A 500 Million Word POS-Tagged Icelandic Corpus

Thomas Eckart[1], Erla Hallsteinsdóttir[2], Sigrún Helgadóttir[3], Uwe Quasthoff[1], Dirk Goldhahn[1]

[1]Natural Language Processing Group, University of Leipzig, Germany

[2] Department of Language and Communication, University of Southern Denmark, Odense, Denmark

[3] The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland

Email: {teckart, quasthoff, dgoldhahn}@informatik.uni-leipzig.de, erla@sdu.dk, sigruhel@hi.is

Abstract

The new POS-tagged Icelandic corpus of the Leipzig Corpora Collection is an extensive resource for the analysis of the Icelandic language. As it contains a large share of all Web documents hosted under the *.is* top-level domain, it is especially valuable for investigations on modern Icelandic and non-standard language varieties. The corpus is accessible via a dedicated web portal and large shares are available for download. Focus of this paper will be the description of the tagging process and evaluation of statistical properties like word form frequencies and part of speech tag distributions. The latter will be in particular compared with values from the Icelandic Frequency Dictionary (IFD) Corpus.

Keywords: Corpus Creation, Part-of-Speech Tagging, Grammar and Syntax

1. History of the Icelandic Corpus

Larger Icelandic corpora have been part of the Leipzig Corpora Collection (LCC) since 2005. The aim of the project is to generate large monolingual corpora based on various material of different genre, where the biggest resources are Web texts provided by the National and University Library of Iceland from autumn 2005 and autumn 2010 (approx. 33 million sentences). Moreover, additional newspaper texts (2 million sentences) and the complete Icelandic Wikipedia is included. For a very large mixed genre corpus, all these resources were combined yielding a corpus of more than 550 million running words. For details of the processing, see (Goldhahn et al., 2012). In 2012, these sentences were POS-tagged as described below.

The generated data can be browsed at a dedicated web portal[1] that provides a Web interface focusing on word form based statistical information. As an example Fig. 1 and 2 show word co-occurrences graphs for the word *skipti*. This word is ambiguous, has two meanings and appears with two possible word classes: as a noun [*time/occasion/separation/change/exchange*] and as a verb [*separate/change/exchange*]. Figure 1 and 2 show the different word co-occurences based on sentences. Note, that the differing co-occurrences illustrate the two different meanings and contexts.

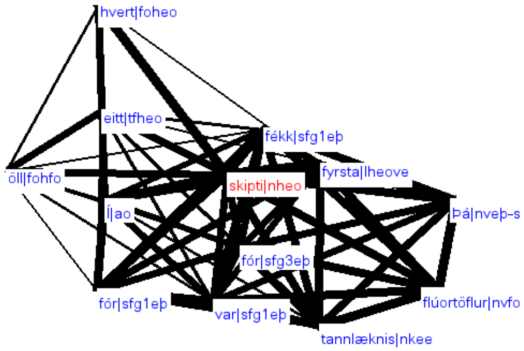

Figure 1: Sentence co-occurrences for the noun skipti

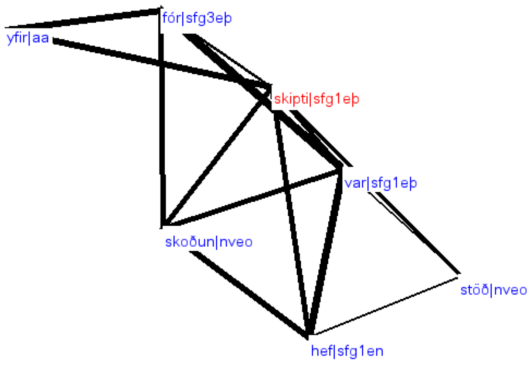

Figure 2: Sentence co-occurrences for the verb skipti

1 http://wortschatz.uni-leipzig.de/ws_isl/

2. Icelandic Part-of-Speech Tagging

2.1. Combined Tagging

The Icelandic corpus of the Leipzig Corpora Collection was annotated using the same procedures and software as the Tagged Icelandic Corpus, *MÍM*, (Helgadóttir et al., 2012). The annotation consists of sentence segmentation, tokenisation and morphosyntactic tagging. The LCC Icelandic corpus was not lemmatized. A special program, *CorpusTagger*, was developed for these tasks for the development of *MIM-GOLD*, a new gold standard for tagging Icelandic (Loftsson et al., 2010). The program uses IceNLP (Loftsson and Rögnvaldsson, 2007) for tokenisation and sentence segmentation. The text was then tagged with four different taggers, after which *CombiTagger* (Henrich et al., 2009) was applied to select a single tag by using simple majority voting. In the original work (Loftsson et al., 2010), the text was tagged with five different taggers (listed in descending order of accuracy when tagging Icelandic text): *IceTagger* (Loftsson, 2008), *Bidir* (Dredze and Wallenberg, 2008), *TnT* (Brants, 2000), *fnTBL* (Ngai and Florian, 2001), and *MXPOST* (Ratnaparkhi, 1996). The *Bidir* tagger had to be dropped from the procedure since it did not seem to be able to handle large quantities of text. The *TnT* tagger was replaced with *TriTagger,* which is Hrafn Loftson's re-implementation of *TnT* (Loftsson et al., 2011).

The tagset used for the corpus was developed for the making of the *IFD* corpus (Pind et al., 1991). The tags are character strings where each character in the tag has a particular function. The first character denotes the word class and the remaining characters (up to 5) denote various morphological features, such as gender, number and case. The *IFD* tagset has about 700 tags. The *IFD* corpus was tagged with a program that used a combination of grammatical rules and frequency information and then all tags were corrected manually. The *IFD* corpus has been used for training the data-driven taggers (*TriTagger*, *MXPOST* and *fnTBL*) as well as developing the rule-based tagger *IceTagger*.

2.2. Tag Frequencies

The tagged corpus contains around 3.9 million tagged types with around 5.6 million different type-tag combinations. Hence, on average every type was tagged with 1.4 different tags. 575 different tags were actually found in the corpus. Table 4 shows the most frequent tags in the corpus with their absolute and relative frequency. All values are based on tagged types, token frequencies are not taken into account.

The most frequent POS tags that do not describe nouns or numerals are "e" (16th most frequent tag, denoting a foreign word) and "lkensf" (34th, adjective (Masc., Sg., Nom., Strong declension, positive)). The most frequent POS tag denoting a verb ("sng") occurs on 40th place with an absolute frequency of 45,744 .

For several reasons, a word might be (correctly or not) tagged with different tags in different sentences. Table 5 shows the number of types being assigned with different numbers of tags. If all combinations of word and type tag that occur less frequently than a certain minimum are removed, a higher percentage of the types are assigned multiple tags.

In the complete corpus 24,758 different combinations of two POS tags were seen that were assigned to the same type. Only 9,637 of these combinations occur more than 10 times (i.e. more than 10 types were tagged with both tags). The following table shows the most frequent assignment of two POS tags to the same type.

Tag1	*Tag2*	*Frequency*
nken-s	nkeo-s	63,090
nken-s	nkeþ-s	59,785
nveo	nveþ	55,495
nhen	nheo	51,774
nkeo-s	nkeþ-s	40,564

Table 1: Typical combinations of POS tags for words having multiple tags

As expected, the high morphological variety in Icelandic leads to multiple assignments. The reason for these assignments are identical word forms for different grammatical categories of the same word (e.g. *woman*: kona|nven – konu|nheo – konu|nveþ – konu|nvee). Furthermore, there are multiple assignments due to identical forms of words of different word classes, that differ in meaning and syntactical characteristics.

In addition, results based on word classes were generated (similar to Petrov's universal tagset (Petrov, 2011) containing only 10 word classes). They can be used to simplify results of the tagging process or to compare results of different taggers using different sets of POS tags. The following tables show the distribution of these word classes in the corpus. Because of the diverse input material used, the tagged sentences contained non-standard Icelandic sentences (as often found in message board entries) and in some rare cases also non Icelandic material which was included due to errors in the preprocessing. To reduce the impact of these problematic parts, the same statistics were generated where only type-wordclass combinations were included that occurred at least 30 times. Apparently, the distribution of word classes in the LCC corpus based on word types has a strong bias towards nouns when compared with values based on the Icelandic Frequency Dictionary (Pind et al., 1991)[2]. These differences almost disappear if the distributions for tokens are compared (cf. Table 6).

2 Higher numbers than the overall number of types are due to ambiguous words.

As before the frequency of types having multiple word class tags is depicted. Table 2 shows the results.

Number of assigned word class tags	*Number of types*
1	3,707,666
2	152,947
3	25,339
4	4,229
5	654
6	125
7	19
8	3

Table 2: Number of types having multiple word class tags

Accordingly the following table takes a closer look at typical combinations of tags for words having multiple word classes.

Word class 1	*Word class 2*	*Absolute Frequency*
ADJ	NOUN	387,542
NOUN	VERB	218,354
ADJ	VERB	114,388
FOREIGN	NOUN	110,298
ADV	NOUN	31,871
ADV	ADJ	20,534
FOREIGN	ADJ	13,034
FOREIGN	VERB	10,306
NOUN	NUM	9,683
ADV	VERB	6,650

Table 3: Typical combinations of POS tags for words having multiple word class tags

3. Applications

3.1. Frequencies for Word Forms and Lemmas

For the frequencies of the corresponding lemmas, the frequencies of all of its inflected forms have to be summed up. For this application, some POS-taggers provide the lemma for each word form. Unfortunately there are two sources for counting errors: sometimes the lemma form provided is wrong. This can be the case for ambiguous word forms belonging to multiple lemmas or errors in lemmatisation. In addition some POS-taggers use the word form itself as lemma or give no lemma at all if the lemmatisation fails. Hence, the frequencies for lemmas generated by POS-taggers should be used with care.

This is the main reason why frequencies for word forms are used in Icelandic Frequency Dictionary (Quasthoff et al. 2012) which was created using the Icelandic corpus described here.

3.2. Text-To-Speech

In autumn 2012 the Icelandic organization of the visually impaired (*Blindrafélagið*[3]) introduced new Icelandic text-to-speech software. The voice software – the male voice Karl and the female voice Dóra – was developed by the Polish company Ivona. The linguistic material in the text-to-speech software was provided by this Icelandic corpus. Hence, the recording corpus compiled by Ivona for building the Icelandic voices was created by using sentences from the corpus. Furthermore also the Icelandic language model used in the text-to-speech software was based on systematic analyses of structures (intonation structures, grammar, syntax and other text structures) in the Icelandic corpus.

3.3. Spellchecking

Skrambi, an Icelandic spellchecker which is currently under development, uses a language model derived, in part, from word frequencies from the corpus. The spellchecker, which is based on the noisy channel model approach to spelling correction (Brill and Moore, 2000), uses a language model as well as an error model in order to estimate the probability that a given suggestion is correct. The error model is trained on 5.000 of the most common nonword errors found in the corpus.

4. References

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing* (ANLC '00). Association for Computational Linguistics, Stroudsburg, PA, USA, 224-231. DOI=10.3115/974147.974178 http://dx.doi.org/10.3115/974147.974178

Brill, E., Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong.

Dredze, M., Wallenberg, J. (2008). Icelandic Data Driven Part of Speech Tagging. In *Proceedings of the 46 th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT, Columbus, OH, USA.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

3 Web page: http://www.blind.is/

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., Richter, M. (2007). Íslenskur orðasjóður - Building a Large Icelandic Corpus. In: *Proceedings of NODALIDA-07*, Tartu, Estonia, 2007.

Helgadóttir, S.; Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages - SaLTMiL 8 - AfLaT2012*, s. 67-72. Istanbúl, Tyrklandi.

Henrich, V., Reuter, T., Loftsson, H. (2009). CombiTagger: A System for DevelopingCombined Taggers. In *Proceedings of the 22nd International FLAIRS Conference*, Special Track: Applied Natural Language Processing, Sanibel Island, Florida, USA.

Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. In *Nordic Journal of Linguistics*, 31(1), 47-72. © 2008 Cambridge University Press.

Loftsson, H., Rögnvaldsson, E. (2007). IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session: "Speech and language technology for less-resourced languages"*. Antwerp, Belgium.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In *Proceedings of Creation and use of basic lexical resources for less-resourced languages, workshop* at the 7th International Conference on Language Resources and Evaluation, LREC 2010. Valetta, Malta.

Loftsson, H.; Helgadóttir, S.;Rögnvaldsson, E. (2011). Using a morphological database to increase the accuracy in PoS tagging. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.

Ngai, G.; Florian, R. (2001). Transformation-based learning in the fast lane. In *Proceedings of North Americal ACL 2001*, pages 40-47, June 2001.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

Pind, J., (ed.); Magnússon, F., Briem, S. (1991). *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

Quasthoff, U., Fiedler, S., Hallsteinsdóttir, E. (eds.) (2012). Frequency Dictionary Icelandic. Leipziger Universitätsverlag.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, Philadelphia, USA.

POS tag	*Description*	*Absolute frequency of types*	*Relative frequency of types in %*
nken-s	Masculine proper name (Nom., Sg.)	353,439	6.4784
ta	Numeral	273,074	5.0054
nken	Masculine noun (Nom., Sg.)	171,016	3.1347
nven	Feminine noun (Nom., Sg.)	152,936	2.8033
nhen	Neuter noun (Nom., Sg.)	148,149	2.7155
nkeþ-s	Masculine proper name (Dat., Sg.)	140,736	2.5796
nveo	Feminine noun (Acc., Sg.)	132,370	2.4263
nveþ	Feminine noun (Dat., Sg.)	131,259	2.4059
nkeo	Masculine noun (Acc., Sg.)	130,846	2.3984
nheo	Neuter noun (Acc., Sg.)	130,294	2.3882

Table 4: Most frequent POS tags

Number of assigned POS tags	*Number of types (Absolute number + Percentage)*	*After removal of all combinations that occurred less than 3 times*	*After removal of all combinations that occurred less than 10 times*
1	3,114,693 (80.05%)	712,536 (72.70%)	277,115 (71.91%)
2	446,511 (11.48%)	151,161 (15.42%)	63,948 (16.59%)
3	165,388 (4.25%)	61,559 (6.28%)	25,693 (6.67%)
4	71,074 (1.83%)	24,937 (2.54%)	9,149 (2.37%)
5	36,009 (0.93%)	12,173 (1.24%)	4,103 (1.06%)
6	19,880 (0.51%)	6,306 (0.64%)	2,050 (0.53%)
7	11,635 (0.30%)	3,602 (0.37%)	1,055 (0.27%)
8	7,349 (0.20%)	2,265 (0.23%)	691 (0.18%)
9	4,986 (0.13%)	1,534 (0.16%)	400 (0.10%)
Sum	3,891,025 (100%)	980,098 (100%)	385,362 (100%)

Table 5: Number of types having multiple POS tags

Word class tag	*For types*	*For token*	*For types (freq>=30)*	*For token (freq>=30)*	*Frequency Dictionary Types*	*Frequency Dictionary Token*
Noun	4,208,385 (**77.14%**)	82,210,473 (**29.33%**)	185,567 (**70.24%**)	71,139,798 (**28.74%**)	**67.7%**	**23.6%**
Adjective	594,693 (**10.90%**)	17,810,552 (**6.35%**)	38,478 (**14.56%**)	60,516,649 (**6.51%**)	**15.9%**	**6.9%**
Numeral	273,976 (**5.02%**)	8,385,095 (**2.99%**)	9,960 (**3.77%**)	7,763,044 (**3.14%**)	**3.4%**	**1.1%**
Verb	253,325 (**4.64%**)	52,572,305 (**18.75%**)	23,427 (**8.87%**)	51,845,267 (**20.95%**)	**7.4%**	**19.9%**
Other (foreign words etc.)	91,592 (**1.68%**)	938,745 (**0.33%**)	2,750 (**1.04%**)	714,604 (**0.29%**)	**0.9%**	**0.1%**
Adverb and preposition	28,711 (**0.53%**)	60,601,142 (**21.62%**)	2,707 (**1.03%**)	60,516,649 (**24.45%**)	**4.2%**	**22.4%**
Conjunction	2,145 (**0.039%**)	32,049,776 (**11.43%**)	89 (**0.03%**)	31,986,948 (**12.92%**)	**0.1%**	**11.6%**
Pronoun	2,078 (**0.038%**)	25,404,021 (**9.06%**)	949 (**0.36%**)	25,398,212 (**10.26%**)	**0.1%**	**14.3%**
Unanalyzed word	628 (**0.012%**)	134,850 (**0.05%**)	68 (**0.03%**)	121,872 (**0.09%**)	**0.2%**	**0.0%**
Determiners	59 (**0.001%**)	234,760 (**0.08%**)	34 (**0.01%**)	234,617 (**0.09%**)	**0.0%**	**0.1%**

Table 6: Distribution of word classes in the LCC Icelandic corpus compared with the IFD corpus