-based method with a measure of pattern and instance reliability. With hybrid methods like this it should be possible to reduce the human validation effort which will be necessary at some point during the building of the semantic database.

6.1. Validation

Many studies on extraction of semantic relations from English text use WordNet to validate the results, e.g. (Pantel and Pennacchiotti, 2008). Neither a WordNet-like resource nor a semantically annotated corpus is available for Icelandic, so some kind of cross-validation between the extraction methods will be used for validation. Like Cederberg and Widdows (2003) we use semantic relatedness values to verify results from pattern extraction. Although we still need to correct and extend data used for the computation of semantic relatedness (see section 5), it seems to be a valuable measure on extracted hypernyms and word pairs from genetive constructions, as shown in table 1. But since semantic relatedness in our sense means belonging to the same topic or domain, incorrectly extracted taxonomic relations like hypernymy can still get high similarity values, as shown for *development - biotechnology company* in table 1. Cederberg and Widdows (2003) achieved a 30% reduction in error using this kind of semantic relatedness to verify extracted hypernyms, precision improved from 40% to 58%. It is a matter of further evaluation to see if we are able to improve our results in this way, despite these findings.

In order to be able to evaluate the results systematically, more data on relatedness is needed since most of the extracted word pairs are not found in the present similarity matrix. For very low frequency words we may not get this data, but with the use of a larger corpus a considerable extension of similarity measures should be possible.

6.2. Extension

As Cederberg and Widdows (2003) showed, it is possible to extend results of hypernym extraction by extracting coordinated nouns of a hyponym. After extracting e.g. *cloves* as a hyponym of *spice* from *...sugar, honey, grape, must, cloves and other spices...* the hyponyms of *spice* can be extended with *nutmeg, cinnamon,* and *coriander* by extract-

ing . . . *nutmeg or cinnamon, cloves or coriander*. They also point out that a seed word like *cloves* has to be chosen carefully, so that different meanings of a seed word don't lead to extraction of words not related to the hypernym in question. We experienced this problem in our tests where we wanted to extract further co-hyponyms of *fugl* 'bird' as a hyponym of *dýralíf* 'animal life'. The word *fugl* can also mean *birdie* (from the domain of golf) and extracted co-hyponyms included *forgjöf* 'handicap' and *par* 'par'.

Another problem concerns the level of the hypernym. One result of the hypernym extraction was *þorskur* 'cod' IS-A *botnfiskur* 'bottom dweller'. This is correct, and so was the extraction of co-hyponyms including various sorts of fish. However, the hypernym here is too narrow: not all of the "co-fishes" are bottom dwellers so that they need to be subsumed under the broader hypernym *fish*.

One question that needs to be further investigated is if certain kind of properties expressing semantic features can be used to extend results. As an example, can a word known to have the property *beginning* be extended to having the property *end*? In examining some extracted words having the property *beginning*, differences regarding modality became apparent. While *a century, an aria* and *a book* all have a certain beginning and an end, there is no necessary or predefined end to *a marriage, a town*, or *the world*, just a potential one.

7. Conclusions and future work

We have addressed several methods for the extraction of semantic relations and given some provisional results in applying these methods to Icelandic. Our current work is based on a small corpus, while further corpus development is taking place. The next steps include thorough testing and evaluation of these methods as well as the implementation of further methods. Together, these should yield a considerable amount of lexical-semantic information about Icelandic nouns. We then face the challenge of combining this information with existing lexical resources, as well as the relations already extracted from the Icelandic dictionary, to build the basis of a semantic database for Icelandic. Throughout this process, we are also guided by the long-term aim of mapping this Icelandic database to WordNet (see e.g. (da Silva et al., 2008)), which has practical ramifications for organization of the resource.

We hope that our efforts will benefit not only the development of Icelandic LT but also other less-resourced languages, by identifying effective methods for addressing the sparse data problem and by contributing necessary open source tools.

8. References

Matthew Berland and Eugene Charniak. 1999. Finding Parts in Very Large Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64.

Lars Borin and Markus Forsberg. 2009. All in the Family: A Comparison of SALDO and WordNet. In Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb, and Ruth Vatvedt Fjeld, editors, *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7 of *NEALT Proceedings Series*, pages 7–12, Odense, Denmark.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Ropert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. In Petr Sojka, K.-S. Choi, Christiane Fellbaum, and P. Vossen, editors, *Proceedings of the GWC*, pages 29–35.

John A. Bullinaria. 2008. Semantic Categorization Using Simple Word Co-occurrence Statistics. In M. Baroni, S. Evert, and A. Lenci, editors, *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 1–8, Hamburg, Germany. ESSLLI.

Sharon Caraballo. 1999. Automatic Construction of a Hypernym-Labeled Noun Hierarchy from Text. In *Proceedings of ACL*, pages 120–126.

Scott Cederberg 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)*, pages 111–118.

Philipp Cimiano, Aleksander Pivk, Lars Schmidt-Thieme, and Steffen Staab. 2005. Learning Taxonomic Relations from Heterogenous Evidence. In Paul Buitelaar et al., editor, *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence*.

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

Bento Carlos Dias da Silva, Ariani Di Felippo, and Maria das Graças Volpe Nunes. 2008. The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portugese WordNet Database. In Nicoletta Calzolari et al., editor, *Proceedings of LREC2008*, pages 1535–1541.

Inderjit Dhillon and Dharmendra S. Modha. 2001. Concept Decompositions for Large Sparse Text Data using Clustering. *Machine Learning*, 42(1):143–175.

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

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge Mass., London.

Ana Fernández-Montraveta, Gloria Vázquez, and Christiane Fellbaum. 2008. The Spanish Version of WordNet 3.0. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Text Resources and Lexical Knowledge*, pages 175–182. Mouton de Gruyter.

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

Erla Hallsteinsdóttir, Thomas Eckart, Chris Biemann, Uwe Quasthoff, and Matthias Richter. 2007. Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07*, Tartu, Estonia.

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

Marti A. Hearst. 1998. Automated Discovery of WordNet Relations. In Christiane Fellbaum, editor, *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge Mass., London.

Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild [A Tagged Icelandic Corpus]. In *Samspil tungu og tækni*, pages 65–71. Ministry of Education, Science and Culture, Reykjavík.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In Bengt Nordström and Aarne Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 205–216, Berlin. Springer.

Adam Kilgarriff and Gregory Grefenstette. 2003. Web as Corpus. *Computational Linguistics*, 29(3):1–15.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit, editors, *Proceedings of the 16th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 128–135, Tartu, Estonia.

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

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

Jane Morris and Graeme Hirst. 2004. Non-classical Semantic Relations. In *Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the ACL*, pages 46–51, Boston, MA.

Anna Björk Nikulásdóttir and Matthew Whelpton. 2009. Automatic Extraction of Semantic Relations for Less-Resourced Languages. In Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb, and Ruth Vatvedt Fjeld, editors, *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, volume 7 of *NEALT Proceedings Series*, pages 1–6, Odense, Denmark.

Anna Björk Nikulásdóttir. 2007. Autmatische 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.

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

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

Jan Pomikálek, Pavel Rychlý, and Adam Kilgarriff. 2009. Scaling to Billion-plus Word Corpora. In *Advances in Computational Linguistics. Special Issue of Research in Computing Science*, volume 41, Mexico City.

Brian Roark and Eugene Charniak. 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic Lexicon Construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Eiríkur Rögnvaldsson, 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, Kimmo Koskenniemi, Steven Krauwer, Bente Maegaard, Eiríkur Rögnvaldsson, and Koenraad de Smedt, editors, *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*, volume 5 of *NEALT Proceedings Series*, pages 27–32, Odense, Denmark.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic Extraction of Semantic Relationships for WordNet by means of Pattern Learning from Wikipedia. In A. Montoyo R. Munos and E. Métais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, volume 3513 of *Lecture Notes in Computer Science*, pages 67–79, Alicante, Spaine, June. Springer.

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

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

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