

Automatic Extraction of Semantic Relations For Less-Resourced Languages

Anna Björk Nikulásdóttir
University of Iceland
Reykjavík, Iceland
abn@hi.is

Matthew Whelpton
University of Iceland
Reykjavík, Iceland
whelpton@hi.is

Abstract

This paper addresses the challenge of creating a network of semantic relations for languages which do not have the resources of investment and manpower which have allowed the development of resources like WordNet for English. We first present a pilot study in this area which used a well-established pattern-based method to extract semantic relations from an Icelandic monolingual dictionary. This proved to have a good success rate for ten semantic relations. We then present a newly funded project which aims to extend and adapt this methodology for use with unstructured tagged corpora. We hope that this will allow the largely automated development of the target semantic resources.

1 Introduction

Although Icelandic language technology (LT) has taken great strides forward in the last ten years (Rögnvaldsson 2008), there are as yet no specifically LT-oriented semantic resources for Icelandic. However, Iceland has a rich lexicographic tradition which provides an excellent starting point for the development of such semantic resources. A pilot study in the exploitation of lexicographic material for the extraction of semantic relations has already been performed by Nikulásdóttir (2007) for Icelandic, building on classic studies for English (Alshawi 1987; Chodorow et al. 1985; Markowitz et al. 1986; Nakamura and Nagao 1988) and more recent work on Basque (Agirre et al. 2000). The pilot study gave promising results, with 94.77% of the analysed definitions being correctly or partly correctly analysed. A Grant of Excellence has just been awarded by the Icelandic Research Fund to the project “Viable Language Technology beyond English – Icelandic as a test case” (hereafter VLT), the first work package of which aims to extend Nikulásdóttir’s work in developing a

semantic network for Icelandic. We hope that the resources developed and the experience acquired in this project will i) lay the foundation for the development of a WordNet-like (cf. Fellbaum 1998) resource for Icelandic, and ii) serve as guidelines for other less-resourced languages for automatically extracting semantic relations.

Sections 2 to 4 of this paper offer an overview of Nikulásdóttir (2007). Nikulásdóttir used definitions in the 2002 3rd Edition of *Íslensk orðabók* ‘Icelandic Dictionary’ (henceforth ÍO) for her pilot study. Section 2 describes the characteristic format of noun definitions in the dictionary (2 main formats) and the issues that relate to these definition formats. Section 3 reviews ten semantic relations which were automatically extracted from definitions of nouns in ÍO: hypernyms, synonyms, holonyms, meronyms, verbal endonyms, adjectival endonyms, attributes, biological family, equivalences, and references. In Section 4, the results of the automatic extraction process are evaluated. The 94.77% success rate of the automatic extraction provides an encouraging basis for further work. However, practical considerations for a less-resourced language like Icelandic require the ability to extract semantic relations from large corpora of free text. The newly-funded VLT Project aims to develop methodologies to address this issue by supplementing the pattern matching methodology of the pilot study with latent semantic and co-ordination techniques and established statistical methods. This is discussed in Section 5.

2 Definition Formats in the Icelandic Dictionary

Definitions of nouns in monolingual dictionaries, such as ÍO, use certain syntactic patterns repeatedly in the formulation of particular kinds of definition (Geeraerts 2003: 89). It is therefore

possible to exploit the correlation between syntactic patterns and semantic relations for automatic extraction. The most common formats for noun definitions in ÍO are a) synonym definitions or synthetic definitions and b) a paraphrase, including *genus proximum* and *differentias specificas* (cf. Geeraerts 2003: 89). The use of the term *genus proximum* is, however, not unproblematic, since it should refer to the closest taxonomical hypernym. The head noun in a definitional paraphrase in a dictionary does not necessarily fit to that description, even if it represents a hypernym (Wiegand 1989: 548). We prefer to describe the paraphrasal definition as including a hypernym with features that distinguish the lemma from its co-hyponyms.

- a) synonym definition:
fagnaður 1 ánægja, gleði
joy 1 *pleasure, gladness*
- b) a paraphrase:
breiðband breitt tíðnisvið [*gen. prox.*] notað til fjarskipta, [...] **broadband** *a broad frequency range used for telecommunication, [...]*

In the ÍO database, definitions are segmented with regard to meaning items. A meaning item is a subpart of a definition serving various lexicographic functions (N.B. a definition may comprise just one meaning item). The following definition for example includes three meaning items:

dílaburkni [1] íslensk burkategund [2] (*Dryopteris assimilis*) [3] af þrílaufungsætt, með fjaðurskiptum blöðum, vex í gjám og kjarri
... [1] *an Icelandic species of fern* [2] (*Dryopteris assimilis*) [3] *of the three-leaved variety, with feathered leaves, grows in crevices and thickets*

319 different functions are defined for meaning items in ÍO, 22 of which are exploited in the present study. For instance, meaning item [1] for **dílaburkni** provides the relation of **hypernym** to *burkategund* but meaning item [2] is discarded as it does not contain Icelandic lexemes.

All relevant meaning items were tagged for part of speech (POS) with the TnT statistical POS-tagger from Brants (2000), which had been trained on Icelandic data. We only used the word class information from the tagger, except for nouns, where case tags were included as well.

This resulted in 106,977 POS-tagged meaning items. The method showed here was developed assuming that the first POS of a meaning item is an important indicator of what semantic relations are likely to be included in the meaning item and how these can be extracted. The vast majority of the meaning items start with a noun, 68.51%. Of these items 48% consist of only one noun. The second largest group of meaning items comprises those starting with an adjective, 13.73% of all analysed meaning items. All POS-tags were extracted from the items to build POS-patterns. These are named according to the first POS-tag, e.g. patterns extracted from items starting with a noun are called *N_Patterns*.

3 Extraction of semantic relations

After analysing the five most important groups of patterns, including those starting with a noun, an adjective, a pronoun, an adverb and a verb, algorithms for extracting semantic relations were developed. As indicated above, the algorithms begin with the first POS-tag of the definition, in order to narrow down the range of possible semantic relations. Every pattern-group is then analysed in a specific way, searching for POS-patterns or lexicosyntactic patterns indicating a semantic relation (cf. Hearst 1998). One aim of the analysis was to extract as many kinds of relations as possible, so that the success of using this methodology on different relation types could be evaluated. The study was therefore not limited to one or two relations, such as hypernymy and synonymy. All in all, ten relations were extracted, including equivalence and references, which are also marked in the ÍO database. These relations are reviewed next.

3.1 Hypernyms

As described in 2, the typical paraphrasal dictionary definition includes a hypernym of the lemma. Among the patterns indicating a hypernym are:

- (1) adj noun
- (2) adj (, |conj) adj noun
- (3) noun .+

In all cases *noun* represents the hypernym:

spenna ótryggt (adj) ástand (noun)
(*tension precarious situation*)
hypernym (spenna, ástand)

The pattern `noun .+` (i.e. noun plus anything) has several exceptions, where e.g. synonyms or more than one hypernym are extracted.

3.2 Synonyms

Normally, in dictionary definitions of nouns which consist of one noun or a list of nouns, each noun in the definition is a synonym of the lemma:

meiðing barsmíð, líkamsárás
beating-up thrashing, physical assault

The synonyms extracted in this way are rarely absolute synonyms and can have quite different connotations. This is characteristic of the general problem of synonymy. Cruse (1986:88) defines propositional synonymy as the relation between two syntactically identical words which, when interchanged in the same context, will not change the proposition of the corresponding sentence. In WordNet, synonymy is defined as the relation between words that are substitutable in *some* contexts (Miller 1998: 24). Synonymy could thus be seen as propositional synonymy where the substitutability only has to be valid in some contexts.

Despite this broad definition of synonymy, not all definitions that have the format of a synonym definition (cf. Section 2) can be seen as containing synonyms but rather represent a hypernym-hyponym relation:

garg fuglahljóð
screech a bird-sound

In this case an underlying “is-a-kind-of” is not explicitly expressed.

In the ÍO database, the meaning items of definitions in this style are mostly marked as equivalences. This is a misleading term, since equivalences normally hold between words in two different languages.

3.3 Equivalences and references

As stated in 3.2, there is a meaning item in ÍO labelled as “equivalence”. An equivalence should consist either of a noun or a listing of nouns, representing synonyms of the lemma. This prescription has not however been followed consistently in ÍO. Sometimes equivalences represent hyper- or hypernyms and even complex sentences are occasionally marked as equivalences.

ÍO also independently labels “references”, which are meaning items containing one noun or a list of nouns, somehow semantically related to the lemma. Randomly selected references represented hyponymy, hypernymy, antonymy and meronymy.

As equivalences and references are independently labelled in ÍO and are inconsistent in semantic type, they are simply extracted by label (and only if they contain only nouns, which are the target of this study).

3.4 Holonyms and meronyms

Holonyms are extracted where, in a first run, a hypernym *hluti* (‘part of’) has been recognized. In these definitions, the hypernym is rejected and the next noun is extracted as a holonym of the lemma:

fingurgómur fremsti hluti fingurs [...]
finger tip the foremost part of a finger
`holonym(fingurgómur, fingurs)`

The lexico-syntactic pattern indicating meronymy is:

`noun(, noun)*og noun.*`

where all nouns are meronyms of the lemma. The extracted meronyms are of different kinds: ‘X is part of Y’, ‘X and Z build Y’ (*bride and groom* build a *bridal couple*), ‘to be a Y includes being X and Z’ (being a *troubadour* includes being a *poet* and a *musician*). Another kind of relation related to meronymy is the member-group relation. As with holonyms, an extracted hypernym is tested to look for a member-group indication. If it is the word *hópur* (‘group’) it will be rejected and the next noun extracted as the members of the group named by the lemma:

leshringur hópur fólks sem [...]
reading group group of people that [...]
`member(leshringur, fólks)`

3.5 Verbal and adjectival endonyms

Sometimes nouns are paronyms of verbs or adjectives, which in turn constitute endonyms of the corresponding nouns (cf. Cruse 1986). These nouns are often defined in terms of the endonyms.

ihugun (n) íhuga (v)
consideration consider
björgun (n) bjarga (v)
rescue rescue

frægð (n)	frægur (adj)
<i>fame</i>	<i>famous</i>
heiðarleiki (n)	heiðarlegur (adj)
<i>honesty</i>	<i>honest</i>

In some cases the extracted endonym is not morphologically related to the noun, but it still has the analogous semantic relation:

eftirför (n)	elta (v)
<i>chase</i>	<i>chase</i>

The basic patterns for the extraction of endonyms are:

- (1) *það að* verb
that to
- (2) noun adv conj verb
- (3) *það að vera* adj(, adj)*
that to be
- (4) e-ð adj.*
sth.

3.6 Attributes

The extracted attributes do not correspond to attribute slots like SIZE or COLOR; they are in fact attribute values, like *big* or *red*. These provide valuable semantic information and can be used as a basis of differentiation between co-hyponyms. Another benefit of the attributes is the possibility of grouping co-hyponyms that have the same attribute, thus allowing extraction of synonymy or near synonymy that would otherwise have been hidden, as shown in table 1.

Lemma	Attribute	Hypernym
<i>skella</i>	<i>hávær (loud)</i>	<i>stúlka (girl)</i>
<i>glumra</i>	<i>hávær</i>	<i>stúlka</i>
<i>bjalla</i>	<i>hávær</i>	<i>stúlka</i>
<i>heimasæta</i>	<i>ógift (unmarried)</i>	<i>stúlka</i>
<i>yngisstúlka</i>	<i>ógift</i>	<i>stúlka</i>
<i>ungfrú</i>	<i>ógift</i>	<i>stúlka</i>

Table 1: lemmata with the same attribute and the same hypernym can be grouped to build a potential synset

The patterns used to extract attributes are the ones starting with an adjective:

- (1) adj (adj)? noun.*
- (2) adj (, |conj) adj noun.*

3.7 Biological family

Definitions of lemmata from the categories of flora and fauna often include encyclopaedic information additionally to a hypernym and an attribute. From these definitions the name of the biological family of the animate being denoted by the lemma can be extracted:

grænlilja íslensk plöntutegund (*Orthilia secunda*) af vetrarliljuætt
... an Icelandic plant species (*Orthilia secunda*) of the wintery lily family

family(grænlilja, vetrarliljuætt)
hypernym(grænlilja, plöntutegund)
attribute(grænlilja, íslensk)

4 Evaluation

The analysis tool is called “MerkOr”, from Icelandic *merking* (‘meaning’) and *orð* (‘word’). The results of MerkOr’s analysis of ÍO include 116,446 semantic relations between a lemma and a word included in its definition, with equivalence as the most frequent relation extracted. Table 2 shows the extracted relations, ordered by frequency:

Relation	Number extracted
Equivalence	51,390
Hypernym	43,066
Attribute	12,771
Biological family	2,817
Reference	2,140
Endonym - verb	1,286
Synonym	1,201
Meronym	731
Endonym - adj	662
Holonym	382
TOTAL	116,446

Table 2: Extracted relations by frequency

Note, however, that the equivalence and reference relations were extracted by the item number in the ÍO definition and not by pattern matching. If these two relations are excluded then there are eight relations extracted 62,916 times, with hypernyms being the largest group.

The first results of MerkOr are promising. First, from a high percentage of the definitions, at least one semantic relation was extracted. Table 3 shows this data.

	Total	Relation extracted
Definitions	77,348	96.45%
Meaning items	106,977	92.61%

Table 3: Analysed definitions and meaning items

A random selection of 1,034 definitions (about 1.34% of the total) was manually analysed as a gold standard against which the MerkOr results could be tested. The evaluation was run with respect to whole definitions rather than individual meaning items, as this information was thought to be more useful for dictionary makers. MerkOr extracted semantic relations from 957 of the definitions in the gold standard (92.55% extraction rate). The evaluation measures are defined as follows: *correct* indicates that all possible semantic relations are identified and no impossible relations are identified; *partly correct* indicates that some but not all possible relations are identified and also that no impossible relations are identified; *false* indicates that at least one impossible relation is identified. Table 4 shows the test results.

correct	82.13%
partly correct	12.64%
false	5.22%
correct + partly correct	94.77%

Table 4: Accuracy of MerkOr for definitions

All in all 94.77% of the analysed definitions were correctly or partly correctly analysed. This is an encouraging result; the next question, however, is whether this methodology can be extended to free text.

5 Grant of Excellence – a database of semantic relations

Given the limited resources (people and money) in a small language community like Iceland, it is essential to develop LT modules in efficient ways. This is especially true for an extensive project like the development of a semantic database. Existing hand-built resources such as WordNet have been decades in the making; for Icelandic there is little alternative but to adopt and adapt more automated methodologies, such as those outlined above.

Building on these results, we aim to develop methods for extracting semantic relations from unstructured Icelandic texts, using lexico-syntactic patterns. As in the ÍO-project, we will strive for the extraction of both lexical and encyclopaedic relations. Such work requires vast

amounts of tagged text and just such resources are being developed at the Árni Magnússon Institute for Icelandic Studies (Helgadóttir 2004).

The central pattern-based methodology will be extended with other techniques such as latent semantic analysis and coordination information (cf. Cederberg and Widdows 2003, Snow et al. 2005) and tested against established statistical methods for automatic thesaurus construction (cf. Grossman and Frieder 2004).

A tool with a graphical user interface will be developed that allows for manual corrections and extensions of the automatically extracted relations.

6 Conclusions and future work

Nikulásdóttir (2007) shows that automatic extraction of semantic relations from a monolingual dictionary works well for Icelandic. The challenge is to extend this work and test the feasibility of applying a similar approach to free text from a tagged corpus of Icelandic. This will be the task undertaken as part of VLT, the recently-awarded Grant of Excellence. We hope that this work will lay the foundation for the development of a WordNet-like resource for Icelandic.

References

- Agirre, Eneko et al. (2000): Extraction of Semantic Relations from a Basque Monolingual Dictionary using Constraint Grammar. In: *CoRR* (cs.CL/0010025)
- Alshawi, Hiyun (1987): Processing Dictionary Definitions with Phraseal Pattern Hierarchies. In: *Computational Linguistics* Vol. 13, Nr. 3-4, pp. 195-202.
- Brants, Thorsten (2000): TnT – A Statistical Part-of-Speech Tagger. In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, pp. 224-231, Seattle, WA.
- 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)*, pp. 111-118.
- Chodorow, Martin; Roy J. Bird and George E. Heidorn (1985): Extracting Semantic Hierarchies from a Large-On-line Dictionary. In: *Proceedings of the 23rd Annual Meeting of the ACL*, pp. 299-304.
- Cruse, Alan (1986): *Lexical Semantics*. Cambridge: Cambridge University Press.

- Fellbaum, Christine (1998): *WordNet. An Electronic Lexical Database*. Cambridge Mass., London: MIT Press.
- Geeraerts, Dirk (2003): Meaning and Definition. In: Piet van Sterkenburg (ed.): *A Practical Guide to Lexicography*. Amsterdam, Philadelphia: John Benjamins, pp. 83-93.
- Grossman, David A. and Ophir Frieder (2004): *Information Retrieval: Algorithms and Heuristics*. Second Edition. Berlin: Springer.
- Hearst, Marti A. (1998). Automated discovery of WordNet relations. In: Christiane Fellbaum (ed): *WordNet. An Electronic Lexical Database*. Cambridge Mass., London: MIT Press, pp.131-151.
- Helgadóttir, Sigrún. (2004). Mörkuð íslensk málheild. [A Tagged Icelandic Corpus]. In Samspil tungu og tækni. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 67-71.
- Íslensk orðabók* (2002). Mörður Árnason (ed.). Reykjavík: Edda.
- Markowitz, Judith; Thomas Ahlswede and Martha Evens (1986): Semantically Significant Patterns in Dictionary Definitions. In: *Proceedings of the 24th Annual Meeting of the ACL*, pp. 112-119.
- Nakamura, Junichi and Makoto Nagao (1988): Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. *Proceedings of the Twelfth International Conference on Computational Linguistics*, pp. 459-464.
- Nikulásdóttir, Anna Björk (2007): *Automatische Extrahierung von semantischen Relationen aus einem einsprachigen isländischen Wörterbuch*. M.A.-Thesis, University of Heidelberg, Germany.
- Rögnvaldsson, Eiríkur (2008). Icelandic Language Technology Ten Years Later. In Collaboration: *Interoperability between People in the Creation of Language Resources for Less-resourced Languages*, pp. 1-5. SALTMIL workshop, LREC 2008. Marrakech.
- Snow, Rion; Daniel Jurafsky and Andrew Y. Ng (2006): Semantic Taxonomy Induction from heterogeneous Evidence. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 801-808, Sydney.
- Wiegand, Herbert Ernst (1989): Die lexicographische Definition im allgemeinen einsprachigen Wörterbuch. In: Franz Josef Hausmann et al. (eds.): *International Encyclopedia of Lexicography*: 003. Berlin, New York. (Handbooks of Linguistics and Communication Science 5.1) pp. 530-588.