
A Semantic Database for Icelandic Language Technology

Anna Björk Nikulásdóttir
University of Iceland



Outline

- ▶ The project *A semantic network with semantic mining*
- ▶ Automatic extraction of semantic relations:
 - ▶ Extraction from dictionaries
 - ▶ Extraction from text
 - ▶ Combination of results
- ▶ Outlook



A Semantic Network with Semantic Mining

- ▶ A work package within the project *Viable Language Technology beyond English – Icelandic as a Test Case*
- ▶ Aims at a semi-automatic construction of a semantic database for Icelandic language technology
- ▶ Two main components:
 - ▶ collection of data through automatic extraction of semantic relations (in progress)
 - ▶ exploiting of possibilities to semi-automatically combine the extracted relations and other available resources to build a semantic database (not started)



A Semantic Network with Semantic Mining

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



Outline

- ▶ The project *A semantic network with semantic mining*
- ▶ Automatic extraction of semantic relations:
 - ▶ **Extraction from dictionaries**
 - ▶ Extraction from text
 - ▶ Combination of results
- ▶ Outlook



Dictionaries and dictionary definitions

- ▶ Standard monolingual dictionaries include semantic information about a considerable part of the vocabulary of a language
- ▶ This information is not directly accessible by computers
- ▶ Dictionary definitions have relatively fixed syntactic structures
- ▶ These structures make it feasible to use *lexico-syntactic patterns* to extract information from the definitions



Lexico-syntactic patterns

Agar is a substance prepared from a mixture of **red algae**, **such as Gelidium**, for laboratory or industrial use.

a. NP_0 *such as* NP_1 {, NP_2 ..., (*and|or*) NP_i } $i \geq 1$

b. for all NP_i , $i \geq 1$, HYPONYM (NP_i , NP_0)

HYPONYM (**Gelidium**, **red algae**)

read: *Gelidium* is a hyponym of *red algae*

(Hearst, 1998)



Extraction from the Icelandic Dictionary

Patterns and relations:

spenna ótryggt(a) ástand(n)

tension *precarious situation*

pattern: a_n

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



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



Outline

- ▶ The project *A semantic network with semantic mining*
- ▶ Automatic extraction of semantic relations:
 - ▶ Extraction from dictionaries
 - ▶ **Extraction from text**
 - ▶ patterns identified by using seed words
 - ▶ evaluation of most common patterns
 - ▶ statistical methods
 - ▶ **Work in progress: preliminary results only!**
 - ▶ Combination of results
- ▶ Outlook



Pattern identification by seed words

- ▶ Words that are known to be in a certain relation are used to identify syntactic or lexico-syntactic patterns
- ▶ Such patterns are often called *Hearst patterns*
- ▶ Example: *England* is known to be a hyponym of *country*. Extract these words occurring together in a parsed corpus and examine their lexico-syntactic environment

$NP_0 \{, \}$ especially $\{NP_1, NP_2 \dots, (and \mid or)\} NP_n$

... most European countries, especially France, England, and Spain

(Hearst, 1992)



Properties of Hearst patterns

- ▶ Hearst patterns are reliable
- ▶ Vast amount of data is needed, because:
 - ▶ one wants the seed words to occur often, at the best in different patterns
 - ▶ Hearst patterns occur relatively seldom
- ▶ Icelandic, being a less-resourced language, does not yet have such amount of prepared data
- ▶ To deal with this sparse data problem, an experiment is being made to harvest patterns in another way



Evaluation of most common patterns

- ▶ Syntactic patterns including NPs and PPs were extracted from a parsed corpus (tagged and parsed using *IceNLP*)
- ▶ Every occurrence of each pattern was saved with the actual text
- ▶ The textual examples of the most common patterns were loosely examined manually to see if they generally represented some semantic relation – no predefined set of semantic relations was used
- ▶ The selected patterns are being tested



Pattern nr: 9/500

1893 examples

COORD. NOUN



[NPs[NP nxen]][CP og c][NP nxen]

coord. noun

Assign relation

Note:

Save note

[NPs [NP Gestrismi nven NP] [CP og c CP] [NP rausn nven NP] NPs]
 [NPs [NP Geymsla nven NP] [CP og c CP] [NP vinnurými nhen NP] NPs]
 [NPs [NP Gjaldþrot nhen NP] [CP og c CP] [NP nauðungaruppböð nhen NP] NPs]
 [NPs [NP Glaðlyndi nhen NP] [CP og c CP] [NP hæverska nven NP] NPs]
 [NPs [NP Glaðlyndi nhen NP] [CP og c CP] [NP hæverska nven NP] NPs]
 [NPs [NP Glaðværð nven NP] [CP og c CP] [NP lífsgleði nven NP] NPs]
 [NPs [NP Glaðværð nven NP] [CP og c CP] [NP lífsgleði nven NP] NPs]
 [NPs [NP Gleði nven NP] [CP og c CP] [NP gaman nhen NP] NPs]
 [NPs [NP Gleði nven NP] [CP og c CP] [NP gaman nhen NP] NPs]
 [NPs [NP Gleði nven NP] [CP og c CP] [NP kátína nven NP] NPs]
 [NPs [NP Gleði nven NP] [CP og c CP] [NP kátína nven NP] NPs]
 [NPs [NP Glimmer nhen NP] [CP og c CP] [NP glæsileikiundirfyrirsögn nven NP] NPs]
 [NPs [NP Glimmer nhen NP] [CP og c CP] [NP glæsileikiundirfyrirsögn nven NP] NPs]
 [NPs [NP Greindarskortur nken NP] [CP og c CP] [NP vitsmunavanþroski nken NP] NPs]
 [NPs [NP Greiðvikni nven NP] [CP og c CP] [NP tillitssemi nven NP] NPs]
 [NPs [NP Greiðvikni nven NP] [CP og c CP] [NP tillitssemi nven NP] NPs]
 [NPs [NP Grænmeti nhen NP] [CP og c CP] [NP salat nhen NP] NPs]
 [NPs [NP Gróðurfar nhen NP] [CP og c CP] [NP tegundasamsetning nven NP] NPs]
 [NPs [NP Gæfa nven NP] [CP og c CP] [NP framtíð nven NP] NPs]
 [NPs [NP Gæsa húð nven NP] [CP og c CP] [NP hrollur nken NP] NPs]
 [NPs [NP Gæska nven NP] [CP og c CP] [NP góðvild nven NP] NPs]
 [NPs [NP Gíraffi nken NP] [CP og c CP] [NP krókódill nken NP] NPs]
 [NPs [NP Gíraffi nken NP] [CP og c CP] [NP krókódillundirfyrirsögn nven NP] NPs]
 [NPs [NP Gíró- nven NP] [CP og c CP] [NP kreditkortþjónusta nven NP] NPs]
 [NPs [NP Gíró- nven NP] [CP og c CP] [NP kreditkortþjónusta nven NP] NPs]
 [NPs [NP Gíró- nven NP] [CP og c CP] [NP kreditkortþjónusta nven NP] NPs]
 [NPs [NP Gíró- nven NP] [CP og c CP] [NP kreditkortþjónusta nven NP] NPs]
 [NPs [NP HRAÐI nken NP] [CP og c CP] [NP spennan nven NP] NPs]
 [NPs [NP Hafsteinn nken NP] [CP og c CP] [NP fjölskylda nven NP] NPs]
 [NPs [NP Hafsteinn nken NP] [CP og c CP] [NP fjölskylda nven NP] NPs]
 [NPs [NP Hagsýni nven NP] [CP og c CP] [NP skynsemi nven NP] NPs]
 [NPs [NP Hagvöxtur nken NP] [CP og c CP] [NP samruni nken NP] NPs]
 [NPs [NP Hagyrðinga- nhen NP] [CP og c CP] [NP skemmtikvöld nhen NP] NPs]
 [NPs [NP Handavinnusýning nven NP] [CP og c CP] [NP sala nven NP] NPs]
 [NPs [NP Handverk nhen NP] [CP og c CP] [NP hönnun nven NP] NPs]
 [NPs [NP Handverk nhen NP] [CP og c CP] [NP hönnun nven NP] NPs]
 [NPs [NP Harmonikuleikur nken NP] [CP og c CP] [NP söngur nken NP] NPs]
 [NPs [NP Harðfiskur nken NP] [CP og c CP] [NP hákarl nken NP] NPs]
 [NPs [NP Heilbrigði nhen NP] [CP og c CP] [NP heilsuvernd nven NP] NPs]
 [NPs [NP Heilbrigðis- nken NP] [CP og c CP] [NP tryggingamálaráðherra nken NP] NPs]

Some preliminary results

- ▶ First tests have been made with patterns possibly representing hypernymy, co-hyponymy / coordinated nouns and relations of properties (e.g. *has-a*) / genitive constructions
- ▶ The tests are currently run on a corpus of 8.8 million tokens, the same corpus the patterns were extracted from
- ▶ For the final test and evaluation phase another bigger corpus has to be prepared



Some preliminary results

- ▶ **Hypernymy:**
 - ▶ 175 extracted noun pairs
 - ▶ very reliable
- ▶ **Co-hyponymy / coordinated nouns:**
 - ▶ 6,176 extracted comma-separated noun chains
 - ▶ 20,863 extracted noun pairs connected with *og* | *eða* (*and* | *or*)
- ▶ **Genitive constructions:**
 - ▶ 58,018 extracted noun pairs



Genitive constructions - properties

- ▶ The results express polysemy of many terms
- ▶ *body* as an organism (prop.: *cell*) or as a thing one can put on the market (*marketization*)
- ▶ *cod* as a fish (*stomach*) or as a product (*market price*)
- ▶ *house* as a building (*roof*), as a property (*running*), as a theatre (*theatre consultant*), or as a restaurant / pub (*band*)



Statistical methods: semantic similarity and clustering

- ▶ Semantic similarity: „You shall know a word by the company it keeps“ (Firth, 1957: 11)
- ▶ Computing of semantic similarity:
 - ▶ 900 most common content-bearing words used as *context words*
 - ▶ Occurrences of nouns in the corpus counted where they appeared max. 12 words before or 12 words after some context word, i.e. co-occurs with a context word
 - ▶ Final number of occurrences for each noun weighted using a logarithmic weighting function (Manning and Schütze, 1999):

$$w(\text{nrOfOccurrences}) = 1 + \log_{10}(\text{nrOfOccurrences})$$

- ▶ Co-occurrence values



Semantic similarity

- ... hafa sérgeymslur inni í **íbúðunum**, það með aðstöðu fyrir þvottavél, auk sameiginlegs **þvottahúss** og sameiginlegrar geymslu ...
- ... **íbúðirnar** afhendast fullbúnar án gólfefna en **baðherbergi** er flísalagt ...

íbúð (*a flat, an apartment*): a context word

þvottahús (*laundry room*) and **baðherbergi** (*bathroom*) share similar context



k-means Clustering

- ▶ Clustering is a method to form groupings of data
- ▶ Unsupervised learning
- ▶ *k-means* clustering used to group nouns by similarities of their vectors of co-occurrence values
- ▶ *Cosine similarity measure* used to compare vectors

$$sim(x,y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} * \sqrt{\sum_{i=1}^n y_i^2}}$$



Preliminary results

- ▶ 7,871 nouns grouped into 60 clusters of max. 200 words
- ▶ 46 clusters can be characterized by a concept, some of which are: RELATIVES, FOOTBALLERS, FINANCES, MUSIC/MUSICIANS, FISH/FISHERY, BIOCHEMISTRY, HOUSE, FILM / ACTORS, VEHICLES, BIOLOGY, SCHOOLS/EDUCATION, POLITICS, SPORT, ... etc.



Cluster HOUSE

baðherbergi, hjónaherbergi, þvottahús, borðstofa, svefnherbergi, sturtuklefi, hol, forstofa, flís, borðkrókur, bílskúr, barnaherbergi, sólpallur, fataskápur, sérinngangur, sturta, parket, ris, salerni, jarðhæð, bað, anddyri, uppvottavél, snyrting, gólfefni, hólf, vistarvera, lofthæð, verönd, dúkur, teppi, stigagangur, setustofa, trappa, ferm, þvottavél, raðhús, bakgarður, fm, marmari, skúffa, álma, timburhús, steinhús

bathroom, master bedroom, laundry room, dining room, bedroom, shower cubicle, hall, entrance hall, tile, dining nook, garage, child's bedroom, sun deck, wardrobe, private entrance, shower, parquet, attic, etc. ...



Outline

- ▶ The project *A semantic network with semantic mining*
- ▶ Automatic extraction of semantic relations:
 - ▶ Extraction from dictionaries
 - ▶ Extraction from text
 - ▶ **Combination of results**
- ▶ Outlook



Combination of results

- ▶ Use semantic similarity to verify extracted relations (Dominic and Widdows, 2003)
- ▶ Use coordination information to extend results (Dominic and Widdows, 2003)
- ▶ Connect results from different extraction procedures from text and dictionaries
- ▶ Use information about semantic features to extend results



Verification of extracted hypernyms

- ▶ $\text{sim}(\text{tónlistarflutningur} - \text{atriði}) = 0.3593$
musical performance – point, item, scene, number (in a show)
- ▶ $\text{sim}(\text{veikindi} - \text{áfall}) = 0.6604$
illness – shock
- ▶ $\text{sim}(\text{þorskur} - \text{botnfiskur}) = 0.4921$
cod – bottom dweller
- ▶ $\text{sim}(\text{fiskur} - \text{sjávarafurð}) = 0.6463$
fish – fishery product
- ▶ $\text{sim}(\text{þröstur} - \text{fugl}) = 0.4923$
thrush - bird
- ▶ $\text{sim}(\text{morð} - \text{glæpur}) = 0.5165$
murder - crime



Verification of extracted properties

▶ $\text{sim}(\text{líkami} - \text{fruma}) = 0.7435$

body - cell

▶ $\text{sim}(\text{dísilvél} - \text{útblástur}) = 0.3263$

diesel engine - exhaust

▶ $\text{sim}(\text{gróðurhúsalofttegund} - \text{útblástur}) = 0.5092$

greenhouse gas - exhaust

▶ $\text{sim}(\text{koltvísýringur} - \text{útblástur}) = 0.3944$

carbon dioxide - exhaust

▶ $\text{sim}(\text{frumskógur} - \text{málari}) = 0.2676$

jungle - painter



Coordination information

- ▶ Don't subsume coordinated terms under a too narrow hypernym:

- ▶ Extracted hypernym for *þorskur* (cod):

þorskur IS-A botnfiskur (bottom dweller)

- ▶ Extracted noun coordination information for *þorskur*:

þorskur, lax, bleikja, langa, skötuselur, steinbítur, ýsa, loðna, síld, flatfiskur, karfi, ufsi, grálúða, barri, hlýri, saltfiskur, makrill
cod, salmon, river trout, ling, ... , catfish, haddock, capelin, ..., flatfish ..., salt fish, ...



Coordination information

- ▶ Choose seed-words carefully:
- ▶ Extracted hypernym for *fugl*:
fugl IS-A *dýralíf* (bird – animal life)
- ▶ Extracted noun coordination information for *fugl*:
fugl, forgjöf, þar (birdie, handicap, þar)



Semantic features information

- ▶ Can relations like „X HAS *beginning*“ be automatically extended to „X HAS *end*“?



Automatic extension of properties?

Extracted information:

The beginning of ...

- ▶ a century
- ▶ an aria
- ▶ a book
- ▶ a town
- ▶ the world
- ▶ a marriage

Possible extensions:

The end of ...

- ▶ *a century*
- ▶ *an aria*
- ▶ *a book*
- ▶ *a town?*
- ▶ *the world?*
- ▶ *a marriage?*



Extension of properties to co-hyponyms?

- ▶ Can properties of a term automatically be extended to its co-hyponyms?
- ▶ Example of *cod* and the polysemy *fish* vs. *product*.
Extension of properties that belong to the meaning *fish* cannot be extended to co-hyponyms in the meaning *product*, like *salt fish*.



Outlook

- ▶ The preliminary results of the relation extraction look promising
- ▶ Necessary to exploit possibilities to verify, combine and extend results
- ▶ Include results from dictionary and encyclopaedias (e.g. *wikipedia*)
- ▶ Next steps include further development of extraction algorithms as well as the preparation of more data



Icelandic Language Technology

- ▶ <http://www.tungutaekni.is>
- ▶ <http://iceblark.wordpress.com/>
- ▶ <http://nlp.ru.is>
- ▶ <http://sourceforge.net/projects/icenlp/>
- ▶ <http://www.linguist.is/>



References

- ▶ Cederberg, Scott and Dominic Widdows (2003): Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. *Proceedings of CoNLL*, pp. 111-118.
- ▶ Firth, J.R. (1957): *Papers in Linguistics 1934-1951*. London: Oxford University Press.
- ▶ Hearst, Marti A. (1998): Automated Discovery of WordNet Relations. Ch. Fellbaum (ed.): *WordNet. An Electronic Lexical Database*. Cambr. Mass, London : MIT Press, pp. 131-151
- ▶ Hearst, Marti A. (1992): Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of COLING-1992*, Nantes, France, pp. 539-545
- ▶ Manning, Christopher and Hinrich Schütze (1999): *Foundations of Statistical Natural Language Processing*. Cambr. Mass., London : MIT Press
- ▶ Nikulásdóttir, Anna and Matthew Whelpton (2009): Automatic Extraction of Semantic Relations For Less-Resourced Languages. In K. Jokinen & E. Bick (eds) *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009 (NEALT PROCEEDINGS SERIES VOL. 4)*. May 14-16, 2009. Odense, Denmark.

