

Automatic Extraction of Semantic Relations

Anna Björk Nikulásdóttir
University of Iceland
abn@hi.is

Matthew Whelpton
University of Iceland
whelpton@hi.is



[Outline]

1. Introduction
2. The Icelandic monolingual dictionary
Íslensk orðabók (ÍO)
3. Automatic extraction of semantic relations from ÍO
4. A semantic network with semantic mining



[NLP and semantics in Iceland]

- No existing specifically LT-oriented semantic resources for Icelandic
- A rich lexicographic tradition
- One pilot study in extracting semantic relations from an Icelandic dictionary



[NLP project in progress]

- The Icelandic Research Fund “RANNÍS” funds and has funded NLP research in Iceland
- A Grant of Excellence has been awarded to the project “Viable Language Technology beyond English – Icelandic as a test case”
- Within this project a work package aiming at the development of a semantic network for Icelandic



[Outline]

1. Introduction
2. The Icelandic monolingual dictionary
Íslensk orðabók (ÍO)
3. Automatic extraction of semantic relations from ÍO
4. A semantic network with semantic mining



[Definitions in ÍO]

- The definition vocabulary in ÍO is unrestricted
- The format of the definitions is mostly either
 - synonym definition
 - or
 - a paraphrase, including a hypernym with distinguishing features



Definitions in ÍO

- synonym definition:

fagnaður ánægja, gleði

joy pleasure, gladness

- a paraphrase:

breiðband breitt tíðnisvið notað til fjarskipta [...]

broadband a broad frequency range used for telecommunication [...]



[Meaning items]

- Meaning item: one part of a definition with a certain lexicographic function or role
- The main purpose of the meaning items in ÍO is to support lexicographers and publishers
- Eases the machine access to different parts of the definitions



[Meaning items]

dílaburkni [1] íslensk burknategund [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*



[Outline]

1. Introduction
2. The Icelandic monolingual dictionary
Íslensk orðabók (ÍO)
3. Automatic extraction of semantic relations from ÍO
4. A semantic network with semantic mining



[Preparation of data]

- Definitions filtered for relevant meaning items
- All relevant meaning items tagged for part of speech (POS) with the TnT statistical POS-tagger from Brants
- Result: 106,977 POS-tagged meaning items



[POS-patterns and rules]

- POS-patterns extracted from the tagged meaning items:

spenna ótryggt(a) ástand(n)

tension precarious situation

pattern: a_n

rule: a_n → hypernym(lemma, n)
attribute(lemma, a)



[POS-patterns]

pattern starts with ...	number	%
noun	73,290	68.51%
adjective	14,684	13.73%
pronoun	7,362	6.89%
adverb / prep.	4,585	4.29%
verb	3,447	3.22%



Extracted relations

Relation	Number
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



Biological family

dílaburkni [1] íslensk burknategund [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

family (dílaburkni, þrílaufungsætt)

hypernym (dílaburkni, burknategund)

attribute (dílaburkni, íslensk)



[Endonyms]

eftirför það að elta
chase (that) to chase

verbalEndonym (eftirför, elta)

frægð það að vera frægur
fame (that) to be famous

adjEndonym (frægð, frægur)



[Results]

	Total	Relation extracted
Definitions	77,348	96.45%
Meaning items	106,977	92.62%

correct	82.13%
partly correct	12.64%
false	5.23%
correct + partly correct	94.77%



[Linked results]



[Outline]

1. Introduction
2. The Icelandic monolingual dictionary
Íslensk orðabók (ÍO)
3. Automatic extraction of semantic relations from ÍO
4. A semantic network with semantic mining



[Pattern generation]

- A pattern-based approach aiming at finding patterns in the style of *Hearst-patterns*
- A tagged corpus of Icelandic texts is being developed at The Árni Magnússon Institute for Icelandic studies (Helgadóttir, 2004)
- A part of this corpus has been chunked with IceParser (Loftsson and Rögnvaldsson, 2007)
- From this data syntactic patterns, including NPs and PPs, have been extracted



[Pattern generation]

- A reasonable amount of the extracted patterns will be checked manually for relation indication
- No pre-defined list of relations
- Lexical and encyclopaedic relations



[Additional methods]

- Extension of the central pattern-based methodology with other techniques
- Statistical methods as well as further symbolic methods like coordination information



References

- Hearst, Marti A. (1998): Automated Discovery of WordNet Relations. 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]. *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 67-71
- Loftsson, Hrafn and Eiríkur Rögnvaldsson (2007): IceNLP: A Natural Language Processing Toolkit for Icelandic. *Proceedings of InterSpeech 2007, Special Session: "Speech and language technology for less-resourced languages."* Antwerp.
- Nikulásdóttir, Anna Björk (2007): *Automatische Extrahierung von semantischen Relationen aus einem einsprachigen isländischen Wörterbuch*. M.A.-Thesis, University of Heidelberg, Germany

