Íslenskur Orðasjóður - Building a Large Icelandic Corpus

Erla Hallsteinsdóttir

erlahall@yahoo.dk

Thomas Eckart, Chris Biemann,
Uwe Quasthoff, Matthias Richter
Natural Language Processing Group
University of Leipzig
Johannisgasse 26, 04103 Leipzig, Germany
{teckart,biem,quasthoff,mrichter}
@informatik.uni-leipzig.de

Abstract

We introduce an Icelandic corpus of more than 250 million running words and describe the methodology to build it. The resource is available for use free of charge. We provide automatically generated monolingual lexicon entries, comprising frequency statistics, samples of usage, co-occurring words and a graphical representation of the word's semantic neighbourhood.

1 Introduction

Corpora are important language resources for a variety of Natural Language Processing tasks, especially in semi-supervised settings, where corpora are used to build e.g. language models. In (Biemann et al., 2004) and (Quasthoff et al., 2006) design and implementation of an architecture capable of building numerous of large corpora for different languages with little manual effort have been introduced. This infrastructure has been used to produce an Icelandic corpus based on web pages. In this paper we present the steps undertaken to digest unstructured clutter of HTML pages into a ready for use linguistic resource that provides rapid access through search indices on different language units.

2 Motivation

Icelandic is considered a small language with around 300,000 native speakers. Most linguistic research on Icelandic has been highly dominated by ideological premises of a linguistic purism driven by the mission of the preservation of the ancient Icelandic language (cf. Kristmannsson, 2004). But, despite a politically initiated language technology campaign in order to strengthen an effective language purity policy, there still has been only little empirical research on the Icelandic language, partially due to the lack of a large corpus. Recent endeavours, as e.g. used by (Helgadóttir, 2004) or the search interface of Iceland's Lexicological Institute[1] operate on small-scale corpora. Here, our data basis is the entirety of all Icelandic websites officially collected by the National and University Library of Iceland.

The corpus "Íslenskur Orðasjóður" is first of all a comprehensive corpus-based lexicon of modern Icelandic. Its data on current language use is also an important and long needed basis for empirical linguistic research (e.g. lexical research on neologisms, the use of word-formation rules, foreign lexical units), computational linguistics and applications in language technology. Furthermore the corpus provides a basis for diachronic comparison as it documents are a to-date snapshot of Icelandic.

3 Data Cleaning

Automated collection of text cannot avoid unwanted junk in the raw data. We apply a heuristic means of separating well-formed from inappropriate sentences. The overall procedure is the same for each language.

3.1 Text Extraction

The National and University Library of Iceland has kindly provided its internet archive of web pages from all .is domains as text basis. Our starting point was approximately 120 GB of zipped

[1] http://www.lexis.hi.is/corpus/leit.pl

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 288–291

Rule	Description	Examples	Hits
too many periods	unseparated sentences gluing words together or incomplete sentences ending with "…"	Upp í flugvél, burt úr kuldanum.....	1,300,000
link artifacts or |	navigation boilerplates	Example: Forsíða > Túlkanir og þýðingar > Þýðingar Heim | Hafa samband | Veftré Leitarvél: Alþjóðahús Gagnlegar upplýsingar Algengar	220,000
begins with number dot blank	enumeration items	1. innkaup hlutu: Gláma/Kím arkitektar ehf., Laugavegi 164.	200,000
too many capital letters or digits in a row	headlines glued together with sentences or enumerations	LEIÐBEININGAR UM NOTKUN Gríptu um borðana og togaðu niður og í sundur. 7.3.2005 Tilkynning frá Högum hf. 7.3.2005 Verslunarrekstur Skeljungs komin til 10-11 25.10.2004 Tilkynning frá Högum hf. 22.6.2004 Tilkynning (...)	198,000
contains too many ":"s	Lists, e.g. of sports results	steini :: Comment :: 10 hugmyndir af bloggi.	166,000
too many {/&:}s	itemizations	Ferðaönd - Svara - Vitna í - Stelpið 31/10/05 - 0:25 Soffía frænka - Svara - Vitna í - aulinn 31/10/05 - 8:39 Kona í bleikum slopp með rúllur í hárinu.	153,000
expression too short	incomplete sentences	10. Valur ? _\åv,c ?	100,000
too many "_"s in a row	clozes	a) _____, b) _____ og c) _____ Hvað myndast í kynhirslunum að lokum?	58,000

Table 1: Text cleaning rules used for dropping undesired sentences, their rationale and impact.

HTML pages downloaded by the library in November 2005. From these pages the document text has been extracted and segmented into sentences leaving 29,718,528 unique sentences after duplicate sentence removal.

3.2 Language Identification

To reliably detect different languages, high text coverage can be achieved with the 10,000 most frequent words per language. We use these words extracted from previously built corpora for sentence-based language identification, see e.g. (Dunning, 1994) for an overview.

From the 29,718,528 unique sentences, merely 19,112,187 of them were identified as Icelandic. The remainder was predominantly identified as English (about 4 million sentences).

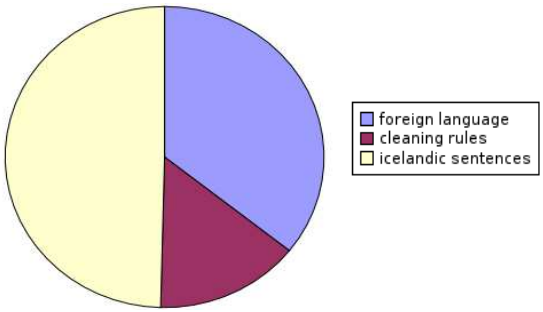

Figure 1: Fractions of the raw material as determined by the cleaning procedures

3.3 Text Cleaning

A set of 65 rules (SQL-statements) was used to get rid of unwanted, ill-formed sentences. Table 1 contains a description of the rules with highest coverage in natural language, their rationale and the number of sentences affected here. Eventually 14.742.802 sentences make up the remaining corpus, approximately half the amount of the raw material. Figure 1 shows the fractions of the raw material removed in the cleaning procedures.

4 Corpus Building

From the remaining sentences, a full form dictionary is computed. Currently, we support two language units in our corpora: word forms and sentences. For sentences, we provide information from which source (e.g. webpage) the respective sentence was obtained, yet we do not store full documents but only well-formed sentences extracted from these (see Section 3). Sentences are indexed by words, so it is possible to quickly access all sentences a word form occurs in. For each word form, we automatically extract the following information: frequency (which indicates common vs. uncommon words) and significant co-occurrences based on neighbouring words (mostly containing typical properties of the corresponding concept or idioms the word is a part of) and based on sentences (mostly containing semantically related word forms). As significance measure, the log-likelihood ratio (cf. Dunning, 1993) is used. Our storing scheme is open for additional, manually provided information, such as grammatical data, lemmatization, thesaurus entries and subject area assignments. *Íslenskur Orðasjóður* contains mappings from word forms to lemmas (source: Bjarnadóttir (2004)), see Appendix.

5 Example Data

Frequency statistics can be used for measuring visibility of a concept and commonness of a word. In lexicography this can be an important criterion to explain judgments. For the language researcher this data can tell whether and to what extent a word was in use in the time interval covered by the corpus. Following are the 100 most frequent words in our Icelandic corpus in descending order by frequency: *og, að, í, á, sem, er, til, við, um, var, en, með, fyrir, ekki, því, það, af, hann, eru, hefur, frá, verið, þar, hafa, ég, eftir, þess, sér, Það, þegar, þá, segir, kl, svo, hún, upp, voru, hafi, eða, sé, úr, fram, verður, Í, þeim, hjá, þeirra, eins, Ég, nú, hans, Hann, þeir, sagði, þetta, sig, út, vera, þau, vel, væri, Við, yfir, okkur, vegna, mjög, okkar, mér, f, allt, ára, Á, einnig, koma, þessu, þó, verði, hér, kom, hefði, ár, hafði, saman, hennar, Þetta, þú, sínum, verða, undir, tíma, Íslands, þessum, En, alltaf, Hún, mun, gera, mikið, dag, má.*

Co-occurrence data is meant to be used extensively as a building block for further applications such as word sense disambiguation, extraction of semantic relations and building of ontologies. For the dictionary user and further uses we provide also networks of co-occurrences like depicted in the Appendix. They can be used for example as a navigation aid or for building topic maps.

The Appendix contains the information as displayed on the corpus website. All words occurring in the corpus are linked to their respective entry, so it is possible to navigate through the resource by clicking on the words. Alternatively, a search mask can be used (not shown).

6 Future Work

With the basic language resource available we aim at including all available types of additional information such as dictionary data, part of speech etc. to the Icelandic language resource. This work is also part of a long term plan to provide language resources in comparable format in the Leipzig Corpora Collection at `http://corpora.uni-leipzig.de`, counting 17 languages at the time writing.

References

Chris Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. 2004. *Language independent Methods for Compiling Monolingual Lexical Data*. In Proceedings of CIcLING, Springer LNCS 2945.

Bjarnadóttir, Kristín. 2004. *Beygingarlýsing íslensks nútímamáls*. Version 2.0, 30. November 2004.

Ted Dunning. 1993. *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, Volume 19, number 1 .

Ted Dunning. 1994. *Statistical identification of language*. In: Technical report CRL MCCS-94-273, New Mexico State University. Computing Research Lab.

Sigrún Helgadóttir. 2004. *Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic*. Nordisk Sprogteknologi. Årbog 2004. Kopenhagen

Gauti Kristmannsson 2004. *Iceland's "Egg of Life" and the Modern Media*. In: Meta, XLIX, 1, 2004, 59-66.

Uwe Quasthoff, Matthias Richter, and Chris Biemann. 2006. *Corpus Portal for Search in Monolingual Corpora*. In: Proceedings of the LREC 2006.

Appendix: Sample Entry from `http://wortschatz.uni-leipzig.de/ws_ice/`

orðmynd: orðabók (English: dictionary)
tíðni: 120
tíðniflokkur: 13 (þ.e. *og* kemur 2^{13} oftar fyrir en þessi orðmynd)

beygingarmyndir: orðabók, orðabókar, orðabókin, orðabókina, orðabókinni, orðabókarinnar, orðabækur, orðabókum, orðabóka, orðabækurnar, orðabókunum, orðabókanna

dæmi:
Ég nefni fyrst Íslenska **orðabók**. (heimild: *Newspaper*)
Einnig fylgir lítil **orðabók** þýðanda með skýringum. (heimild: *Newspaper*)
Að gefast upp eða tapa, - það var ekki til í hans **orðabók** eða fasi. (heimild: *Newspaper*)
fleiri dæmi

orð með háa tíðni sem nágrannar orðabók:
Menningarsjóðs (85), Íslenskri (64), Íslensk (42), Orðabók (23), Mörður (21), ritstjórn (20), Marðar (20), Í (17), Orðastað (17), lýsingarorðið (15), heiðinn (15), orðabækur (14), orð (14), Háskólans (14), þreyja (13), Árnasonar (13), stórfiskaleikur (13), prentútgáfna (13), lýðveldistímans (13), klyfberi (13), Bókaútgáfu (13), ÍSLENSK (12), uppgjöf (12), delicious (12), útgáfudegi (11), merking (11), lsquo (11), gefast (11), eða (11), Freysteins (11), orðsins (10), orðið (10), orðinu (10), merkir (10), hugum (10), færeyskt (10), fletta (10), ekki (10), dægrastytting (10), Grunnavík (10), Örlygs (9), Íslenska (9), syndrome (9), glöggva (9), forsölu (9), bók (9), Árna (8), viðhorfa (8), slangur (8), samkvæmt (8), samanlögðu (8), ríkjandi (8), nýrri (8), metsölubók (8), heimspekideild (8), forrit (8), endurbætt (8), Blöndals (8), nefni (7), merkingar (7), keyptum (7), er (7), alist (7), Starfaði (7), Orðið (7), Böðvarssonar (7), íslenskri (6), Ö (6), skýringum (6), selst (6), samantekt (6), lektor (6), hinni (6), hin (6), gefin (6), eintök (6), dósent (6), Færeyingar (6), íslensku (5), íslenska (5)

orð með háa tíðni sem vinstri nágrannar orðabók:
Íslenskri (64), Íslensk (34), ÍSLENSK (12), íslenskri (4), Úr (4), Íslenska (4), íslenska (3), samkvæmt (3)

orð með háa tíðni sem hægri nágrannar orðabók:
Menningarsjóðs (50), Freysteins (11), Blöndals (8), ríkjandi (5), Háskólans (5)